Martin, R. K., Wastvedt, S., Pareek, A., Persson, A., Visnes, H., Fenstad, A. M., Moatshe, G., Wolfson, J., Lind, M., Engebretsen, L. (2023). Ceiling Effect of the Combined Norwegian and Danish Knee Ligament Registers Limits Anterior Cruciate Ligament Reconstruction Outcome Prediction. *American Journal of Sports Medicine, 51*(9), 2324 – 2332. http://dx.doi.org/10.1177/03635465231177905

Ceiling Effect of the Combined Norwegian and Danish Knee Ligament Registers Limits Anterior Cruciate Ligament Reconstruction Outcome Prediction

ABSTRACT

Background: Clinical tools based on machine learning analysis now exist for outcome prediction following primary anterior cruciate ligament reconstruction (ACLR). Relying partly on data volume, a general principle is that more data may lead to improved model accuracy.

Purpose/Hypothesis: To apply machine learning to a combined dataset comprised of the Norwegian and Danish knee ligament registers (KLR) with the aim of producing an algorithm that can predict subsequent revision surgery with improved accuracy relative to a previously published model developed using only the Norwegian register (NKLR). The hypothesis was that the additional patient data would result in an algorithm that is more accurate.

Study Design: Level IV retrospective review of a prospective cohort registry

Methods: Machine learning analysis was performed on the combined KLR. The primary outcome was the probability of revision ACLR within 1, 2, and 5 years. Data were split randomly into training sets (75%) and test sets (25%). Four machine learning models were tested: Cox Lasso, survival random forest, gradient boosted regression, and super learner. Concordance and calibration were calculated for all four models.

Results: The data set included 62,955 patients, where 5.1% underwent a revision surgical procedure during a mean follow-up of 7.6 ± 4.5 years. The three nonparametric models (random survival forest, gradient boosted regression, and super learner) performed best, demonstrating concordance in the moderate range (0.67, 95% CI 0.64-0.70) and were well calibrated at 1 and 2 years. Model performance was similar to the previously published model (concordance 0.67-0.69; well calibrated).

Conclusion: Machine learning analysis of the combined Norwegian and Danish knee ligament registers enabled prediction of revision ACLR risk with moderate accuracy. However, the resulting algorithms were less user-friendly and did not demonstrate superior accuracy in comparison to the previously developed model based on patients from the NKLR, despite the analysis of nearly 63,000 patients. This ceiling effect suggests that simply adding more patients to the current national knee ligament registers is unlikely to improve predictive capability and may prompt future changes to increase variable inclusion.

Clinical Relevance: For an improvement in the ability to predict outcome based on knee ligament registry data, an evolution in the variables recorded by the registers is required and the present study may prompt changes to increase variable collection.

Key Terms: ACL revision; outcome prediction; machine learning; artificial intelligence

What is known about the subject: Knee ligament registers enable estimation of a patient's specific risk of subsequent revision surgery.

What this study adds to existing knowledge: A ceiling effect has been reached and evolution of the registers (adding or changing recorded variables) is required to improve predictive capability.

INTRODUCTION

There has been an increased focus on outcome prediction using machine learning in orthopaedic surgery recently[22]. The primary goal of these early clinical predictive models is to enable patient-specific risk estimation to guide management discussions and expectations. Clinical tools based on machine learning analysis now exist for outcome prediction following anterior cruciate ligament (ACL) reconstruction (ACLR) including subsequent revision surgery[32] and inferior patient reported outcome[33]. These models were developed from analysis of the Norwegian Knee Ligament Register (NKLR) and the revision prediction model has also been externally validated using the Danish Knee Ligament Register (DKLR)[31].

Accurate prediction of outcome following ACLR holds value for both the patient and surgeon. However, with so many interrelated variables contributing to the risk of a poor outcome, it can be challenging for a clinician to quantify that risk for the patient in the office, regardless of experience level. Machine learning represents a novel approach to this problem and can facilitate patient-specific risk quantification through the analysis and interpretation of large volumes of data in ways that were previously unrealistic.

Relying partly on data volume to develop the predictive algorithms, a general principle is that more data may lead to improved model accuracy. The rationale for this is that more data presents more opportunity for the models to "learn" the association between predictors and outcome. Therefore, the purpose of this study was to apply machine learning to a combined NKLR and DKLR dataset with the aim of predicting subsequent revision surgery with improved accuracy relative to the previously published model[32]. The original NKLR model was developed through machine learning analysis of approximately 25,000 patients while the combined NKLR and DKLR dataset approaches 65,000 patients. The hypothesis was that the additional patient data would result in more accurate prediction of revision ACLR risk.

MATERIALS AND METHODS

This manuscript was written in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement[6]. The statement includes a 22-item checklist with the goal of improving the transparency of prediction model studies through full and clear reporting.

Ethics

All patients provide informed consent for the NKLR and the Norwegian Data Inspectorate grants permission for the register to collect, analyze, and publish on health data. Data registration was performed confidentially according to European Union (EU) data protection rules, with all data de-identified prior to retrieval. The Regional Ethics Committee (REK) states that it is not necessary to obtain further ethical approval[11]. Similarly, the DKLR obtains informed consent at the time of enrollment and patient data was de-identified prior to retrieval with no further ethical approval required.

Data preparation

Patients with primary ACLR surgery dates from June 2004 through December 2020 were included. Patients missing values for graft choice, those with graft choice recorded as "direct suture", and those missing values for the indicator of revision surgery were excluded. Variables considered for analysis are shown in Table 1.

Table 1: Demographics and Included Variables

Variable	Combined data N = 62,955
Revision†	3,205 (5.1%)
Follow-up time or time to revision*	7.6 (4.5)
Age at surgery‡	26 (20, 36)

Age at injury‡	24 (18, 34)
Missing	1,870
Sex†	
Male	36,509 (58%)
Female	26,446 (42%)
Pre-surgery KOOS QOL score (out of 10)*	3.63 (1.80)
Missing	29,512
Pre-surgery KOOS Sports score (out of 10)*	4.12 (2.69)
Missing	29,708
Below median on all pre-surgery KOOS†	6,372 (19%)
Missing	29,323
Activity that led to injury†	
Non-pivoting	20,391 (33%)
Pivoting	35,851 (57%)
Other	6,162 (9.9%)
Missing	551
Meniscus injury†	
Injury without repair	20,328 (32%)
Injury with repair	10,554 (17%)
None	32,061 (51%)
Missing	12
Cartilage injury†	
Grade 1-2	8,766 (14%)
Grade 3-4	3,223 (5.1%)
None	50,878 (81%)
Missing	88
Graft choice†	
Bone Patellar Tendon Bone	15,639 (25%)
Hamstring	43,518 (69%)
Quadriceps Tendon	2,520 (4.0%)
Other	1,278 (2.0%)
Tibia fixation device†	
Interference screw	55,792 (90%)
Suspension/cortical device	3,643 (5.9%)
Other	2,356 (3.8%)
Missing	1,164
Femur fixation device†	
Interference screw	16,434 (27%)
Suspension/cortical device	39,742 (65%)
Other	4,822 (7.9%)
Missing	1,957

Fixation device combination†	
Interference screw x2	15,865 (26%)
Interference/Suspension	236 (0.4%)
Suspension/cortical device x2	2,994 (4.9%)
Suspension/Interference	34,895 (58%)
Other	6,529 (11%)
Missing	2,436
Injured side†	
Right	32,147 (51%)
Left	30,807 (49%)
Missing	1
Previous surgery on opposite knee†	4,839 (8.1%)
Missing	2,946
Previous surgery on same knee†	10,312 (17%)
Missing	673
Time injury to surgery (years)‡	0.61 (0.33, 1.32)
Missing	2,083
Registry†	
DKLR	34,554 (55%)
NKLR	28,401 (45%)

* Mean (SD)
‡ Median (Interquartile range)
† n (%)
KOOS: Knee Injury and Osteoarthritis Outcome Score
QOL: Quality of Life
DKLR: Danish Knee Ligament Register
NKLR: Norwegian Knee Ligament Register

A predictor indicating if a patient was below the median score in the respective registry on all pre-surgery KOOS variables was created. Patients undergoing a revision ACLR prior to the follow-up time were considered to have experienced the event.

Machine learning modeling

The NKLR and DKLR data were combined then randomly split into training (75%) and test (25%) sets used to fit and evaluate the models, respectively. The primary outcome was probability of revision ACLR within 1, 2, and 5 years. R (version: 4.1.1, R Core Team 2021) was used to fit machine learning models that are adapted for censored, time-to-event data. "Censoring" refers to the fact that patients who have not yet reached a given follow-up time point may still contribute partial information toward that end point. For example, a patient who has been revision-free for four years has not yet reached the five-year selected outcome time-point, but their revision-free time can still be considered in the analysis for the five-year revision risk. Censoring also accounts for the fact that patients who have not yet had a revision may ultimately undergo a revision surgery in the future.

Four models intended for this type of data were used: Cox lasso, random survival forest, gradient boosted regression (GBM), and super learner. These models represent a range of approaches regarding flexibility of model fit and the number of variables incorporated. The Cox lasso is a semiparametric, penalized regression model that selects a subset of the most important predictor variables for inclusion[42]. The random survival forest is a nonparametric model, meaning that it does not require pre-specification of a model structure, and uses all available variables. This model is an adaptation of the widely used tree-based random forest method for censored data[17]. The GBM is also tree-based, nonparametric, and adapted for censored data; this model iteratively updates to improve the fit using all available variables[9]. The super learner model is an "ensemble" model that creates a weighted average of other machine learning techniques, combining them into one overall fit and thereby providing an even more flexible approach[23]. The super learner model combined random survival forest and GBM models. Further description of each model is included in Appendix A.

Variables with non-zero coefficients were selected using the L1-regularized Cox model ("Cox lasso", package *glmnet*, lambda value selected via cross-validation), retaining the variables shown in the top panel of Figure 1.

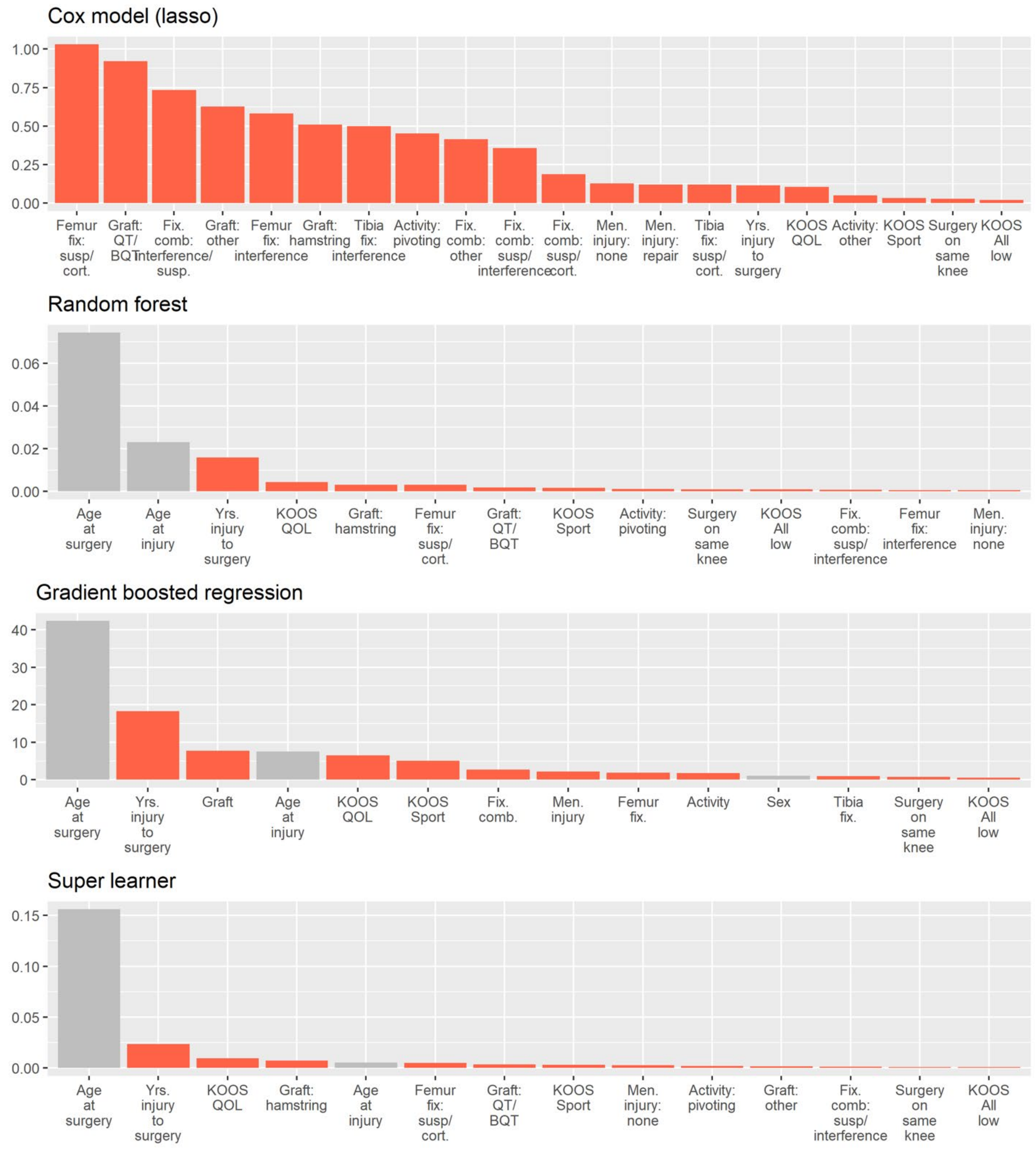

Figure 1: Feature importance

The four plots show relative feature importance in each of the machine learning models. The highlighted bars indicate features selected into the Cox Lasso model. Random forest, gradient boosted, and super learner plots show features in the top half by importance score, for readability. Feature importance is measured on a different scale for each model, and thus only rankings of features, rather than scores, should be compared among the models. The Cox Lasso measures feature importance by absolute effect size. The random forest and super learner models use permutation-based importance, which measures the relative change in model performance upon randomly permuting values of the given feature. The GBM uses difference in error rate were the feature to be removed, normalized to sum to 100.

For the random survival forest, GBM, and super learner models, a grid search method was used to determine hyperparameters (package *MachineShop)*. This method compares all combinations of a range of possible hyperparameter values and chooses the optimal combination based on a performance metric, in this case the C-index, described below. A random survival forest (package *randomForestSRC*) was trained using the selected hyperparameters: node size 300, 10 variables per split, and 500 trees. A GBM (package *gbm*) was trained with a shrinkage parameter of 0.01, interaction depth of 3, minimum node size 100, and 1000 trees. The super learner model was trained using the same hyperparameter values for the random survival forest and GBM and using the *SuperModel* function (package *MachineShop*) to determine, via cross-validation, the optimal weighting of the component models. All four models were restricted to patients with complete data for the predictors used (see Table 1 and Missing Data section below).

Model evaluation

Model performance was evaluated by calculating survival probabilities with each model for observations in the hold-out test set. Concordance and calibration were then calculated using methods adapted for censored data. Concordance was measured using Harrell's C-Index at 1,2, and 5-year follow-up times. The C-Index is a generalization of the common area under the receiver operator characteristic curve (AUC) metric. As with AUC, it ranges from 0 to 1 with 1 indicating perfect concordance. The C-Index measures the proportion of pairs of observations in which predicted ranking of survival probabilities corresponds to actual ranking[14]. Further, the C-index calculation is limited to pairs of patients with sufficient information to determine the true ordering – either both patients must have known times-to-revision, or one has had revision surgery while the other is censored (no revision yet, with time-since-surgery at least as long as the other patient's time-to-revision). For example, a concordance of 0.80 would mean that for a random pair of patients, risk estimates match the true ordering of times-to-revision approximately 80% of the time.

Calibration is a measure of the accuracy of predicted probabilities that compares expected to actual outcomes. We calculated calibration using a version of the Hosmer-Lemeshow statistic that accounts for censoring[47]. This statistic sums average misclassification in each predicted risk quintile and converts the sum into a chi-squared statistic. Larger values of the calibration statistic indicate worse accuracy and correspond to smaller p-values, with statistical significance indicating rejection of the null hypothesis of perfect calibration.

Missing data

Models were trained using observations from the training set with complete data on all variables. The models were then evaluated using observations from the test set with complete data on all variables needed for a given model. To assess the impact of restricting data to complete cases, the models were re-trained and re-evaluated using multiple imputation. This is a common technique for dealing with missing data that fills in incomplete values based on patterns in the data. Multiple imputation allowed the assessment of the reasonableness of restricting the analysis to complete cases. Multivariate imputation by chained equations (MICE) was conducted with five imputations on training and test data (package *mice*). The variables with non-zero coefficients on the complete case Cox lasso were used to re-fit the model with each imputed training data set, averaging predictions over the five imputations. The random survival forest, GBM, and super learner models were similarly refit. A bootstrap procedure was used to compare the calibration statistics between the complete case and imputed models.

RESULTS

Data characteristics

Table 1 describes characteristics of the population at the time of surgery and shows all variables included for analysis. After data cleaning, the combined registry population consisted of 62,955 patients, 55% from the DKLR and 45% from the NKLR. The primary outcome, revision surgery, occurred in 5.1% of patients during an average follow-up time of 7.6 years (SD 4.5). The population was 55% male with a median age at primary injury of 24 years (IQR 18, 34) and median age at surgery of 26 years (IQR 20, 36).

Model performance

The three nonparametric models – random survival forest, GBM, and super learner – had concordance in the moderate range (0.67) at all follow-up times with 95% confidence intervals ranging from (0.64, 0.69) to (0.65, 0.70) (Table 2).

Table 2: Model performance, complete-case training data

Outcome	Model	Concordance	Concordance 95% CI	Calibration statistic	Calibration p-value
1 year	Cox model (lasso)	0.59	(0.56, 0.61)	7.19	0.066
	Random survival forest	0.67	(0.64, 0.69)	5.54	0.136
	Gradient boosted regression	0.67	(0.65, 0.70)	7.48	0.058
	Super learner	0.67	(0.65, 0.69)	8.67	0.034
2 years	Cox model (lasso)	0.58	(0.56, 0.61)	8.17	0.043
	Random survival forest	0.67	(0.64, 0.69)	6.42	0.093
	Gradient boosted regression	0.67	(0.64, 0.69)	4.53	0.210
	Super learner	0.67	(0.64, 0.69)	4.10	0.250
5 years	Cox model (lasso)	0.58	(0.56, 0.61)	11.37	0.010
	Random survival forest	0.67	(0.65, 0.69)	9.27	0.026
	Gradient boosted regression	0.67	(0.64, 0.69)	11.07	0.011
	Super learner	0.67	(0.64, 0.69)	11.82	0.008

The Cox lasso performed more poorly with concordance of 0.58. The Cox lasso showed moderate evidence of mis-calibration (p-value 0.01-0.05) at 2 and 5 years. The other three models were better calibrated, with the exception of the super learner at 1 (p=0.034) and 5 years (p=0.008). The random survival forest and gradient boosted regression models also demonstrated moderate evidence of mis-calibration at 5 years.

Model performance with imputation is presented in Table 3.

Table 3: Model performance, multiply imputed training data

Outcome	Model	Concordance	Concordance 95% CI	Calibration statistic	Calibration p-value
1 year	Cox model (lasso)	0.59	(0.56, 0.61)	8.35	0.039
	Random survival forest	0.66	(0.64, 0.69)	4.17	0.244
	Gradient boosted regression	0.68	(0.65, 0.70)	7.57	0.056
	Super learner	0.67	(0.65, 0.70)	7.99	0.046
2 years	Cox model (lasso)	0.59	(0.56, 0.61)	8.81	0.032
	Random survival forest	0.67	(0.65, 0.70)	8.96	0.030
	Gradient boosted regression	0.67	(0.65, 0.70)	8.98	0.030
	Super learner	0.67	(0.65, 0.70)	8.34	0.039
5 years	Cox model (lasso)	0.58	(0.56, 0.61)	8.30	0.040
	Random survival forest	0.67	(0.65, 0.70)	8.95	0.030
	Gradient boosted regression	0.67	(0.65, 0.69)	11.53	0.009
	Super learner	0.67	(0.65, 0.69)	14.05	0.003

Multiply imputed data did not show notable differences from the complete case analysis. The concordance confidence intervals were nearly identical in all cases. Observed calibration ratios from all four models were compared to the bootstrap distribution, and all the observed ratios were within a 95% confidence interval. This suggests that there is no significant difference in calibration between the complete case and imputed models.

Factors predicting outcome

The most important factors predicting revision surgery, according to the three best-performing models, included age at surgery and injury, years between injury and surgery, graft choice, and pre-surgery KOOS Quality of Life and Sports scores. Variables in approximately the top half by feature importance in the random survival forest, GBM, and super learner models are shown in the bottom three panels of Figure 1. Variables with non-zero coefficients in the Cox lasso model are shown in the top panel of Figure 1. The Cox lasso model quantifies feature importance in terms of absolute value of the associated effect size. The GBM uses difference in error rate were the feature to be removed. The random survival forest and super learner models use permutation-based variable importance, measuring the relative change in model performance upon randomly permuting values of the given variable.

DISCUSSION

Machine learning analysis of the combined NKLR and DKLR enabled the prediction of subsequent revision surgery risk after primary ACLR with moderate accuracy. The most important finding of this study, however, was that this analysis of nearly 63,000 patients yielded similar prediction accuracy as a previous study of approximately 25,000 patients[31,32]. This suggests a so-called ceiling effect of the registries has been reached, and the addition of more patients is unlikely to appreciably improve prediction accuracy. This information can be used to inform further evolution of the national ACL registries regarding variable inclusion and data collection.

Machine learning applications within orthopaedic surgery have been increasing at an exponential rate in recent years[22]. These advanced statistical techniques can evaluate large datasets and realize complex interactions between variables[29]. "Learning" from these interactions, machine learning models can create algorithms capable of predicting outcome for future patients, often at a level of accuracy superior to expert humans[3,8,38,40,41,46,50].

Similar to how humans learn through repetition and experience, machine learning algorithms often require large volumes of data to optimize model accuracy. Data volume, however, is not the only factor contributing to the accuracy of the model. Just as important is the quality of the data. If the dataset used for model creation does not consider variables that are associated with the outcome of interest, then the full potential of the model may not be reached. Poor data quality can also manifest as substantial missing or incomplete data, which impacts the ability of the model to learn and form accurate associations between predictors and outcome. Techniques such as imputation can address some data quality inadequacies, but there are limits to what may be overcome[2].

After nearly 20 years of data collection by the Norwegian and Danish knee ligament registers, the data quantity is superb with satisfactory completeness and data accuracy[7,35–37]. However, the present study suggests that for an improvement in our ability to predict outcome based on the registry data, an evolution in the variables collected is required. This represents a significant challenge as the balance between optimal variable collection and surgeon compliance is a delicate one[11,30]. Data collection must be streamlined to avoid survey fatigue, and the addition of variables to the registry must be carefully considered, weighing the added value against the additional onus on the surgeons which may affect compliance.

Factors that may improve prediction accuracy and could be considered for addition to the national registers include data regarding radiographic findings[4,12,13,18,24,34,48], adjunctive surgical procedures, clinical examination, rehabilitation details[39], and alternative patient reported outcome measures including psychological factors[5]. Pre- and post-operative radiographic indices could be manually captured, for example tibial slope and coronal alignment, or included as raw image files which could then be evaluated through computer vision machine learning techniques[21]. The recording of additional surgical details such as graft diameter/size, ligament augmentation, lateral extra articular tenodesis, or anterolateral ligament reconstruction may also be of value given their recent association with outcome[1,10,15,16,25,27,43,52]. Clinical examination and rehabilitation information such as pre-operative knee laxity grade[26,44] could be obtained through third party sources such as physiotherapists or via natural language processing of patient chart notes[49]. Finally, the KOOS may not reflect the most appropriate patient reported outcome tool for the patient population and an alternative measurement of patient function, such as baseline Marx activity level, could be considered for inclusion in the registries moving forward[19,28].

It is worth mentioning that an algorithm for the prediction of revision surgery after primary ACLR will likely never achieve perfect or even excellent performance in the traditional sense. There are two main reasons for this. First, re-injury events leading to revision surgery may occur randomly such as after a slip on the ice or collision on the playing field. That randomness, combined with the variance related to uncollected variables, limits the predictive capability of ACLR failure models. The second reason is that the outcome, in this case revision surgery, is itself imperfect. That is, not everyone who has experienced a failure will undergo revision surgery. This is a major consideration for most clinical predictive models which are limited by the chosen endpoint. While discrimination has often been interpreted as performance greater than 0.9 being excellent, 0.8-0.9 good, 0.7-0.8 fair, and less than 0.7 poor[45], most clinically useful algorithms demonstrate performance in the 0.65-0.8 range[51]. In fact, discrimination greater than 0.8 for clinical predictive models may represent data mismanagement or model overfitting[20].

Modeling using the combined DKLR and NKLR data revealed some notable differences between the two registries. The poor performance of the Cox lasso model is due in part to the fact that, when modeled separately, the two registry populations led to selection of different variables and different effect sizes for the selected variables. The model fit to the combined data therefore is unable to achieve either of these individually optimal fits and thus performs more poorly. The nonparametric models did not have this limitation since they were able to fit the data with more flexibility. This observation helps explain the fact that although the Cox lasso was the best model in the previous study of the NKLR[32], here the more flexible models performed better.

The present study has some limitations. First, while several machine learning methods were considered, it is possible that another model may have performed differently. Second, there was a high proportion of missing pre-operative KOOS data (47%) and most patients with this missing variable were from the DKLR. Since the pre-operative KOOS data has been important in predicting outcome based on previous studies, this substantial missingness likely contributed to the limited improvement in outcome prediction accuracy. In addition, patients were pooled across the entire time period, from 2004 through 2020. Therefore, this analysis may inherit bias related to temporal changes in revision surgery risk as the surgical indications, techniques, and trends have evolved over time. These changes were not directly accounted for in the present study but likely represent low risk of bias given the stable revision surgery rate observed in the registries.

Regarding clinical limitations of this study, more variables are required for revision prediction using this algorithm than the previously published NKLR calculator which only required the input of five variables. This means the present algorithms are more onerous to use in the office setting with no appreciable improvement in prediction accuracy compared with the

NKLR model. It therefore is likely of limited clinical value unless future external validation demonstrates superiority with different patient populations.

CONCLUSION

Machine learning analysis of the combined Norwegian and Danish knee ligament registers enabled prediction of revision ACLR risk with moderate accuracy. However, the resulting algorithms were less user-friendly and did not demonstrate superior accuracy in comparison to the previously developed model based on patients from the NKLR, despite the analysis of nearly 63,000 patients. This ceiling effect suggests that simply adding more patients to the current national knee ligament registers is unlikely to improve predictive capability and may prompt future changes to increase variable inclusion.

REFERENCES

1. Beckers L, Vivacqua T, Firth AD, Getgood AMJ. Clinical outcomes of contemporary lateral augmentation techniques in primary ACL reconstruction: a systematic review and meta-analysis. *J Exp Orthop*. 2021;8(1):59.

2. Buuren S van, Groothuis-Oudshoorn K. **mice**: Multivariate Imputation by Chained Equations in *R*. *J Stat Softw*. 2011;45(3).

3. Choi JW, Cho YJ, Lee S, et al. Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography. *Invest Radiol*. 2020;55(2):101-110.

4. Christensen JJ, Krych AJ, Engasser WM, Vanhees MK, Collins MS, Dahm DL. Lateral Tibial Posterior Slope Is Increased in Patients With Early Graft Failure After Anterior Cruciate Ligament Reconstruction. *Am J Sports Med*. 2015;43(10):2510-2514.

5. Christino MA, Fleming BC, Machan JT, Shalvoy RM. Psychological Factors Associated With Anterior Cruciate Ligament Reconstruction Recovery. *Orthop J Sports Med*. 2016;4(3):2325967116638341.

6. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.

7. *Dansk Korsbånds Rekonstruktions Register Årsrapport 2020/2021*.

8. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.

9. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367-378.

10. Getgood AMJ, Bryant DM, Litchfield R, et al. Lateral Extra-articular Tenodesis Reduces Failure of Hamstring Tendon Autograft Anterior Cruciate Ligament Reconstruction: 2-Year Outcomes From the STABILITY Study Randomized Clinical Trial. *Am J Sports Med*. 2020;48(2):285-297.

11. Granan LP, Bahr R, Steindal K, Furnes O, Engebretsen L. Development of a national cruciate ligament surgery registry: the Norwegian National Knee Ligament Registry. *Am J Sports Med*. 2008;36(2):308-315.

12. Grassi A, Macchiarola L, Urrizola Barrientos F, et al. Steep Posterior Tibial Slope, Anterior Tibial Subluxation, Deep Posterior Lateral Femoral Condyle, and Meniscal Deficiency Are Common Findings in Multiple Anterior Cruciate Ligament Failures: An MRI Case-Control Study. *Am J Sports Med*. 2019;47(2):285-295.

13. Grassi A, Signorelli C, Urrizola F, et al. Patients With Failed Anterior Cruciate Ligament Reconstruction Have an Increased Posterior Lateral Tibial Plateau Slope: A Case-Controlled Study. *Arthroscopy*. 2019;35(4):1172-1182.

14. Harrell FE. Evaluating the Yield of Medical Tests. *JAMA J Am Med Assoc*. 1982;247(18):2543.

15. Heusdens CHW, Blockhuys K, Roelant E, Dossche L, Van Glabbeek F, Van Dyck P. Suture tape augmentation ACL repair, stable knee, and favorable PROMs, but a re-rupture rate of 11% within 2 years. *Knee Surg Sports Traumatol Arthrosc*. 2021;29(11):3706-3714.

16. Hopper GP, Aithie JMS, Jenkins JM, Wilson WT, Mackay GM. Combined Anterior Cruciate Ligament Repair and Anterolateral Ligament Internal Brace Augmentation: Minimum 2-Year Patient-Reported Outcome Measures. *Orthop J Sports Med*. 2020;8(12):2325967120968557.

17. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860.

18. Jaecker V, Drouven S, Naendrup JH, Kanakamedala AC, Pfeiffer T, Shafizadeh S. Increased medial and lateral tibial posterior slopes are independent risk factors for graft failure following ACL reconstruction. *Arch Orthop Trauma Surg*. 2018;138(10):1423-1431.

19. Kaeding CC, Pedroza AD, Reinke EK, Huston LJ, MOON Consortium, Spindler KP. Risk Factors and Predictors of Subsequent ACL Injury in Either Knee After ACL Reconstruction: Prospective Analysis of 2488 Primary ACL Reconstructions From the MOON Cohort. *Am J Sports Med*. 2015;43(7):1583-1590.

20. Kernbach JM, Staartjes VE. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II-Generalization and Overfitting. *Acta Neurochir Suppl*. 2022;134:15-21.

21. Ko S, Pareek A, Ro DH, et al. Artificial intelligence in orthopedics: three strategies for deep learning with orthopedic specific imaging. *Knee Surg Sports Traumatol Arthrosc.* 2022;30(3):758-761.

22. Kunze KN, Krivicich LM, Clapp IM, et al. Machine Learning Algorithms Predict Achievement of Clinically Significant Outcomes After Orthopaedic Surgery: A Systematic Review. *Arthroscopy*. Published online December 27, 2021:S0749-8063(21)01121-X.

23. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1).

24. Lee CC, Youm YS, Cho SD, et al. Does Posterior Tibial Slope Affect Graft Rupture Following Anterior Cruciate Ligament Reconstruction? *Arthroscopy*. 2018;34(7):2152-2155.

25. Magnussen RA, Lawrence JTR, West RL, Toth AP, Taylor DC, Garrett WE. Graft size and patient age are predictors of early revision after anterior cruciate ligament reconstruction with hamstring autograft. *Arthroscopy*. 2012;28(4):526-531.

26. Magnussen RA, Reinke EK, Huston LJ, et al. Effect of High-Grade Preoperative Knee Laxity on 6-Year Anterior Cruciate Ligament Reconstruction Outcomes. *Am J Sports Med*. 2018;46(12):2865-2872.

27. Mariscalco MW, Flanigan DC, Mitchell J, et al. The influence of hamstring autograft size on patient-reported outcomes and risk of revision after anterior cruciate ligament reconstruction: a Multicenter Orthopaedic Outcomes Network (MOON) Cohort Study. *Arthroscopy.* 2013;29(12):1948-1953.

28. Marmura H, Tremblay PF, Getgood AMJ, Bryant DM. The Knee Injury and Osteoarthritis Outcome Score Does Not Have Adequate Structural Validity for Use with Young, Active Patients with ACL Tears. *Clin Orthop*. Published online March 2, 2022.

29. Martin RK, Ley C, Pareek A, Groll A, Tischer T, Seil R. Artificial intelligence and machine learning: an introduction for orthopaedic surgeons. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(2):361-364.

30. Martin RK, Persson A, Visnes H, Engebretsen L. Registries. In: Musahl V, Karlsson J, Hirschmann MT, et al., eds. *Basic Methods Handbook for Clinical Orthopaedic Research*. Springer Berlin Heidelberg; 2019:359-369.

31. Martin RK, Wastvedt S, Pareek A, et al. Machine learning algorithm to predict anterior cruciate ligament revision demonstrates external validity. *Knee Surg Sports Traumatol Arthrosc*. Published online January 1, 2022.

32. Martin RK, Wastvedt S, Pareek A, et al. Predicting Anterior Cruciate Ligament Reconstruction Revision: A Machine Learning Analysis Utilizing the Norwegian Knee Ligament Register. *J Bone Joint Surg Am*. Published online October 18, 2021.

33. Martin RK, Wastvedt S, Pareek A, et al. Predicting Subjective Failure of ACL Reconstruction: A Machine Learning Analysis of the Norwegian Knee Ligament Register and Patient Reported Outcomes. *J ISAKOS*. Published online January 2022:S2059775422000013.

34. Mehl J, Otto A, Kia C, et al. Osseous valgus alignment and posteromedial ligament complex deficiency lead to increased ACL graft forces. *Knee Surg Sports Traumatol Arthrosc*. 2020;28(4):1119-1129.

35. Midttun E, Andersen MT, Engebretsen L, et al. Good validity in the Norwegian Knee Ligament Register: assessment of data quality for key variables in primary and revision cruciate ligament reconstructions from 2004 to 2013. *BMC Musculoskelet Disord*. 2022;23(1):231.

36. *Norwegian Arthroplasty Register, Norwegian Cruciate Ligament Register, Norwegian Hip Fracture Register, and Norwegian Paediatric Hip Register 2020 Annual Report*. Norwegian National Advisory Unit on Arthroplasty and Hip Fractures; 2020:376.

37. Rahr-Wagner L, Thillemann TM, Lind MC, Pedersen AB. Validation of 14,500 operated knees registered in the Danish Knee Ligament Reconstruction Register: registration completeness and validity of key variables. *Clin Epidemiol*. 2013;5:219-228.

38. Rouzrokh P, Wyles CC, Philbrick KA, et al. A Deep Learning Tool for Automated Radiographic Measurement of Acetabular Component Inclination and Version After Total Hip Arthroplasty. *J Arthroplasty*. Published online February 16, 2021:2510-2517.

39. Samitier G, Marcano AI, Alentorn-Geli E, Cugat R, Farmer KW, Moser MW. Failure of Anterior Cruciate Ligament Reconstruction. *Arch Bone Jt Surg*. 2015;3(4):220-240.

40. Schock J, Truhn D, Abrar DB, et al. Automated Analysis of Alignment in Long-Leg Radiographs by Using a Fully Automated Support System Based on Artificial Intelligence. *Radiol Artif Intell*. 2021;3(2):e200198.

41. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577(7792):706-710.

42. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw*. 2011;39(5).

43. Snaebjörnsson T, Hamrin-Senorski E, Svantesson E, et al. Graft Diameter and Graft Type as Predictors of Anterior Cruciate Ligament Revision: A Cohort Study Including 18,425 Patients from the Swedish and Norwegian National Knee Ligament Registries. *J Bone Joint Surg Am*. 2019;101(20):1812-1820.

44. Sonnery-Cottet B, Saithna A, Cavalier M, et al. Anterolateral Ligament Reconstruction Is Associated With Significantly Reduced ACL Graft Rupture Rates at a Minimum Follow-up of 2 Years: A Prospective Comparative Study of 502 Patients From the SANTI Study Group. *Am J Sports Med*. 2017;45(7):1547-1557.

45. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240(4857):1285-1293.

46. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019;48(2):239-244.

47. Vock DM, Wolfson J, Bandyopadhyay S, et al. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *J Biomed Inform*. 2016;61:119-131.

48. Webb JM, Salmon LJ, Leclerc E, Pinczewski LA, Roe JP. Posterior tibial slope and further anterior cruciate ligament injuries in the anterior cruciate ligament-reconstructed patient. *Am J Sports Med*. 2013;41(12):2800-2804.

49. Wyatt JM, Booth GJ, Goldman AH. Natural Language Processing and Its Use in Orthopaedic Research. *Curr Rev Musculoskelet Med*. 2021;14(6):392-396.

50. Yamada Y, Maki S, Kishida S, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. *Acta Orthop*. 2020;91(6):699-704.

51. Youngstrom EA. A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: we are ready to ROC. *J Pediatr Psychol*. 2014;39(2):204-221.

52. Zhao D, Pan JK, Lin FZ, et al. Risk Factors for Revision or Rerupture After Anterior Cruciate Ligament Reconstruction: A Systematic Review and Meta-analysis. *Am J Sports Med*. Published online October 3, 2022:3635465221119787.

APPENDIX A

Cox Lasso[1]

The Cox Lasso applies Lasso (L1) regularization to the Cox proportional hazards model for regression on right-censored time-to-event outcomes. The method performs variable selection by applying a penalty during model fitting that sets less important predictor coefficients to zero. The remaining (non-zero) coefficients comprise the selected predictors. A tuning parameter controls the extent of this shrinkage: larger values of the tuning parameter correspond to more shrinkage and thus the selection of fewer predictors. We fit the Cox Lasso using the *glmnet* package in R, with the tuning parameter selected via cross-validation to balance model simplicity and fit.

Survival Random Forest[2]

The survival random forest, as implemented in the *randomForestSRC* R package, uses an ensemble tree method designed for right-censored time-to-event data. A log-rank split rule is used, and the estimates associated with each terminal node are computed using the Kaplan-Meier estimator (survival estimate) and the Nelson-Aalen estimator (cumulative hazard estimate). Estimates for an individual are averaged over all bootstrap samples for which the individual is out of bag (OOB). Prediction error for the forest is measured by 1-C, where C is Harrell's concordance index, a measure of accuracy in ranking pairs in terms of their predicted and actual survival.

Gradient boosted regression[3,4]

Gradient boosting uses an iterative method to fit a regression function to the data. At each iteration, the gradient, or the derivative of the loss function with respect to the current regression function, is calculated. The regression function is then updated in the direction of this gradient, improving the fit. Gradient boosted regression as implemented in the R package *gbm*, which we used for our model, uses regression trees as the functions. To accommodate right-censored time-to-event data, the model uses the negative log partial likelihood under the Cox proportional hazards model as the loss function.

Super learner[5]

The super learner is an ensemble method that combines other machine learning models to increase flexibility. The super learner produces a weighted average of its component models by using cross-validation to obtain predictions for each component model, and then training the overall weighted average model to minimize prediction error. The user may specify many different machine learning models as components for the super learner. In this analysis, the super learner combined random survival forest and gradient boosted regression models.

References:

1. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw*. 2011;39(5). doi:10.18637/jss.v039.i05

2. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860. doi:10.1214/08-AOAS169

3. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5). doi:10.1214/aos/1013203451

4. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367-378. doi:10.1016/S0167-9473(01)00065-2

5. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1). doi:10.2202/1544-6115.1309