Knee Surgery, Sports Traumatology, Arthroscopy (2020) 28:3207–3216
https://doi.org/10.1007/s00167-019-05822-7

KNEE

Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty

Yong-Hao Pua[1] · Hakmook Kang[2] · Julian Thumboo[3] · Ross Allan Clark[4] · Eleanor Shu-Xian Chew[1] · Cheryl Lian-Li Poon[1] · Hwei-Chi Chong[5] · Seng-Jin Yeo[6]

Received: 16 October 2019 / Accepted: 5 December 2019 / Published online: 12 December 2019

Abstract

Purpose Machine-learning methods are flexible prediction algorithms with potential advantages over conventional regression. This study aimed to use machine learning methods to predict post-total knee arthroplasty (TKA) walking limitation, and to compare their performance with that of logistic regression.

Methods From the department's clinical registry, a cohort of 4026 patients who underwent elective, primary TKA between July 2013 and July 2017 was identified. Candidate predictors included demographics and preoperative clinical, psychosocial, and outcome measures. The primary outcome was severe walking limitation at 6 months post-TKA, defined as a maximum walk time $\leq$ 15 min. Eight common regression (logistic, penalized logistic, and ordinal logistic with natural splines) and ensemble machine learning (random forest, extreme gradient boosting, and SuperLearner) methods were implemented to predict the probability of severe walking limitation. Models were compared on discrimination and calibration metrics.

Results At 6 months post-TKA, 13% of patients had severe walking limitation. Machine learning and logistic regression models performed moderately [mean area under the ROC curves (AUC) 0.73–0.75]. Overall, the ordinal logistic regression model performed best while the SuperLearner performed best among machine learning methods, with negligible differences between them (Brier score difference, $<$ 0.001; 95% CI [$-$0.0025, 0.002]).

Conclusions When predicting post-TKA physical function, several machine learning methods did not outperform logistic regression—in particular, ordinal logistic regression that does not assume linearity in its predictors.

Level of evidence Prognostic level II

Keywords Knee · Replacement · Prediction · Machine learning · Arthroplasty · Artificial intelligence · Algorithms

Introduction

Previous studies have indicated that 11–20% of patients reported dissatisfaction following total knee arthroplasty (TKA), and that patient dissatisfaction was associated with persistent functional limitations [15]. Thus, early and accurate identification of patients at risk for poor post-TKA functional outcomes would be preferable in terms of directing resources toward preventive care.

Reviewing the literature, most clinical prediction models for post-TKA physical function [6, 30, 34], including one of ours [30], have been developed using conventional regression analyses. However, machine learning, a data analysis technique that develops algorithms to predict outcomes by iteratively "learning" from data, is increasingly emphasized in orthopaedics [3] and rheumatology [24] as a competitive alternative to regression analysis. Importantly, machine learning has the potential to outperform conventional regression, possibly through its ability to capture nonlinearities and complex interactions among multiple predictor variables [12]. Despite this, only three studies [10, 16, 22] have used machine learning algorithms to predict post-TKA physical

Yong-Hao Pua and Hakmook Kang contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00167-019-05822-7) contains supplementary material, which is available to authorized users.

✉ Yong-Hao Pua
pua.yong.hao@sgh.com.sg

Extended author information available on the last page of the article

function, and only two studies [10, 22] compared their performance with that of logistic regression.

Given the clinical importance of identifying patients who are at risk for poor functional outcomes and given the paucity of machine learning studies in TKA, this study aimed to use machine learning methods to predict post-TKA walking limitation, and to compare their performance with that of logistic regression. It is hypothesized that machine learning algorithms outperform multivariable logistic regression models in terms of discrimination between severe and non-severe walking limitations.

Methods

This was a single-centre, cohort study at Singapore General Hospital—the largest tertiary teaching hospital in Singapore which performed half of all knee arthroplasties in a nation of 5.6 million people. From the department's database, 5491 patients aged ≥ 50 years who underwent a unilateral primary TKA between July 2013 and July 2017 were identified. Patients who underwent a revision knee surgery within 6 months post-TKA ($n = 16$) were excluded. Patients who had a history of rheumatoid arthritis ($n = 58$) and patients with stroke or Parkinson disease ($n = 108$) were also excluded. For patients with consecutive admissions for TKA ($n = 863$), only data from their first admission were used. Of the remaining 5309 patients, a cohort of 4026 patients with non-missing 6-month follow-up outcomes were selected. Included patients were similar to those who were excluded because of missing data (Appendix Table 1 in ESM). All data were collected by physiotherapists and data technicians trained in the testing procedures and entered into an electronic registry database as per routine practice policies of the institution. Data were de-identified prior to analyses. The institutional review board approved the study with a waiver of informed consent (SingHealth CIRB 2014/2027, Singapore).

Outcome

The primary outcome was severe postoperative (6 months) walking limitation. An intermediate (6 months) time point was chosen because (1) model prediction accuracy may decrease with a longer time horizon and (2) knowledge of intermediate-term (6 months) risk for poor TKA outcomes will aid patient education and assist in rehabilitation planning. Patients were asked to estimate the time they were able to walk (without a rest) before they had severe difficulty with the operated knee. This variable had four categories: (1) > 30 min, (2) 16–30 min, (3) 5–15 min, and (4) around the house only. Severe walking limitation was defined as a maximum walk time of ≤ 15 min (severe walking limitation = 1 for those who were in categories 3 and 4 and severe walking limitation = 0 otherwise).

Predictor variables

Predictor variables were selected based on clinical expertise, literature review [6, 16, 38], and data availability in the department's databases. To improve the practicality of the prediction models, variables which were less equipment dependent and were routinely and easily measured in the clinical setting were considered. Altogether, 25 predictors were identified and they included demographics and preoperative clinical, psychosocial, and outcome measures (Table 1). Of note, these clinical, psychosocial, and outcome measures were mainly derived from the Short Form 36 (SF-36) health survey, Oxford Knee Questionnaire, and Knee Society Clinical Rating Scale, and previous studies [7, 23, 26] have demonstrated good test–retest reliability for these instruments in patients with TKA (intraclass correlation coefficients 0.80–0.92).

Model development

Apart from the "education-level" variable which was missing at 7.7%, all other predictors were missing at very low levels (0.02% to 0.5%). Thus, the *transcan* function in the R [31] *Hmisc* [19] package was used to perform single imputation. Eight common regression (logistic, ordinal logistic with splines, L-1 penalized logistic [35], L-2 penalized logistic [20], and L-1/L-2 penalized logistic [42]) and ensemble machine learning (random forest, extreme gradient boosting, and SuperLearner) methods were implemented to predict the probability of severe walking limitation. Notably, these machine learning methods were chosen because they were successfully used in clinical research [9, 27]. All analyses were done with the *rms* [18], *Superlearner* [29], *caret* [25], and *vip* [13] *R* packages (http://www.r-project.org).

Logistic regression

A logistic regression model that included all variables was first fitted. To create a reference model against which performance of all other models can be compared, additive predictor effects were assumed and the (regression) coefficients for continuous predictors were linearly associated with the logit of the probability of having severe walking limitations. A proportional odds ordinal logistic regression model was then fitted on the ordinal (non-dichotomized) walking limitation outcome, and all continuous predictors in the ordinal model were modelled as restricted cubic splines with three knots [8, 17]. Three penalized logistic models were further fitted. The first model was a logistic regression with least absolute shrinkage and selection

Table 1 Patient demographics and preoperative clinical characteristics

Variables	Month-6 severe walking limitation			*P* value
	Absent ($n=3488$)	Present ($n=538$)	Overall ($n=4026$)	
Age	62.5 **67.4** 72.9 (67.5 ± 7.4)	65.1 **70.5** 76.1 (70.2 ± 8.2)	62.8 **67.8** 73.4 (67.9 ± 7.5)	< 0.001[1]
Weight (kg)	58 **66** 74 (67 ± 13)	58 **67** 76 (68 ± 14)	58 **66** 74 (67 ± 13)	0.15[1]
Height (m)	1.5 **1.6** 1.6 (1.6 ± 0.1)	1.5 **1.5** 1.6 (1.5 ± 0.1)	1.5 **1.6** 1.6 (1.6 ± 0.1)	< 0.001[1]
BMI (kg/m^2)	24.4 **27.0** 30.1 (27.5 ± 4.4)	24.8 **28.1** 31.7 (28.5 ± 5.2)	24.4 **27.1** 30.3 (27.6 ± 4.6)	< 0.001[1]
Race				< 0.001[2]
Chinese	86% (3009)	77% (415)	85% (3424)	
Malay	7% (245)	11% (61)	8% (306)	
Indian	5% (172)	10% (53)	6% (225)	
Others	2% (62)	2% (9)	2% (71)	
Women	74% (2574)	80% (429)	75% (3003)	0.003[2]
Contralateral knee pain	59% (2054)	68% (368)	60% (2422)	< 0.001[2]
Hypertension	60% (2082)	66% (354)	61% (2436)	0.007[2]
Dyslipidemia	40% (1410)	43% (233)	41% (1643)	0.2[2]
Diabetes	19% (662)	25% (132)	20% (794)	0.003[2]
Adult recon specialist	65% (2261)	59% (319)	64% (2580)	0.013[2]
Caregiver available	73% (2550)	69% (369)	73% (2919)	0.029[2]
Education Level				< 0.001[2]
None	19% (648)	32% (174)	20% (822)	
Primary	38% (1327)	36% (196)	38% (1523)	
Secondary	33% (1139)	24% (130)	32% (1269)	
Tertiary	11% (374)	7% (38)	10% (412)	
Gait aids				< 0.001[3]
None	73% (2555)	48% (258)	70% (2813)	
Stick	22% (768)	37% (199)	24% (967)	
Quadstick	3% (98)	9% (47)	4% (145)	
Walking frame	2% (67)	6% (34)	3% (101)	
Knee pain				< 0.001[3]
None or very mild	2% (71)	1% (7)	2% (78)	
Mild	16% (554)	9% (48)	15% (602)	
Moderate	46% (1607)	35% (187)	45% (1794)	
Severe	36% (1256)	55% (296)	39% (1552)	
Depression level				< 0.001[3]
Most or all	3% (112)	6% (33)	4% (145)	
A good bit	5% (174)	9% (48)	6% (222)	
Some	16% (573)	20% (108)	17% (681)	
A little	12% (422)	13% (71)	12% (493)	
None	63% (2207)	52% (278)	62% (2485)	
Anxiety level				0.56[3]
Most or all	4% (132)	4% (22)	4% (154)	
A good bit	4% (141)	4% (21)	4% (162)	
Some	14% (497)	16% (87)	15% (584)	
A little	16% (569)	15% (81)	16% (650)	
None	62% (2149)	61% (327)	62% (2476)	
Difficulty when climbing down stairs				< 0.001[3]
None	8% (283)	4% (20)	8% (303)	
Little	20% (705)	11% (58)	19% (763)	
Moderate	28% (977)	21% (113)	27% (1090)	
Extreme	37% (1293)	46% (249)	38% (1542)	

Table 1 (continued)

Variables	Month-6 severe walking limitation			*P* value
	Absent ($n=3488$)	Present ($n=538$)	Overall ($n=4026$)	
Unable	7% (230)	18% (98)	8% (328)	
Difficulty when kneeling and getting up				< 0.001[3]
None	2% (66)	1% (4)	2% (70)	
Little	3% (108)	1% (6)	3% (114)	
Moderate	4% (136)	3% (16)	4% (152)	
Extreme	5% (186)	4% (20)	5% (206)	
Unable	86% (2992)	91% (492)	87% (3484)	
Knee flexion	109 **121** 131 (118 ± 18)	105 **119** 129 (116 ± 19)	108 **121** 131 (118 ± 18)	0.004[1]
Knee extension	2.0 **6.0** 10.0 (7.1 ± 6.9)	3.0 **6.0** 11.0 (7.8 ± 8.1)	3.0 **6.0** 10.0 (7.2 ± 7.1)	0.26[1]
SF-36 physical function	25 **40** 55 (41 ± 23)	10 **20** 35 (23 ± 19)	20 **35** 55 (38 ± 23)	< 0.001[1]
Walking limitation				< 0.001[3]
> 30 min	19% (650)	4% (21)	17% (671)	
16–30 min	31% (1068)	14% (74)	28% (1142)	
5–15 min	42% (1469)	52% (282)	43% (1751)	
Around house	9% (301)	30% (161)	11% (462)	

The median 50th percentile values are in bold

[1]Continuous variables are summarized as 25th, **50**th, 75th percentiles (mean ± SD), and tested with the Wilcoxon–Mann–Whitney test

[2]Categorical variables are summarized as percentages and frequencies (*N*), and tested with the Pearson's χ^2 test

[3]The proportional odds likelihood ratio test

operator (LASSO)—a state-of-the-art variable selection and shrinkage method. The second model was a logistic regression with ridge regularization—a shrinkage method that constraints all regression coefficients toward zero to reduce model variance. The third model was a logistic regression with elastic net regularization—a variable selection and shrinkage method that combines the LASSO and the ridge regularization penalties.

Random forest

Random forest is an ensemble tree-based machine learning method [1] that fits multiple classification and regression trees on bootstrap samples of the data. When fitting a tree, the random forest algorithm considers a random subset of the predictors at each node and iteratively identifies optimal splits in them to maximally separate the outcome into two groups with disparate outcome probabilities. Using a random number of predictor variables, diverse trees that are less correlated with one another are created, potentially increasing prediction accuracy. To reduce model variance, the random forest algorithm uses a "bagging" (bootstrap aggregation) procedure that averages predictions from diverse trees grown on bootstrap samples. The optimal number of predictor variables considered at each node was determined using repeated cross-validation of the training dataset (described later).

Extreme gradient boosting machine

Gradient boosting is an ensemble tree-based machine learning method that sequentially fits a series of classification and regression trees, with each tree created to predict the outcomes misclassified by the previous tree [11]. By creating trees to predict residuals of previous trees, the gradient boosting process focuses on predicting more difficult cases and corrects its own shortcomings. This "boosting" process continues iteratively, with the tree depth, learning rate, and number of trees were optimized using repeated cross validation. Extreme gradient boosting (XGBoost) is a specific implementation of the gradient boosting process [4], and uses memory-efficient algorithms to improve computational speed and model performance.

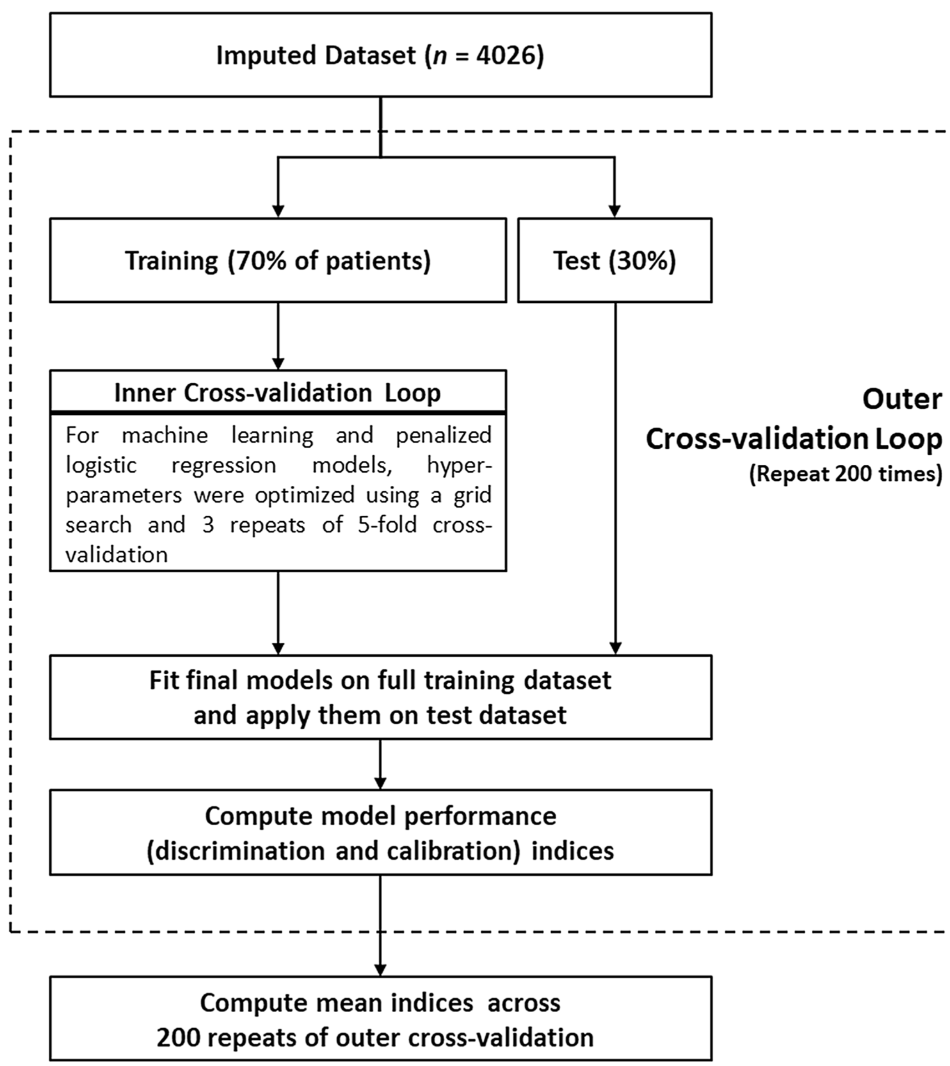
SuperLearner

SuperLearner is an ensemble machine learning method that creates an optimal prediction algorithm from a set of prediction algorithms [36]. By cross-validating these candidate algorithms, the SuperLearner optimally weighs and combines predictions from them, and this "stacking" process has been shown to be asymptotically as accurate as the best individual candidate algorithm. In this study, XGBoost, random forest, binary logistic regression, logistic regression with LASSO, and logistic regression with ridge regularization were employed as candidate prediction algorithms.

Model performance

To validate the models, nested cross-validation which comprised an outer and an inner cross-validation loop was used [39] (Fig. 1). In the outer loop, 200 repeated random splits of the dataset were performed, dividing it into training (70% of sample) and validation (30%) datasets. In the inner loop, hyperparameters of the machine learning and penalized logistic regression models were optimized using three repeats of fivefold cross-validation, and the final models (with the optimized hyperparameters) were fitted on the entire training dataset. These models were applied to the validation datasets and their performance was assessed in three ways. First, model discrimination was measured by the area under the receiver operating characteristic (AUC) curve, where a value of 1.0 represents perfect discrimination and 0.5 represents no discrimination ('coin flip'). Specifically, the AUC is the probability that a randomly chosen patient with the event (severe walking limitation) will have a higher predicted risk than a randomly chosen patient without the event. Second, model calibration was assessed using loess-smoothed calibration plots, and the *val.prob* function [18] implemented in *R* software was used to compute the mean absolute error in predicted and loess calibrated probabilities. Third, the Brier score [2], where a value of 0 represents perfect overall model performance was computed. Of note, while the AUC measures model discrimination, the Brier score is the mean quadratic difference between predicted probabilities and observed binary outcomes and thus, includes both discrimination and calibration aspects. From the outer cross-validation loop, the mean performance indices and their 95% confidence limits were computed. Finally, to gain insights into the relative contribution of the predictor variables in the best performing logistic regression and machine learning methods, the Wald χ^2 statistic minus the degrees-of-freedom ($\chi^2 - df$) [17] was computed for the best performing logistic regression and the AUC-based

Fig. 1 Analysis pipeline



permutation importance measure [1, 13, 14] was computed for the best performing machine learning method.

Results

Table 1 shows the demographics and preoperative characteristics of the patients. Preoperatively, half of the patients (55%, 95% CI 53–57%) reported an inability to walk for more than 15 min; at 6 months post-surgery, just over one of every ten patients (13%, 95% CI 12–14%).

Figure 2 summarizes the discrimination and calibration performance of machine learning and logistic regression models. In terms of model discrimination, the cross-validated AUCs of the best performing logistic regression model (ordinal regression) and the best machine learning method (XGBoost) were similar (AUC difference, 0.002; 95% CI [−0.015, 0.018]). Notably, compared with the binary logistic regression model (mean AUC: 0.751), the AUC difference was greatest for ordinal logistic regression model with spline terms (AUC difference, 0.006; 95% CI [−0.008, 0.019]). The random forest model was least discriminative.

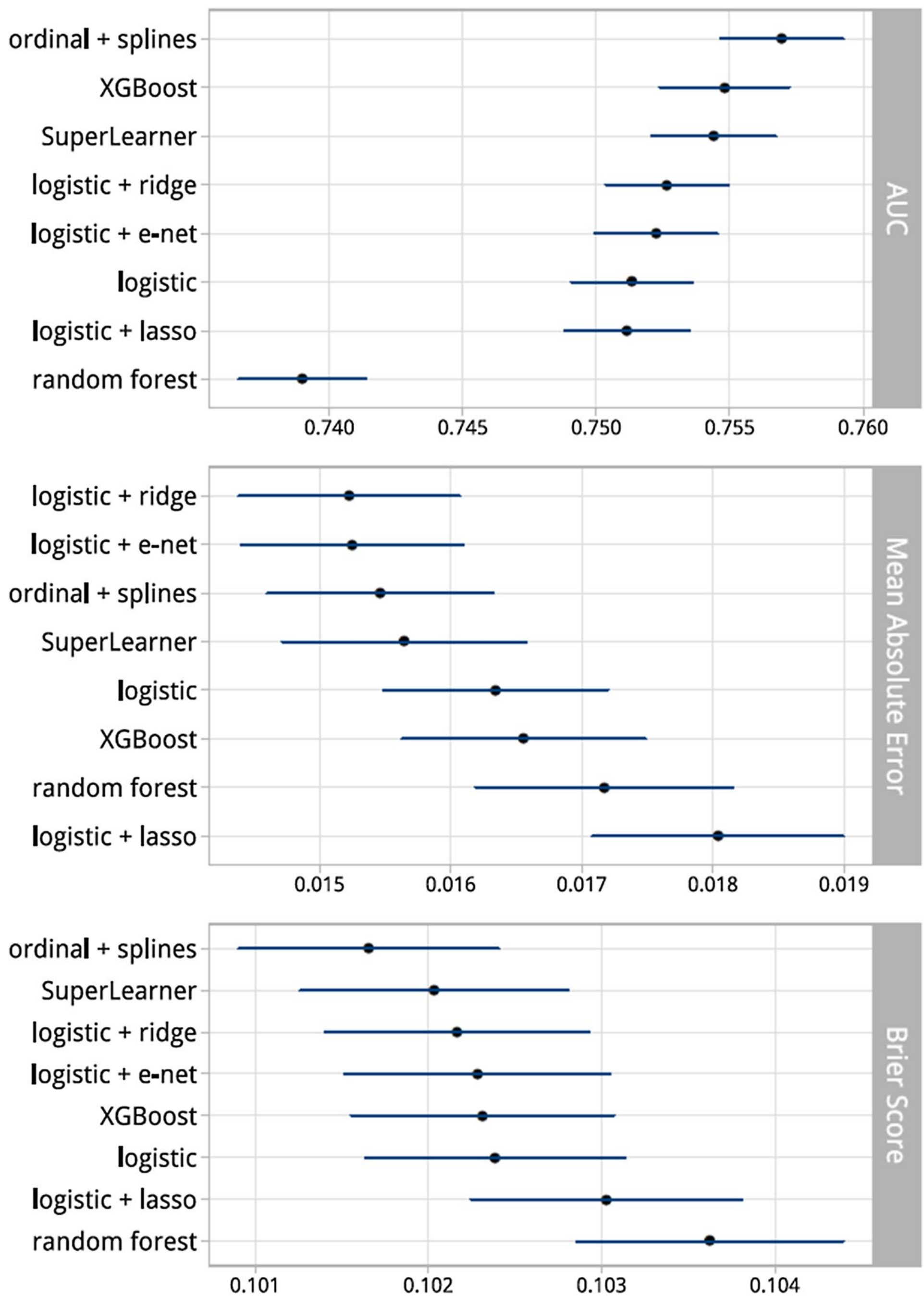


Fig. 2 Results of discrimination and calibration metrics of machine learning and logistic regression models computed from nested cross-validation. Area under the receiver operating characteristic curve (AUC), mean absolute error, and Brier score values are represented for each model by mean and 95% confidence intervals (95% CIs). XGBoost: extreme gradient boosting; e-net: elastic net regression

In terms of calibration performance, indexed by mean absolute error in predicted and loess calibrated probabilities, ordinal and penalized (L-1/L-2 and L-2 norms) logistic models and the SuperLearner were amongst the best calibrated models. Similarly, in terms of overall performance, indexed by the Brier score, ordinal logistic regression model (with spline terms) was the best performing model while the SuperLearner was the best performing machine learning method. Figure 3 shows the top ten predictor variables in the ordinal regression model and the SuperLearner. For both methods, age and preoperative physical function (that is, preoperative walking limitation levels, type of gait aids used preoperatively, and preoperative SF-36 physical function) were among the most important predictors. Although the top four predictors were the same (with different ranks) between the two methods, some predictors were unique in the ordinal logistic model such as "Operated by adult reconstruction specialist", "Depression", and "Race:Malay" while "Anxiety" and "Kneeling difficulty" were only listed in the SuperLeaner method. To help clinical readers interpret the potential contributions of individual predictors to the SuperLearner predictions, a linear regression model was fitted to approximate predictions from the SuperLearner, and Appendix Table 2 in ESM gives the odds ratios from both ordinal logistic and SuperLearner (approximated) models.

Discussion

The key finding of this cohort study was that machine learning algorithms did not improve the predictions of post-TKA severe walking limitation compared with logistic regression models. Thus, the study hypothesis was not confirmed. All models showed moderate discrimination, with AUC statistics above 0.73. Similar to the logistic regression model, the best performing machine learning model identified older age and poorer preoperative physical function as important predictors of more severe walking limitations (Fig. 3 and Appendix Table 2 in ESM). Thus, these findings give clinicians and researchers confidence in the machine learning approach.

Reviewing the literature, the AUC values are generally similar to those of Fontana et al. [10], who investigated logistic regression with LASSO (0.75), random forest (0.75), and support vector machine (0.73) in predicting post-TKA Knee Disability and Osteoarthritis Outcome Scores for joint replacement (KOOS-Jr) scores. The findings also agree with the conclusion of a recent systematic review that machine learning was not superior to logistic regression in clinical prediction modelling [5]. More specifically, they are consistent with those of Huber et al. [22]—a TKA study that was similar in design to this study. Huber et al. [22] investigated

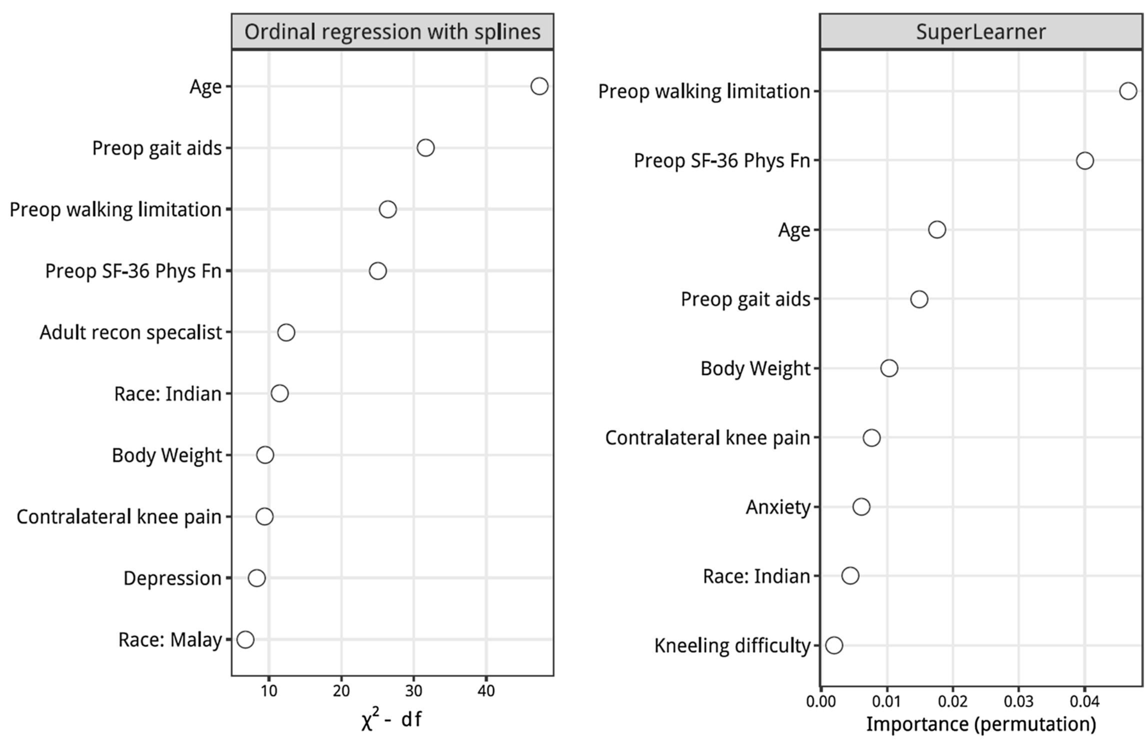


Fig. 3 Relative contribution of the top ten predictor variables in ordinal regression model (ranked by their $\chi^2 - df$ values) and SuperLearner model (ranked by their AUC-based permutation importance)

eight machine learning and logistic regression models to predict post-TKA physical function, and they reported that the AUC values of XGBoost and logistic regression were nearly identical (0.70). Because machine learning models could potentially outperform logistic regression by allowing nonlinearities and interactions among predictors, this study and Huber et al.'s indicate that in the prediction of post-TKA physical function, predictors mainly act additively, with complex non-additive effects and nonlinearities not sufficiently pronounced to make machine learning methods beneficial.

In this study, the best performing method was ordinal logistic regression model (Fig. 2). At a time when machine learning papers tend to dichotomize (ordinal) outcomes and present "conventional" regression simply as a method that assumes linearity, this finding is timely and unsurprising: the proportional odds ordinal regression model shares the same structure as the binary logistic model but uses the full (ordinal) information in the outcome and thus has greater statistical power than a binary logistic regression [17]. Furthermore, because assuming linearity may reduce model performance [17], all continuous predictors in the ordinal model were expanded using restricted cubic splines. Given that ordinal regression has other positive elements, such as its ability to give exceedance probabilities at any outcome cut-points, this finding supports previous calls [32] for a wider adoption of ordinal outcome analysis.

Among the machine learning methods, the AUC of XGBoost was best and second only to that of ordinal regression overall. Although this finding is consistent with that of Huber et al. [22], discrimination is just one aspect of model performance [5, 35]. Given that model calibration accuracy is also important in the field of orthopaedics, this study went further and found that XGBoost was not the best calibrated model. Instead, the SuperLearner was, overall, the best performing machine learning algorithm, and this finding is consistent with prior literature in other clinical settings [28, 33]. As it is unlikely that one prediction algorithm will be most accurate across all scenarios [41], a major strength of the SuperLearner framework is that it does not require the analyst to rely on a single algorithm [36]. However, as generally applies to machine learning algorithms, a potential drawback of the SuperLearner is that its "black box" nature may limit its interpretability and thus, its acceptance by the clinical community. Nevertheless, work is going on [14, 21] to develop methods for visualizing predictor variable importance. Figure 3 shows the top ten predictor variables from the best performing logistic and machine learning methods.

Limitations

This study has limitations. First, the data come from only one institution, though it does deliver care to a large segment of the nation's population. Second, a relatively small number of predictors were studied. Thus, one potential criticism is that we have not harnessed the ability of machine learning methods to handle numerous predictors, which may limit the ability of machine learning techniques to outperform traditional logistic regression modelling. That said, Huber et al. [22] considered 81 pre-specified predictors in their analyses and arrived at the same conclusion as this study. Furthermore, simulation analyses [37] have suggested that machine learning methods may require substantially larger effective sample sizes than logistic regression to avoid model overfitting and produce small optimism in cross-validated AUC. Third, although deep learning is the most rapidly emerging tool in biomedical research, deep learning was not employed to analyse the data because current application of deep learning focuses on imaging data analysis [40] and requires extremely large datasets. Thus, it is possible that in studies with radiographic or imaging predictors, deep learning or machine learning may outperform logistic regression. Due to limited performance of support vector machine (SVM) [9, 27], SVM was also not employed in this study.

Conclusion

In conclusion, when predicting post-TKA physical function, this study found that several machine learning methods were not more accurate than logistic regression—in particular, ordinal logistic regression that did not assume linear predictor effects. These findings suggest that both ordinal logistic regression and machine learning methods may be used to identify patients who are at high risk for severe walking limitations. Furthermore, to facilitate clinical interpretation of the machine learning model, variable importance and potential predictor effects were illustrated (Fig. 3 and Appendix Table 2 in ESM). It is hoped that this study will encourage future head-to-head comparisons between machine learning and "well-done" logistic regression, and the R codes used in this study are publicly available in an online repository (https://github.com/yhpua/RvML).

Acknowledgements We thank Brandon Greenwell for his generous help with the *vip* R package and Michael W. Wade at Vanderbilt University Medical Center for his editorial work on this article. We acknowledge the support from Jennifer Liaw, the head of the Department of Physiotherapy, Singapore General Hospital. We thank William Yeo from the Orthopaedic Diagnostic Centre, Singapore General Hospital, for his assistance. Finally, we thank Ee-Lin Woon, Felicia Jie-Ting Seah, Nai-Hong Chan, and the therapy assistants (Penny Teh and Hamidah Binti Hanib) for their kind assistance.

Funding No funding was provided for the completion of this study.

Compliance with ethical standards

Conflict of interest The authors have no professional or financial affiliations that may be perceived to have biased the presentation. Each author certifies that he or she has no commercial associations that might pose a conflict of interest in connection with the submitted article.

Ethical approval Ethical approval was provided by the SingHealth Centralized IRB (SingHealth CIRB 2014/2027, Singapore).

References

1. Breiman L (2001) Random forests. Mach Learn 45:5–32
2. Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78:1–3
3. Cabitza F, Locoro A, Banfi G (2018) Machine learning in orthopedics: a literature review. Front Bioeng Biotechnol 6:75
4. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794
5. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 110:12–22
6. Dowsey MM, Spelman T, Choong PF (2016) Development of a prognostic nomogram for predicting the probability of nonresponse to total knee arthroplasty 1 year after surgery. J Arthroplast 31:1654–1660
7. Dunbar M, Robertsson O, Ryd L, Lidgren L (2001) Appropriate questionnaires for knee arthroplasty: results of a survey of 3600 patients from The Swedish Knee Arthroplasty Registry. J Bone Joint Surg Br 83:339–344
8. Durrleman S, Simon R (1989) Flexible regression models with cubic splines. Stat Med 8:551–561
9. Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15:3133–3181
10. Fontana MA, Lyman S, Sarker GK, Padgett DE, MacLean CH (2019) Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? Clin Orthop 477:1267–1279
11. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
12. Goldstein BA, Navar AM, Carter RE (2017) Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 38:1805–1814
13. Greenwell B, Boehmke B, Gray B (2018) vip: variable importance plots. R package version 0.1.2. https://CRAN.R-project.org/package=vip. Accessed 10 Jan 2019
14. Greenwell BM, Boehmke BC, McCarthy AJ (2018) A simple and effective model-based variable importance measure. arXiv preprint. arXiv:1805.04755
15. Gunaratne R, Pratt DN, Banda J, Fick DP, Khan RJK, Robertson BW (2017) Patient dissatisfaction following total knee arthroplasty: a systematic review of the literature. J Arthroplast 32:3854–3860
16. Gutacker N, Street A (2017) Use of large-scale HRQoL datasets to generate individualised predictions and inform patients about the likely benefit of surgery. Qual Life Res 26:2497–2505
17. Harrell FE Jr (2015) Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, New York
18. Harrell Jr FE (2019) rms: regression modeling strategies. R package version 5.1-3. http://CRAN.R-project.org/package=rms. Accessed 10 Jan 2019
19. Harrell Jr FE, with contributions from Charles Dupont and many others (2019) Hmisc: Harrell Miscellaneous. R package version 4.2-0. https://CRAN.R-project.org/package=Hmisc. Accessed 10 Jan 2019
20. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67
21. Hubbard A, Kennedy C (2018) varimpact: variable importance estimation using targeted causal inference (TMLE). R package version 1.3.0-9004. http://github.com/ck37/varimpact. Accessed 10 Jan 2019
22. Huber M, Kurz C, Leidl R (2019) Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. BMC Med Inform Decis Mak 19:1–13
23. Impellizzeri FM, Mannion AF, Leunig M, Bizzini M, Naal FD (2011) Comparison of the reliability, responsiveness, and construct validity of 4 different questionnaires for evaluating outcomes after total knee arthroplasty. J Arthroplast 26:861–869
24. Jamshidi A, Pelletier JP, Martel-Pelletier J (2019) Machine-learning-based patient-specific prediction models for knee osteoarthritis. Nat Rev Rheumatol 15:49–60
25. Kuhn M (2019) caret: classification and regression training. R package version 6.0-82. https://CRAN.R-project.org/package=caret. Accessed 10 Jan 2019
26. Martimbianco ALC, Calabrese FR, Iha LAN, Petrilli M, Lira Neto O, Carneiro Filho M (2012) Reliability of the "American Knee Society Score"(AKSS). Acta Ortop Bras 20:34–38
27. Ogutu JO, Piepho HP, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc 5(Suppl 3):1–5
28. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ (2015) Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med 3:42–52
29. Polley E, LeDell E, Kennedy C, van der Laan M (2018) SuperLearner: super learner prediction. R package version 2.0-24. https://CRAN.R-project.org/package=SuperLearner
30. Pua YH, Poon CL, Seah FJ, Thumboo J, Clark RA, Tan MH et al (2019) Predicting individual knee range of motion, knee pain, and walking limitation outcomes following total knee arthroplasty. Acta Orthop 90:179–186
31. R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. Accessed 10 Jan 2019
32. Roozenbeek B, Lingsma HF, Perel P, Edwards P, Roberts I, Murray GD et al (2011) The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. Crit Care 15:1–7
33. Rose S (2013) Mortality risk score prediction in an elderly population using machine learning. Am J Epidemiol 177:443–452
34. Sanchez-Santos MT, Garriga C, Judge A, Batra RN, Price AJ, Liddle AD et al (2018) Development and validation of a clinical prediction model for patient-reported pain and function after primary total knee replacement surgery. Sci Rep 8:1–9
35. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 6:267–288
36. Van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Stat Appl Genet Mol Biol 6:1–21
37. van der Ploeg T, Austin PC, Steyerberg EW (2014) Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 14:1–13
38. Van Onsem S, Van Der Straeten C, Arnout N, Deprez P, Van Damme G, Victor J (2016) A new prediction model for patient satisfaction after total knee arthroplasty. J Arthroplast 31:2660–2667

39. van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruyt ND et al (2018) Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. Front Neurol 9:1–8
40. Wainberg M, Merico D, Delong A, Frey BJ (2018) Deep learning in biomedicine. Nat Biotechnol 36:829–838
41. Wolpert DH (1992) Stacked generalization. Neural Netw 5:241–259
42. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 67:301–320

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

Affiliations

Yong-Hao Pua[1] · Hakmook Kang[2] · Julian Thumboo[3] · Ross Allan Clark[4] · Eleanor Shu-Xian Chew[1] · Cheryl Lian-Li Poon[1] · Hwei-Chi Chong[5] · Seng-Jin Yeo[6]

Hakmook Kang
h.kang@vumc.org

Julian Thumboo
julian.thumboo@singhealth.com.sg

Ross Allan Clark
rclark@usc.edu.au

Eleanor Shu-Xian Chew
eleanor.chew.s.x@sgh.com.sg

Cheryl Lian-Li Poon
cheryl.poon.l.l@sgh.com.sg

Hwei-Chi Chong
hwei_chi_chong@cgh.com.sg

Seng-Jin Yeo
yeo.seng.jin@singhealth.com.sg

[1] Department of Physiotherapy, Singapore General Hospital, Singapore, Singapore

[2] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[3] Department of Rheumatology and Immunology, Singapore General Hospital, Singapore, Singapore

[4] Research Health Institute, University of the Sunshine Coast, Sunshine Coast, Australia

[5] Department of Physiotherapy, Changi General Hospital, Singapore, Singapore

[6] Department of Orthopaedic Surgery, Singapore General Hospital, Singapore, Singapore