

R. Kyle Martin

# **Predicting Anterior Cruciate Ligament Reconstruction Outcome**

Machine Learning Analysis of National Knee  
Ligament Registries

DISSERTATION FROM THE NORWEGIAN SCHOOL OF SPORT SCIENCES • 2025

ISBN 978-82-502-0634-2



For my Nan, who instilled in me the importance of empathy and a good bedside manner.

## Table of Contents

<b>Acknowledgements .....</b>	<b>I</b>
<b>List of Papers .....</b>	<b>III</b>
<b>Abbreviations .....</b>	<b>IV</b>
<b>List of Figures .....</b>	<b>V</b>
<b>List of Tables .....</b>	<b>VII</b>
<b>Summary .....</b>	<b>VIII</b>
<b>Introduction .....</b>	<b>11</b>
Background .....	11
The Challenge with Anterior Cruciate Ligament Reconstruction Outcome Prediction .....	12
Machine Learning for Outcome Prediction .....	15
Machine Learning and National Registries to Predict ACL Reconstruction Outcome .....	18
<b>Specific Aims of the Dissertation .....</b>	<b>21</b>
<b>Materials and Methods .....</b>	<b>22</b>
Ethics .....	22
General Comments .....	22
Prediction Model Development (Papers I-III) .....	23
Patient Population .....	23
Data Preparation .....	24
Machine Learning Analysis .....	26
Missing Data .....	28
Model Performance Evaluation .....	29
External Validation (Papers IV and V) .....	30
Patient Population .....	30
Data Preparation .....	31
Missing Data .....	32
Model Performance Evaluation .....	32
Unsupervised Learning (Paper VI) .....	32
Missing Data .....	33
Unsupervised Learning Analysis .....	33
Model Output Evaluation .....	35
<b>Results .....</b>	<b>37</b>
Paper I – Norwegian Revision Risk Prediction .....	37
Paper II – Norwegian Inferior Patient Reported Outcome Risk Prediction .....	40



Paper III – Combined Norwegian and Danish Revision Risk Prediction .....	43
Paper IV – External Validation Using the Danish Registry.....	44
Paper V – External Validation Using the STABILITY I Patients .....	45
Paper VI – Combined Norwegian and Danish Unsupervised Machine Learning Analysis .....	48
<b>Discussion .....</b>	<b>50</b>
Prediction Model Development (Papers I-III) .....	50
Main Findings.....	50
Model Performance.....	52
Outcome Measures.....	55
Factors Associated with Outcome .....	56
Other ACLR Outcome Prediction Models .....	57
External Validation (Papers IV and V).....	59
Main Findings.....	59
Model Performance.....	59
The Effect of Lateral Extra-Articular Tenodesis.....	60
The Importance of External Validation.....	61
Barriers to External Validation.....	61
Unsupervised Learning Analysis (Paper VI).....	62
Main Findings.....	62
The Challenge with Cluster Interpretation – The Black-Box Effect.....	62
Putting it all Together.....	64
Clinical Relevance .....	64
Other Limitations .....	68
<b>Ethical Considerations .....</b>	<b>69</b>
<b>Future Opportunities and Next Steps .....</b>	<b>73</b>
Automated Registry Data Collection.....	73
Prospective External Validation.....	77
Testing ACL Reconstruction Outcome Predictions (TAROT) Study.....	77
<b>Conclusions .....</b>	<b>80</b>
Key Points.....	83
<b>References .....</b>	<b>84</b>
<b>Papers I-VI .....</b>	<b>100</b>



## Acknowledgements

This thesis began in 2019 at the University of Minnesota with initial funding from the Norwegian Arthroplasty & Knee Ligament Register and the University of Oslo School of Medicine. In 2020 we received a Norwegian Centennial Chair seed grant to support the project. The completion of this dissertation was made possible by the collaborative efforts of colleagues from the Norwegian School of Sport Sciences, Oslo Sports Trauma Research Center, University of Oslo, Norwegian Knee Ligament Register, Haukeland University Hospital, Danish Knee Reconstruction Registry, Aarhus University Hospital, University of Western Ontario, Mayo Clinic, Hospital for Special Surgery, and the University of Minnesota.

Many people have inspired and supported me throughout this journey, in particular:

Lars Engebretsen: Quite simply, this work would not have been possible without you. Thank you for taking me as a fellow in 2017 and for introducing me to the power of national registries. You are a legend in the world of orthopaedic surgery and sports medicine, and your clinical and academic mentorship has been invaluable as I progress through the early stages of my career. Your dedication to making everyone around you better is inspiring, and I feel so fortunate to have you as my mentor and friend.

Gilbert Moatshe, my main supervisor and dear friend: I have learned so much from your integrity and your thoughtful approach to research, surgery, leadership, and life. Thank you for motivating me to be a better researcher, clinician, and person and for helping me keep perspective on what is most important.

Roald Bahr: Thank you for your guidance and support with this thesis and in navigating the logistics of completing the PhD program as an international student.

Andreas Persson: Thank you for your friendship, for helping to drive these projects forward, and for sharing the vision of leveraging novel technology to improve the knee ligament registry.

Ayoosh Pareek: What began with beer and wings in 2018 has blossomed into a highly productive collaboration resulting in over 25 AI-related publications, one PhD, and even a memorable engagement in Paris. I look forward to finding out what we can accomplish together over the next seven years (and to reading your PhD thesis soon). Thank you for your visionary leadership, your loyalty, and most of all your friendship.

Solveig Wastvedt: Thank you for applying your expertise to these studies. Your thoughtful approach to machine learning analysis not only propelled our productivity but also significantly

enhanced the clarity and impact of our findings. Your ability to bridge complex technical concepts with practical applications has been invaluable.

Julian Wolfson, my fellow Canadian: I approached you with a concept in 2019 and you made it a reality. Thank you for your enthusiasm regarding our collaboration that has not wavered since we began working together, and for your eagerness to share your expertise through grant writing, statistical oversight, clinical interpretation, and innovative brainstorming sessions. I look forward to continuing our collaboration.

To my co-authors Håvard Visnes, Anne Marie Fenstad, Martin Lind, Hanna Marmura, Dianne Bryant, and Alan Getgood: Thank you for your critical contributions to this trans-Atlantic collaboration. Your time, expertise, and teamwork have been invaluable, and I look forward to collaborating with you all again in the future.

Peter MacDonald: You are the reason that I went into orthopaedic surgery, sports medicine, and academia. Thank you for taking me under your wing and sending me on a path that has led to immense personal and professional satisfaction.

Brad Nelson and Denis Clohisy: Thank you for giving me an opportunity to grow an academic practice in Central Minnesota and for supporting me as I pursue this PhD and other projects.

My friends and family: Thank you for the love and support that has enabled me to pursue my goals and keep balance in my life.

My parents: Thank you for giving me opportunities, morals, work ethic, and the belief that I can do anything I put my mind to.

Claire, my much better half: Thank you for your unwavering patience as you join me on this crazy journey and for the many sacrifices you have made along the way. Thank you also for unlocking my potential as a writer by teaching me about “word vomit” – my life has not been the same ever since. You often ask me what’s next, and even if I don’t know the answer, at least I know we’re in it together.

William and Oliver: Your laughter and curiosity have been a constant reminder of what truly matters. Thank you for grounding me and bringing joy into every day.

R. Kyle Martin

Oslo, 4<sup>th</sup> June 2025

## List of Papers

This dissertation is based on the following papers, which are referred to in the text by their Roman numerals:

- I. Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Engebretsen L. **Predicting Anterior Cruciate Ligament Reconstruction Revision: A Machine Learning Analysis Utilizing the Norwegian Knee Ligament Register.** *J Bone Joint Surg Am.* 2022;104(2):145-153. doi:10.2106/JBJS.21.00113
- II. Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Engebretsen L. **Predicting subjective failure of ACL reconstruction: a machine learning analysis of the Norwegian Knee Ligament Register and patient reported outcomes.** *J ISAKOS.* 2022;7(3):1-9. doi:10.1016/j.jisako.2021.12.005
- III. Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Lind M, Engebretsen L. **Ceiling Effect of the Combined Norwegian and Danish Knee Ligament Registers Limits Anterior Cruciate Ligament Reconstruction Outcome Prediction.** *Am J Sports Med.* 2023;51(9):2324-2332. doi:10.1177/03635465231177905
- IV. Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Lind M, Engebretsen L. **Machine learning algorithm to predict anterior cruciate ligament revision demonstrates external validity.** *Knee Surg Sports Traumatol Arthrosc.* 2022;30(2):368-375. doi:10.1007/s00167-021-06828-w
- V. Martin RK, Marmura H, Wastvedt S, Pareek A, Persson A, Moatshe G, Bryant D, Wolfson J, Engebretsen L, Getgood A. **External validation of the Norwegian anterior cruciate ligament reconstruction revision prediction model using patients from the STABILITY 1 Trial.** *Knee Surg Sports Traumatol Arthrosc.* 2024;32(2):206-213. doi:10.1002/ksa.12031
- VI. Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Lind M, Engebretsen L. **Unsupervised Machine Learning of the Combined Danish and Norwegian Knee Ligament Registers: Identification of 5 Distinct Patient Groups With Differing ACL Revision Rates.** *Am J Sports Med.* 2024;52(4):881-891. doi:10.1177/03635465231225215

## Abbreviations

ACL	Anterior cruciate ligament
ACLR	Anterior cruciate ligament reconstruction
AHC	Agglomerative hierarchical clustering
AUC	Area under the receiver operator characteristic curve
BMI	Body Mass Index
BPTB	Bone-patellar tendon-bone autograft
ChatGPT	Chat Generative Pre-Trained Transformer
CT	Computed tomography
DKRR	Danish Knee Reconstruction Registry
EHR	Electronic health record
FCL	Fibular collateral ligament
GAM	Generalized additive model
GBM	Gradient boosted regression
HT	Hamstring tendon autograft
IKDC	International Knee Documentation Committee
IRB	Institutional review board
k	Number of clusters
KOOS	Knee Injury and Osteoarthritis Outcome Score
LET	Lateral extra-articular tenodesis
MCID	Minimal Clinically Important Difference
MCL	Medial collateral ligament
MICE	Multiple imputation by chained equations
MOON	Multicenter Orthopaedic Outcomes Network
MRI	Magnetic resonance imaging
NKLR	Norwegian Knee Ligament Register
PASS	Patient acceptable symptom state
PCL	Posterior cruciate ligament
PLC	Posterolateral corner
PROM	Patient reported outcome measures
QoL	Quality of Life subscale
QT	Quadriceps tendon autograft
RCT	Randomized controlled trial
SHAP	SHapley Additive exPlanations
TAROT	Testing ACL Reconstruction Outcome PredicTions
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

## List of Figures

Figure 1: Artificial intelligence and its subsets machine learning and deep learning. Reproduced with permission from Pareek A, Ro DH, Karlsson J, Martin RK. Machine learning/artificial intelligence in sports medicine: state of the art and future directions. *J ISAKOS*. 2024;9(4):635-644. doi:10.1016/j.jisako.2024.01.013<sup>1</sup>

Figure 2: The three main branches of machine learning – supervised, unsupervised, and reinforcement learning. Reproduced with permission from Pruneski JA, Williams RJ, Nwachukwu BU, et al. The development and deployment of machine learning models. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(12):3917-3923. doi:10.1007/s00167-022-07155-4<sup>2</sup>.

Figure 3: Elbow method to determine the optimal number of clusters. The circled point is the “elbow” at which further increases in the number of clusters no longer significantly reduces the within-cluster variance.

Figure 4: Calibration plots for the revision prediction models at each time point.

CL: cox lasso; RF: survival random forest; GAM: generalized additive model; GBM: gradient boosted regression.

Figure 5: QR Code for revision risk calculator

Figure 6: QR Code for inferior patient reported outcome risk calculator

Figure 7: Mean absolute SHapley Additive exPlanations (SHAP) values by variable for each cluster. Colours represent the contributions of the variables assigned to each cluster.

BPTB: bone–patellar tendon–bone autograft; comb: combined; fix.: fixation; ICRS: International Cartilage Regeneration & Joint Preservation Society; KOOS: Knee injury and Osteoarthritis Outcome Score; QOL: Quality of Life subscale; QT/BQT: quadriceps tendon autograft (with or without bone); Sports: Sport and Recreation subscale.

Figure 8: Kaplan-Meier survival curve for all 5 clusters.

Figure 9: Tree diagram for approximate patient classification by cluster.

BPTB: bone–patellar tendon–bone autograft; KOOS: Knee injury and Osteoarthritis Outcome Score (Sports subscale); QT: quadriceps tendon autograft (with or without bone).

Figure 10: Example output of the online revision risk calculator. The patient is 20 years old with a pre-operative KOOS QoL score of 50 undergoing an ACL reconstruction with hamstring tendon autograft and suspension fixation on the femur four months after ACL injury. Patient-specific risk estimates are shown on the right, along with the median level of risk with 25<sup>th</sup> to 75<sup>th</sup> percentiles based on the Norwegian Knee Ligament Register patient population.

KOOS: Knee Injury and Osteoarthritis Outcome Score; QoL: Quality of Life subscale; ACL: anterior cruciate ligament

Figure 11: Example output of the online calculator to predict two-year post-operative KOOS QoL score less than 44 after ACL reconstruction. The patient is 20 years old, with a BMI of 23, and pre-operative KOOS score of 50 on all subscales. There is no history of previous ipsilateral knee surgery, no concomitant cartilage injury, and the injury occurred during a pivoting activity. Patient-specific risk estimate is shown on the right, along with the median level of risk with 25<sup>th</sup> to 75<sup>th</sup> percentiles based on the Norwegian Knee Ligament Register patient population.

KOOS: Knee Injury and Osteoarthritis Outcome Score; QoL: Quality of Life subscale; ACL: anterior cruciate ligament; BMI: Body Mass Index

Figure 12: Clinical relevance pyramid for clinical prediction modelling

Figure 13: Testing ACL Reconstruction Outcome Predictions (TAROT) Study



## List of Tables

Table 1: List of variables from the Norwegian and Danish Knee Ligament Registries considered for machine learning analysis during prediction model development.

Table 2: Model Performance Measures – Norwegian Revision Risk Analysis

Table 3: Model Performance Measures – Norwegian Inferior Patient Reported Outcome Analysis

Table 4: Model Performance Measures – Combined Registry Revision Risk Analysis

Table 5: Comparison of model performance of the revision risk algorithm between the original Norwegian internal validation and the Danish external validation cohorts.

Table 6: Comparison of model performance of the revision risk algorithm between the original Norwegian internal validation and the STABILITY I RCT external validation cohorts with patients randomized to hamstring tendon autograft plus lateral extra-articular tenodesis coded three different ways.

## Summary

### Introduction:

Anterior cruciate ligament (ACL) injuries are common, and surgery is often performed to improve function. Many factors have been identified that may influence the risk of a poor outcome following ACL reconstruction (ACLR). However, putting those risk factors into context and applying them to an individual patient to accurately estimate their specific risk of a poor outcome is challenging. The ability to accurately quantify risk at a patient-specific level is desirable as it can lead to more informed discussions and surgical decision-making, and can guide efforts at decreasing risk.

Machine learning is a branch of artificial intelligence that enables the development of algorithms capable of predicting clinical outcomes based on analysis of large databases. These novel techniques can tease out relationships between variables that may be more complex than can be realized through traditional statistical analyses. The purpose of this thesis was to apply machine learning analysis to the Norwegian Knee Ligament Register (NKLK) and Danish Knee Reconstruction Registry (DKRR) to develop easy-to-use models capable of predicting post-operative outcomes (revision surgery and inferior patient reported outcome) for patients undergoing ACLR and identify the factors that are most important for making the outcome predictions. The hypothesis was that this analysis would lead to the development of accurate and externally valid clinical prediction tools that clinicians could use to predict the risk of revision surgery or inferior patient reported outcome for their patients undergoing ACLR.

### Methods:

Four methods of supervised machine learning were performed on the NKLK data that had first been split into training and test sets. In Paper I, models were trained to predict revision surgery and in Paper II the models were trained to predict inferior patient reported outcome, defined as a score of less than 44 on the Knee Injury and Osteoarthritis Outcome Score (KOOS) Quality of Life subscale (QoL). In Paper III, the DKRR data was merged with data from the NKLK and four methods of supervised machine learning were performed on the combined dataset with the goal of improving the accuracy of the revision prediction model through the addition of more data. The revision prediction model developed using the NKLK-only data in Paper I was subsequently tested on the DKRR data (Paper IV) and on patients from the STABILITY I randomized controlled trial (Paper V) to determine external validity of the model on different

patient groups. Performance of the supervised machine learning models was evaluated using measures of discrimination and calibration in all cases.

In Paper VI, unsupervised machine learning was performed on the combined NKLR and DKRR dataset to generate distinct clusters of patients with similar intra-cluster characteristics. The optimal number of clusters was determined via a combination of data-driven and domain knowledge assessment. The clusters were then interpreted through the aid of SHapley Additive exPlanations (SHAP) analysis to determine the distinguishing characteristics of each cluster and enable future patients to be assigned to the most appropriate cluster. Revision rates of each cluster were then evaluated to determine if clusters had varying risks of revision surgery following primary ACLR.

**Main results:**

Supervised machine learning analysis of the NKLR produced prediction models with area under the receiver operator characteristic curve (AUC) and concordance (two measures of discrimination) values of 0.67-0.69. The models were generally well-calibrated, with modest evidence of mis-calibration only for the two-year prediction of revision risk. Factors required for revision surgery prediction included: graft choice, femoral fixation device, pre-operative KOOS QoL score, time between the injury and surgery, and age at the time of surgery. Factors required for prediction of inferior patient reported outcome two-years after ACLR were: pre-operative KOOS subscale scores, grade of cartilage injury, activity leading to injury, previous ipsilateral knee surgery, Body Mass Index (BMI) at surgery, and age at injury. Both algorithms were converted into easy-to-use online calculators.

Accuracy of the revision prediction model did not improve when the DKRR data was merged with the NKLR data. The discrimination performance of the revision model did not change when it was evaluated using the DKRR patients, while the calibration worsened for the one-year and five-year predictions. When the revision model was evaluated on the STABILITY I cohort, the model performed best, with discrimination and calibration similar to the original model testing, when the addition of a lateral extra-articular tenodesis (LET) to a hamstring tendon autograft (HT) ACLR was entered into the algorithm as a bone-patellar tendon-bone autograft (BPTB). However, the discrimination value confidence interval was wide.

Five clusters were found to be optimal and were subsequently created through k-prototypes unsupervised machine learning analysis. Each cluster demonstrated a unique revision rate and the clusters were divided into high-risk (Cluster 1, revision rate: 9.9%), medium-risk (Cluster 2, revision rate: 6.9%), and low-risk (Clusters 3-5, revision rate: 3.1-4.7%) groups. A tree diagram was created to facilitate rapid risk stratification based on three variables: age, graft choice, and KOOS Sports subscale score.

**Conclusion:**

The most significant findings from these studies are: 1) machine learning analysis of the NKLR and DKRR enabled the development and validation of prediction models that demonstrated moderate accuracy for predicting revision surgery and inferior outcome following ACLR and identified the most important factors used to predict these outcomes, 2) a rigorous approach to clinical prediction modeling has been described, laying the foundation for future innovation, 3) more work is needed to evaluate the performance of the prediction models on patients from outside Scandinavia and to determine the threshold for clinical relevance regarding ACLR outcome prediction, 4) the development and validation of clinical prediction tools may be limited by both the quality and quantity of the available data, and 5) the data collected by the registries should be expanded to include more variables that have been associated with outcome.

Although these studies enabled the development of several risk estimation tools for patients undergoing ACLR, the performance of these models was limited by the data contained within the registries. More specifically, they were limited by the lack of some important relevant variables associated with outcome such as pre-operative knee laxity, posterior tibial slope, and rehabilitation factors. The choice of outcomes (revision surgery and low KOOS scores) may have also limited the model performance. In addition, external validation outside of Scandinavia was limited by poor data quantity in the STABILITY I cohort. Evolution of the national knee ligament registries to capture more variables is required to improve the ability to predict outcome using these databases. Overall, the processes outlined in these studies can serve as a guide for the pursuit of clinical prediction models in the future; however, the current clinical utility of the ACLR prediction models remains unknown. Prior to widespread adoption and implementation of these prediction algorithms, their performance relative to predictions made by surgeons must be ascertained. This represents an important next step because until it is known how well surgeons can predict outcome, it will never be known if prediction tools driven by artificial intelligence confer an advantage and, therefore, are clinically relevant.

## Introduction

### Background

Physicians strive to optimize the outcome for each individual patient they care for. To accomplish this, the physician evaluates the information available to them to first identify the most likely diagnosis and then determine the most appropriate course of action, which may include further investigation or the initiation of a treatment. The data used to inform these decisions often come from a wide variety of sources which may include the patient history, physical examination, and available imaging or laboratory studies. In addition, physicians must simultaneously consider any potential barriers or risk factors that may impact the eventual outcome and strategize ways to minimize these challenges.

In essence, physicians develop an algorithmic approach to patient care based on pattern recognition that is influenced, and limited, by their own experience and understanding of the best available evidence. Although these clinical decisions may occasionally be relatively simple and straightforward, there are often nuances that require a more individualized approach. In orthopaedic surgery, anterior cruciate ligament (ACL) ruptures represent one condition that demands such an approach in the pursuit of optimal outcome.

The ACL is a central stabilizer of the knee that is important for maintaining normal knee biomechanics and function. Injuries to the ACL are common and can lead to persistent pain, instability, and significantly increased risk of post-traumatic osteoarthritis<sup>3,4</sup>. Surgical reconstruction of the ACL (ACLR) is often performed to restore knee function and stability. In the United States, more than 120,000 ACL reconstructions are performed every year in young athletes and rates have been rising<sup>5-9</sup>. This observed increase has been tied to widespread growth in sports participation and specialization and is expected to continue. With increased ACL injuries, the associated societal costs are also rising. Cost-utility analyses have reported lifetime costs of greater than \$30,000 for those who have ACLR and greater than \$80,000 for those treated non-operatively<sup>10</sup>.

Anterior cruciate ligament injuries primarily affect young, active persons who have a desire to return to an active lifestyle or sports participation and are often entering or are in the early stages of their careers. Unfortunately, despite many advancements in surgical and rehabilitation techniques over time, the failure rate of ACLR remains a concern – the risk of a patient experiencing a second ACL injury (ipsilateral or contralateral) following ACLR has been reported to be between 8-35%<sup>11-13</sup>. Further, outcomes following revision ACLR have been found to be inferior when compared with primary ACLR<sup>14-18</sup>. Due to the high prevalence and potential morbidity of these injuries, the study of ACLR outcomes and the risk factors associated with inferior outcomes has garnered substantial attention in the literature as surgeons strive to optimize results for their patients<sup>19</sup>. However, individual outcome optimization remains limited due to the inability to accurately predict the expected outcome of treatment strategies.

## **The Challenge with Anterior Cruciate Ligament Reconstruction Outcome Prediction**

Accurate prognostication for patients undergoing ACLR is a crucial component of individualized outcome optimization, enabling the identification of patients who are at an increased risk of experiencing failure. This information can be applied in a clinical setting to guide discussions with patients, align expectations with reality, and may influence surgical decision-making and post-surgical care. From a research perspective, risk quantification also enables the evaluation of targeted strategies aimed at reducing risk overall and particularly among those deemed high-risk.

Several intrinsic and extrinsic factors have been identified that place individuals at risk for sustaining an ACL injury<sup>20</sup>. These include female sex, hormones, landing mechanism, ligamentous laxity, anatomical variation, and neuromuscular control during activity<sup>20</sup>. Additionally, the identification of risk factors associated with failure of ACLR has been the focus of multiple studies in the orthopaedic literature<sup>21</sup>.

Young age and graft-related factors are some of the most consistently reported risk factors associated with ACLR graft failure<sup>21</sup>. Analysis of data from the Norwegian, Swedish, and Danish

knee ligament registries have identified young age, high body mass index (BMI), graft diameter, and graft type as risk factors associated with failure of ACLR, defined as revision surgery<sup>22,23</sup>. The finding that young age and small graft size increase the risk of ACLR revision was also reported by Magnussen et al<sup>24</sup>. Similarly, the Multicenter Orthopaedic Outcomes Network (MOON) group reported that young age, high activity level, and the use of allograft were associated with higher odds of a graft failure<sup>11</sup>.

The association between age and graft failure after ACLR was explored further by Grindem et al<sup>25</sup>. They found that when adjusting for return to pivoting sports within one year after surgery and performance on return to sport functional criteria, age was no longer an independent risk factor. The authors opined that adequate rehabilitation and return to sport are confounders that can explain the often-cited association between age and ACLR failure. In a previous study, Grindem et al. also highlighted the importance of completing a full rehabilitation protocol prior to return to sport<sup>26</sup>. They reported an ACLR failure rate of 38.2% among patients who did not pass both time-based and functional return to sport criteria compared with only 5.6% who did. The importance of adequate rehabilitation and functional readiness for return to sport was also identified by Kyritsis et al., finding a four-times higher graft failure rate among athletes who did not complete their return to sport functional criteria<sup>27</sup>. Overall, post-operative rehabilitation seems to significantly impact the outcome of ACLR.

Several other studies from the national knee ligament registries in Scandinavia have evaluated the factors associated with subsequent ACLR revision. In their review of patients from the Norwegian Knee Ligament Register (NKLR), Persson et al. found that the use of hamstring tendon autograft (HT) was associated with higher revision surgery rates than bone-patellar tendon-bone autograft (BPTB)<sup>28</sup>, and that revision rates were highest when the combination of suspension fixation on the femur and an absorbable interference screw on the tibia were used<sup>29</sup>. Subsequent analysis of the combined Scandinavian knee ligament registry data revealed similar findings<sup>30,31</sup>. Interestingly, following the publication of these studies practice patterns changed substantially in Norway. Prior to 2013, HT was used in 73% of all primary ACLR recorded in the NKLR<sup>28</sup>. In contrast, the use of HT dropped to 33% in 2016, with BPTB accounting for 63% of all primary ACLR that year<sup>32</sup>.

More recently, an increased posterior tibial slope has been suggested as a possible risk factor for ACL injury and ACLR failure. Hashemi et al. found that posterior tibial slope, in particular involving the lateral side, was increased in patients who sustained an ACL injury when compared with uninjured controls<sup>33</sup>. Jaecker et al. also reported increased posterior tibial slope both medially and laterally was observed to be an independent risk factor for ACLR graft failure<sup>34</sup>. A proposed cutoff of 12° has been made for the posterior tibial slope, above which the risk of ACLR graft rupture may be significantly increased, and this finding may be even more pronounced in adolescents<sup>35,36</sup>. The influence of the posterior tibial slope was summarized in a recent systematic review and meta-analysis by Duerr et al<sup>37</sup>. They aggregated 15 studies that compared posterior tibial slope in those with and without an ACL graft failure and found that the posterior tibial slope was increased significantly in the failure groups. Despite a growing number of studies that have suggested this association between increased tibial slope and ACLR failure, a few studies have challenged this conclusion and the true impact of posterior tibial slope on ACLR outcome remains uncertain<sup>38–40</sup>.

In addition to the posterior tibial slope, other anatomic risk factors for ACL injury and failure of primary ACLR have been suggested. A narrow femoral intercondylar notch width has been identified as a risk factor for ACL injury<sup>41–43</sup>. Hughes et al. have subsequently reported a five-times higher ACL graft failure rate in patients with a narrow intercondylar notch, which they defined as a width less than 16 mm<sup>44</sup>. Increased knee hyperextension, defined as greater than 5° of passive knee hyperextension on the contralateral side, was also identified as a risk factor for graft failure after ACLR by Guimarães et al. in their study<sup>45</sup>.

The multitude of risk factors for ACLR revision or re-rupture were recently summarized in a systematic review and meta-analysis by Zhao et al<sup>21</sup>. The authors found that these factors include male sex, young age, low BMI, a family history of revision or failure of ACLR, white race, increased tibial slope, high-grade pre-operative knee laxity, higher baseline Marx activity level, return to a high activity level or sport, surgery performed less than one year from injury, the presence of a concomitant medial collateral ligament (MCL) injury, the use of an anteromedial or transportal technique, HT or allograft, and a smaller graft diameter. However, the ability to



synthesize this information and accurately quantify a patient's risk of experiencing subsequent graft failure leading to revision surgery and other inferior outcomes following ACLR remains challenging and imprecise due to the complex interactions between these many factors that influence and contribute to an individual's outcome.

## **Machine Learning for Outcome Prediction**

Artificial intelligence, and in particular machine learning, has been identified as a potential solution for the problem of outcome prediction. Over the past several years, artificial intelligence applications have become ubiquitous throughout society. Self-driving cars are being tested on roads throughout the world. Voice recognition software has become commonplace on our phones and in our homes. Generative text programs have passed medical and legal licensure examinations. Targeted content and advertisements are now a routine expectation on social media and web browsers. However, despite these rapid advancements in our day-to-day lives, there is a paucity of clinically relevant artificial intelligence applications within the field of orthopaedics and sports medicine.

The widespread capabilities of artificial intelligence and machine learning have recently become more appreciated in orthopaedics and sports medicine. This is the age of big data, and a promising feature of machine learning is its capacity to use historical clinical data to inform future care delivery<sup>46</sup>. One of the most important areas in which these innovations can impact the field is related to the use of machine learning to facilitate outcome prediction. Predictive models driven by machine learning have the ability to provide synthesized data, validated predictions, and the basis for clinical decision-making to health care providers.

Broadly speaking, the term “artificial intelligence” refers to any technique in which machines are said to mimic human behaviour<sup>1</sup>. This often involves automation with minimal human programming. Within artificial intelligence is the subset known as “machine learning,” whereby a statistical model or algorithm is developed that can “learn” through experience and apply that knowledge to new or future data. A further subset within machine learning is called “deep learning,” which enables more complex pattern recognition and automation using neural networks (Figure 1).

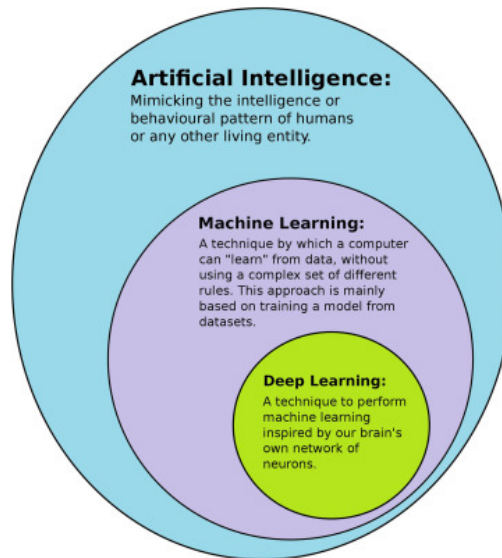


Figure 1: Artificial intelligence and its subsets machine learning and deep learning. Reproduced with permission from Pareek A, Ro DH, Karlsson J, Martin RK. Machine learning/artificial intelligence in sports medicine: state of the art and future directions. J ISAKOS. 2024;9(4):635-644. doi:10.1016/j.jisako.2024.01.013<sup>1</sup>.

Although the concept of machine learning has been around since the mid 20<sup>th</sup> century<sup>47–49</sup>, it is only recently that improved computational power and data availability has enabled the advancement in algorithm creation and applications throughout society<sup>1,50,51</sup>. These algorithms can be trained to identify patterns in a dataset and make predictions through several different “learning” techniques. The three main branches of machine learning are supervised learning, unsupervised learning, and reinforcement learning (Figure 2)<sup>2</sup>.

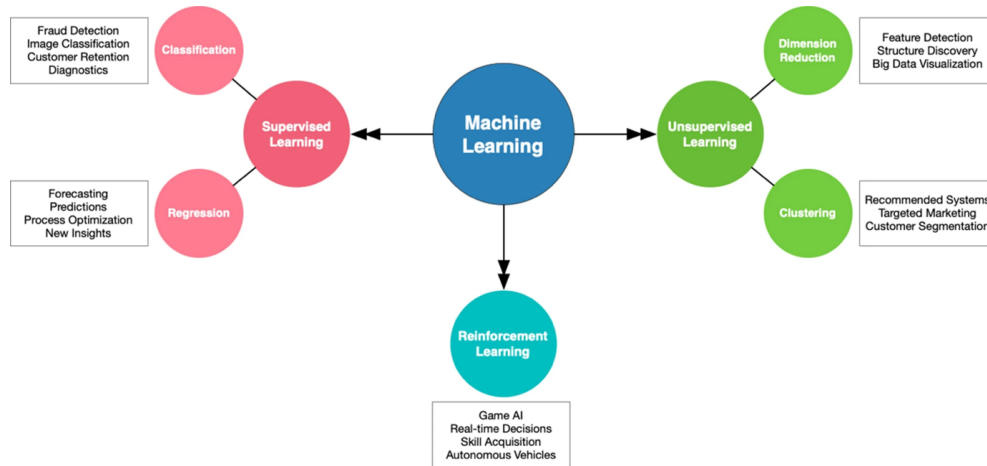


Figure 2: The three main branches of machine learning – supervised, unsupervised, and reinforcement learning. Reproduced with permission from Pruneski JA, Williams RJ, Nwachukwu BU, et al. The development and deployment of machine learning models. *Knee Surg Sports Traumatol Arthrosc.* 2022;30(12):3917-3923. doi:10.1007/s00167-022-07155-4<sup>2</sup>.

Supervised learning can be thought of as an algorithm that learns from labeled examples<sup>52</sup>. The model is provided with all the variables in a dataset, which are labeled as either "predictors" (input features) or "outcomes" (target variables). By training on the examples in the dataset, the model learns to recognize the relationships between predictor variables and the outcome of interest. The goal is to generalize this knowledge to make accurate predictions on new, unseen data. In supervised learning, classification is used to predict discrete or categorical outcomes, while regression is used to predict continuous numeric values.

In contrast, unsupervised learning is a type of machine learning where the model is given data without labeled outcomes or target variables<sup>53</sup>. Unlike supervised learning, which relies on known "predictors" and "outcomes," unsupervised learning has no predefined labels to guide the model. Instead, the model analyzes the data to identify hidden patterns or groupings within it. This is particularly useful for exploratory analysis or finding natural clusters in data, such as groups of patients that share similar characteristics. Two common applications of unsupervised learning are clustering, where the model groups similar data points together, and dimensionality reduction, which simplifies data by reducing the number of features while retaining important information. Although unsupervised learning is not applied to predict specific outcomes, the approach may be

used to discover structure and relationships within unlabeled data and, when applied to health care data, may identify groups or clusters of patients with differing outcome or risk profiles.

Reinforcement learning is the third type of machine learning where an algorithm learns to make decisions by interacting with an environment<sup>1,54,55</sup>. Unlike supervised learning, which learns from labeled data, or unsupervised learning, which finds patterns in unlabeled data, reinforcement learning is driven by a system of rewards and penalties. The algorithm takes actions in the environment, receives feedback in the form of rewards (positive feedback) or penalties (negative feedback), and uses this feedback to improve its future actions. The goal of reinforcement learning is for the model to maximize cumulative reward over time, developing an optimal strategy or policy for the task, such as playing a game, managing resources, or controlling a robot.

In recent years, machine learning has driven significant breakthroughs in medical outcome prediction. Models trained on large datasets of medical records and imaging data have enabled highly accurate predictions in diverse areas. Models have been developed that can detect early signs of cancer with accuracy on par with or better than human radiologists<sup>56,57</sup>, and to help predict in-hospital mortality due to sepsis<sup>58</sup>. Similarly, predictive models are also being used to identify patients at high risk of developing chronic conditions like heart disease, enabling proactive intervention<sup>59</sup>. With these and many other advancements has come a lot of enthusiasm surrounding the possibilities of machine learning within orthopaedic surgery. However, clinically useful orthopaedic applications of the approach have lagged behind the other medical specialties<sup>60,61</sup>.

## **Machine Learning and National Registries to Predict ACL Reconstruction Outcome**

One of the barriers to the successful implementation of machine learning into healthcare is the scarcity of relevant large data repositories<sup>62</sup>. National knee ligament registries therefore present an opportunity for exploration given their comprehensive data collection and labelling systems. The NKLRL was founded in 2004, representing the world's first national registry focused on knee ligament surgery<sup>63</sup>. The Danish Knee Reconstruction Registry (DKRR) followed in 2005, and

both registries have been prospectively collecting data regarding cruciate ligament surgeries since their inception<sup>64</sup>.

The knee ligament registries record patient demographic and injury details, information regarding the surgical findings and procedures performed, and most importantly, include long-term tracking of outcome<sup>65</sup>. Two primary outcomes are recorded in the registries – subsequent revision surgery and patient reported outcome measures (PROM). Patients in the registries are linked using their unique personal health identification number which enables the detection of any subsequent revision surgery, even if it is performed in another centre or with another surgeon, provided it occurs within the same country. Attrition due to death or emigration can also be measured using the unique patient identification numbers. The Knee Injury and Osteoarthritis Outcome Score (KOOS)<sup>66</sup> is a PROM that is obtained before surgery and at pre-specified timepoints following ACLR. In Norway, the KOOS is collected two, five, and ten years after ACLR, while in Denmark patients complete the questionnaire one, two, and ten years post-operatively<sup>63,64,67,68</sup>. As a PROM, the KOOS has been validated for measuring knee function in patients with osteoarthritis and for other knee conditions, including ACL, chondral, and meniscal injuries<sup>66,69</sup>.

Reporting to the registries is mandatory for surgeons, and data collection has demonstrated high levels of completeness and validity<sup>64,70,71</sup>. Registry data have been instrumental in guiding the management of ACL rupture through the publications of several previous studies that have enhanced understanding of ACL injuries, surgical techniques, and ACLR outcome<sup>22,28–31,68,72–75</sup>. These studies have identified that some factors such as patient age, along with graft choice and fixation technique, have been associated with variable and quantifiable rates of failure. However, it has remained unclear how these factors translate to the clinic in a patient-specific manner.

Machine learning predictive models have the ability to consider all patient, injury, and surgical factors while generating not only a patient-specific likelihood of a poor outcome, but also the magnitude of effect for each variable. Additionally, machine learning algorithms using feature selection can narrow all variables down to those of highest importance while maintaining accuracy in the predictive capabilities. This is in contrast to non-machine learning analyses whose

validity relies on researchers' underlying assumptions to correctly specify which variables to include, how they interact, and their functional relationship to the outcome. Therefore, the results of machine learning analysis can both assess variable interactions within a model and place importance onto these variables in quantifiable terms, allowing more accurate predictions to be obtained.

This background establishes the basis for the hypothesis of this thesis: that machine learning can be applied to national knee ligament registry data in Norway and Denmark to enable the prediction of outcome following ACLR and to identify the most important factors used to predict these outcomes.

## Specific Aims of the Dissertation

The overall objective of this thesis is to apply machine learning to the NKLR and DKRR to create and validate machine learning algorithms capable of predicting outcome following ACLR with particular emphasis on ease of use and clinical applicability.

The specific aims of the six papers were:

1. To identify the most important risk factors associated with subsequent revision following primary ACLR using supervised machine learning analysis of the NKLR (Paper I)
2. To develop a clinically useful prediction model to estimate patient-specific risk of subsequent revision following primary ACLR using supervised machine learning analysis of the NKLR (Paper I)
3. To identify the most important risk factors associated with inferior patient reported outcome following primary ACLR using supervised machine learning analysis of the NKLR (Paper II)
4. To develop a clinically useful prediction model to estimate patient-specific risk of inferior patient reported outcome following primary ACLR using supervised machine learning analysis of the NKLR (Paper II)
5. To improve the accuracy of the revision prediction model through amalgamation of the NKLR and DKRR databases (Paper III)
6. To evaluate the external validity of the NKLR revision prediction model when applied to patients from the DKRR (Paper IV)
7. To evaluate the external validity of the NKLR revision prediction model when applied to patients from the STABILITY I randomized clinical trial (Paper V)
8. To identify distinct subgroups (clusters) of patients within the NKLR and DKRR with similar characteristics using an unsupervised learning technique, and determine how the rate of subsequent revision ACLR differs between them (Paper VI)
9. To develop a clinically relevant rapid risk-stratification algorithm based on the unsupervised learning clusters (Paper VI)

## Materials and Methods

### Ethics

All patients provide informed consent at the time of enrollment in the NKLR while informed consent is not required for the DKRR. All data are stored securely, and only de-identified information is made available for research. For this reason, the Data Inspectorate and Regional Ethics Committee in Norway and the General Data Protection Regulation in Denmark do not require additional ethical review board evaluation or approval for studies utilizing the registry data. Institutional review board (IRB) evaluation at the University of Minnesota similarly determined that the studies comprising this thesis were exempt from IRB review. For Papers I-IV and VI, deidentified data from the NKLR and DKRR were transferred to the University of Minnesota research team for machine learning analysis and external validation. For Paper V, the revision prediction algorithm was shared with the Fowler Kennedy Sports Medicine Clinic research team at the University of Western Ontario for external validation using patients from the STABILITY I randomized controlled trial (RCT). The STABILITY 1 trial was previously approved by the Western Ontario Health Sciences Research Ethics Board (#104524) and no further IRB approval was necessary<sup>76</sup>.

### General Comments

When planning the development and deployment of clinical outcome prediction models it is important to have a rigorous and methodical approach. This thesis reflects a stepwise approach that was taken intentionally, to instill confidence in the findings while establishing a framework for future projects of a similar nature. The studies were carried out in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement<sup>77</sup>. This comprehensive set of recommendations for prediction model development or validation studies aims to improve the transparency of these studies through full and clear information reporting, independent of study methods.

The first step involved the initial model development with internal validation using the NKLR (Papers I and II). Next, additional data (from the DKRR) was combined with the NKLR in an



attempt to improve the accuracy of the prediction model (Paper III). Once a working model was created, the third step was to assess the external validity of the model when applied to patients that were independent from the original patient population used to develop the algorithm (Papers IV and V). Finally, unsupervised learning was applied to the combined registry data to develop a risk stratification tool using a novel approach (Paper VI).

Three distinct study methods were employed: prediction model development using supervised machine learning (Papers I-III), external validation (Papers IV and V), and unsupervised machine learning analysis (Paper VI). What follows is a discussion regarding the methodology that is grouped accordingly.

### **Prediction Model Development (Papers I-III)**

The first three studies share a common goal of developing predictive models using supervised machine learning to forecast outcomes for patients undergoing ACLR. All studies utilized national knee ligament registry data, with the first two relying solely on the NKLR, while the third combined data from both the NKLR and the DKRR. The primary outcome in the first and third studies was the risk of subsequent revision ACL surgery, whereas the second study focused on predicting inferior PROMs using the KOOS Quality of Life subscale (QoL). This summary will explore the methods employed across the studies, with a focus on the patient populations, data preparation, machine learning model selection, missing data, and model performance evaluation techniques.

### **Patient Population**

For Papers I and II, all patients from the NKLR who underwent primary ACLR between January 2004 and December 2018 were included. Paper III occurred later and subsequently extended the timeframe to December 2020 while combining data from the NKLR and DKRR to increase sample size. Patients with missing or incomplete outcome data were excluded, including those with missing revision status (Papers I and III) or missing KOOS scores (Paper II).

### Data Preparation

The first step in a machine learning approach involves preparing the data for analysis. In supervised machine learning, this means variables must be defined as either “predictor” or “outcome” variables. All three studies considered a wide range of predictors related to patient demographics, injury characteristics, and surgical details that were available in the registries. In total, 24 predictor variables were considered for analysis in Paper I, 19 were considered for Paper II, and 17 were considered for Paper III. In addition to variables extracted directly from the registries, this also included composite indicators such as whether or not a patient was below the median KOOS in all five subscales (Table 1). The outcome variable for Papers I and III was revision surgery, while the outcome variable for Paper II was a KOOS QoL score <44, which has previously been defined as subjective failure<sup>73,78</sup>.

Table 1: List of variables from the Norwegian and Danish Knee Ligament Registries considered for machine learning analysis during prediction model development.

<b><u>Predictor Variables Considered for Machine Learning Analysis</u></b>	
Age at injury	Presence of PCL injury†
Age at surgery	Graft choice
Sex	Tibial fixation device
Body Mass Index (kg/m <sup>2</sup> )‡	Femoral fixation device
Pre-operative KOOS QoL subscale score	Fixation device combination*
Pre-operative KOOS Sports subscale score	Injured side
Below median on all pre-operative KOOS subscale scores*	History of surgery to contralateral knee
Activity leading to injury	History of surgery to ipsilateral knee
Presence of meniscus tear	Time from injury to surgery
Presence and grade of cartilage injury	Pre-operative systemic antibiotics‡
Presence of MCL injury†	Hospital geographic region†
Presence of FCL or PLC injury†	Hospital type (public or private)†

\* Composite measure

† Not considered for analysis in Paper II or Paper III

‡ Not considered for analysis in Paper III

KOOS: Knee Injury and Osteoarthritis Outcome Score; QoL: Quality of Life; MCL: medial collateral ligament; FCL: fibular collateral ligament; PLC: posterolateral corner; PCL: posterior cruciate ligament

### Machine Learning Analysis

To properly train and test a machine learning model, the data must first be split into training and testing sets. The training sets are used to fit various machine learning models, while the test sets are reserved for model evaluation. These sets are separate from one another, meaning each patient appears in only one set and there is no crossover between these groups. For all three studies, the complete dataset was divided such that 75% of the data was placed into a training set while the remaining 25% was allocated to the hold-out test set. These divisions were performed randomly for each study, meaning the composition of the groups varied between studies. All models were trained and tested using the program R (R Core Team).

Several different types of supervised learning approaches exist for the purposes of developing outcome prediction models. The most commonly applied include regression models, support-vector machines, decision trees, and ensemble methods<sup>52</sup>.

Regularized regression models represent modifications to simple linear and logistic regression that regularize and constrain the weights of the model to decrease overfitting<sup>52</sup>. Regression models are relatively simple machine learning techniques to implement and interpret, and have been shown to outperform more complex methods in the right setting<sup>79,80</sup>. However, the regression models are often inadequate when there is a large volume of complex and interacting variables<sup>52</sup>.

Support-vector machines are a supervised learning approach that constructs a hyperplane, or decision boundary, to separate classes of data<sup>52</sup>. These are most commonly applied to predict binary outcomes.

Decision trees generate several dichotomous questions that are used to split and isolate the patients into their respective classes<sup>52</sup>. An Ensemble method is any machine learning approach that combines multiple methods, and random forests represent one of the most common examples of this. In random forest machine learning, multiple decision trees are combined into a

single “forest.” These forests are capable of producing highly accurate prediction algorithms that are relatively easy to interpret.

In addition to random forests, other ensemble methods have been developed and are often employed for prediction tasks<sup>52</sup>. Bagging is a term that refers to an ensemble method based on several models that were created in parallel, while boosting denotes an ensemble method that refines models sequentially to create a final model.

In each of the three prediction model development studies, four machine learning approaches were utilized. In all cases, the models were adapted for censored time-to-event data<sup>81</sup>. Censoring enables the consideration of patients who have not yet reached a given follow-up time, by including their event-free time in the model development. The following machine learning models were chosen for each study as they represent a variety of different approaches intended for this type of data and analysis:

1. Cox Lasso (Papers I and III)

The Cox Lasso model applies Lasso (L1) regularization to the Cox proportional hazards model, used for time-to-event data<sup>82</sup>. It selects important predictors by setting less significant ones to zero. The extent of this shrinkage is controlled by a tuning parameter, which is optimized via cross-validation to balance model simplicity and accuracy.

2. Lasso Logistic Regression (Paper II)

The Lasso logistic regression model uses L1 regularization to perform variable selection in a logistic regression framework, setting less important predictors to zero<sup>82</sup>. A tuning parameter, optimized via cross-validation, controls the degree of shrinkage to balance simplicity and fit.

3. Survival Random Forest (Papers I and III)

The survival random forest uses an ensemble of decision trees for time-to-event data, using the log-rank split rule and estimating survival via Kaplan-Meier and Nelson-Aalen estimators<sup>83</sup>. Individual predictions are averaged across all "out-of-bag" bootstrap samples. Accuracy is measured by 1-C, where C is Harrell's concordance index, reflecting prediction ranking quality.

#### 4. Random Forest (Paper II)

The random forest model for binary classification is an ensemble of decision trees built from bootstrap samples with randomly selected variables at each split<sup>83</sup>. Predictions are averaged across "out-of-bag" samples, and model accuracy is assessed by the overall out-of-bag error rate.

#### 5. Generalized Additive Model (Papers I and II)

A generalized additive model (GAM) is a flexible regression model that allows for non-linear relationships between predictors and outcomes, using smooth terms fit with penalized splines<sup>84</sup>. For time-to-event data, the model uses a Cox proportional hazards framework with smooth terms.

#### 6. Gradient Boosted Regression (Papers I-III)

Gradient boosted regression (GBM) iteratively fits a series of regression trees to minimize prediction error<sup>85,86</sup>. For time-to-event data, it optimizes the negative log partial likelihood under a Cox model. Each iteration updates the model in the direction of the loss function's gradient.

#### 7. Super Learner (Paper III)

The Super Learner is an ensemble method that combines multiple machine learning models to improve prediction accuracy<sup>87</sup>. It creates a weighted average of its component models by cross-validating each and optimizing the weights to minimize error. In Paper III, the Super Learner combined survival random forest and gradient boosted regression models.

### **Missing Data**

Missing data can significantly impact clinical prediction modeling, and all three model development studies employed methods to deal with missing data. First, models were trained and tested using only patients without missing data (complete cases). Then, multiple imputation by chained equations (MICE) was applied to assess the impact of excluding patients with incomplete data<sup>88</sup>. This method fills in missing data based on observed patterns and uses this expanded dataset to retrain and test the models. Model performance was then compared between the complete case and the imputed datasets to evaluate whether imputation improved performance. For Paper I, variable distributions were also compared between the complete case

and full datasets to evaluate for differences between these patient groups. In Paper II, inverse probability weighting was performed to evaluate for differences between those with and without two-year follow-up KOOS QoL scores.

### Model Performance Evaluation

There are many different ways to assess the performance of a clinical prediction model, but measures of discrimination and calibration represent the most common and important to report<sup>89,90</sup>. Accordingly, the evaluation of model performance was consistent across all three prediction model development studies, which reported both discrimination and calibration metrics for the test sets. Concordance and the area under the receiver operating characteristic curve (AUC) are measures of model discrimination, measuring how well the algorithm ranks patients in terms of their risk<sup>90,91</sup>. Calibration on the other hand is a measure of model accuracy (goodness-of-fit), referring to how well the predicted probabilities of subsequent revision surgery or inferior patient reported outcome match the actual outcomes observed in the test data<sup>90,92</sup>.

Discrimination was measured using Harrell's C-index in Papers I and III, while the AUC was reported in Paper II. The C-index is a generalization of the AUC and is particularly suited for time-to-event data where some patients have incomplete follow-up<sup>81,89,91,93,94</sup>. Discrimination results range from 0 to 1, with 1 indicating perfect agreement between the predicted risk rankings and the true outcomes. In Papers I and III, concordance was calculated at one, two, and five-year follow-up periods, while Paper II focused on predicting two-year PROMs. In short, the discrimination metric seeks to answer the question: "do patients who experience the outcome have higher risk predictions than those who do not?"

For calibration, all three studies used a modified version of the Hosmer-Lemeshow statistic that accounts for censored data. This method compares the predicted risk with the actual outcomes, grouped into quintiles, and converts the mean squared error of the differences into a chi-squared statistic<sup>89,92</sup>. A larger statistic indicates poorer calibration, and a significant p-value means that the model's predictions are statistically different from the actual outcomes, suggesting mis-calibration. In a well calibrated model, close to x patients out of 100 with a risk prediction of x%

would be expected to experience the outcome. For example, if 100 patients each have a risk estimate of 9% for experiencing subsequent revision surgery, the model would be considered well calibrated if close to 9 patients truly underwent revision surgery.

## **External Validation (Papers IV and V)**

The revision prediction model developed in Paper I was selected for external validation using two external datasets. The NKLR-based Cox Lasso revision model was chosen for further validation for two main reasons. The first is that it demonstrated similar performance while being easier to use, and was therefore more clinically applicable, in comparison with the prediction model developed using the combined NKLR and DKRR data in Paper III. The second, was that the first external dataset (DKRR) did not contain the outcome variable (two-year follow-up KOOS QoL score) required to validate the inferior patient reported outcome model from Paper II. The DKRR records follow-up KOOS at one, five, and ten years post-operatively<sup>64</sup>. The external validity of the NKLR-based prediction model was further assessed using the STABILITY I cohort due to the desire to further evaluate its performance on patients from outside of Scandinavia.

The goal of these two studies was to evaluate the external validity of the revision prediction model and the primary outcome measure was the performance of the model (discrimination and calibration) when applied to the two external datasets. This summary will explore the methods employed across both external validation studies, with a focus on the patient populations, data preparation, missing data, and model performance evaluation.

## **Patient Population**

All patients contained within the DKRR and STABILITY I study were included if they had known values for the five variables that are required for outcome prediction using the Cox Lasso model developed in Paper I.



The five variables required for outcome prediction using the Cox Lasso model are:

- Patient age at primary ACLR
- Pre-operative KOOS QoL score
- Graft choice
- Femur fixation method used for ACLR
- Time between injury and primary ACLR

The makeup of the DKRR was previously reviewed, and patients from 2005-2020 were included. The STABILITY I RCT was a study evaluating the effect of a lateral extra-articular tenodesis (LET) on outcome when added to an ACLR performed using a HT in high-risk patients<sup>76</sup>. The definition of high risk in the STABILITY I study was any patient who met at least two of the following criteria: pivot shift grade  $\geq 2$ , desire to return to high-risk or pivoting sports, and/or generalized ligamentous laxity. The STABILITY I trial included patients from seven sites in Canada and two in Europe (Belgium and United Kingdom). Variable distribution from the NKLR patients was compared with that of the DKRR and STABILITY I external validation cohorts.

### Data Preparation

Data from the DKRR and STABILITY I dataset were recoded to match the definitions used for the NKLR prediction model. Specifically, femoral fixation devices were defined as either “Suspension/cortical fixation device,” “Interference screw,” or “Other,” while graft choice was coded as “BPTB,” “HT,” or “Other.”

All patients in the STABILITY I trial received HT for their ACLR with or without the addition of a LET. To evaluate whether the addition of a LET altered the patients’ revision risk profile, and therefore the accuracy of the prediction model, the graft choice for the STABILITY I patients was entered into the revision prediction model in three different ways:

- All patients coded as HT
- HT plus LET = BPTB
- HT plus LET = Other graft choice

### **Missing Data**

Patients with missing data for any of the five variables required for risk prediction were excluded from the analysis. Patients with complete data for the five variables were compared to the full dataset to evaluate for differences.

### **Model Performance Evaluation**

The approach to model performance evaluation during external validation was similar to the technique used in the original prediction model development studies, including assessment of the discrimination and calibration that accounted for censoring. Paper IV mirrored Paper I and assessed model discrimination using Harrell's C-index while calibration was evaluated using the Hosmer-Lemeshow statistic. Paper V required a modification to the calibration assessment due to the smaller sample size, creating three groups instead of five for the Hosmer-Lemeshow calculation. This change ensured sufficient distribution of revisions in each group while retaining validity of the method. Concordance was calculated using Harrell's C-index as per the original model development study. Model performance for the STABILITY I cohort was compared using all three coding approaches for the graft choice variable listed above, to assess which method was most accurate when considering the addition of LET to the ACLR.

### **Unsupervised Learning (Paper VI)**

To approach the problem of outcome prediction using a different method, unsupervised machine learning was applied to the combined NKLR and DKRR dataset comprised of patients included in the two registries from 2004-2020. The purpose was to identify discrete subgroups of patients with common characteristics and to compare the rate of subsequent revision ACLR between these groups. The goal was to be able to categorize patients into one of the subgroups to enable rapid risk estimation in the clinical setting.

Unsupervised learning is a machine learning approach designed to find patterns or groupings in data without relying on pre-labeled outcomes. Unlike supervised learning, which predicts a target variable, unsupervised methods analyze the relationships among predictor variables to reveal inherent structures in the dataset, with no consideration or knowledge of the outcome<sup>53</sup>.

In clinical applications, this typically follows a three-step approach. First, the model creates the clusters based on the predictor variables. Second, the clusters are evaluated with respect to their defining characteristics. In other words, the clusters are interpreted to determine what factors the model used to define each one and therefore, how future patients should be assigned to a specific cluster. Finally, the outcome of interest can be assessed in each cluster to determine if the risk or event rate varies across clusters.

The dataset used for unsupervised learning was identical to the one used in Paper III. This summary will explore the methodological approach applied in Paper VI with a focus on missing data, unsupervised learning technique, and model output evaluation.

### **Missing Data**

Instead of using imputation to fill in missing values, this study relied on complete-case analysis. This decision was based on the previous analyses in Paper III which demonstrated no notable difference between the multiply imputed data and the complete case dataset.

### **Unsupervised Learning Analysis**

In Paper VI, unsupervised clustering techniques were used to group patients with similar characteristics based on the predictor variables contained within the NKLR and DKRR. Three clustering algorithms were applied:

1. K-Means Clustering

This method groups data points to minimize the variance (sum of squared distances) within clusters<sup>95,96</sup>. The number of clusters (k) must be pre-specified. K-means clustering only accommodates continuous predictor variables.

2. Agglomerative Hierarchical Clustering (AHC)

Similar to k-means, AHC only considers continuous variables, but unlike k-means, this technique builds clusters incrementally<sup>97</sup>. Each data point starts as its own cluster, and pairs of clusters are iteratively merged based on their similarity. The process produces a hierarchical structure that represents possible clustering solutions. The optimal number of

clusters is then decided by identifying a level of complexity that balances simplicity and meaningful subgrouping.

### 3. K-Prototypes Clustering

Designed to handle mixed data types (continuous and categorical variables), k-prototypes extends k-means by accommodating both continuous and categorical variables<sup>98</sup>. It assigns patients to clusters by minimizing a weighted distance metric. The weighting parameter ensures that categorical variables are appropriately balanced with continuous ones. Similar to k-means, the k is pre-determined.

To determine the optimal k, the elbow and silhouette methods were used. The elbow plots the within-cluster sum of squares against k<sup>95,96</sup>. The "elbow point," where additional clusters cease to significantly reduce within-cluster variance, indicates the ideal number of clusters (Figure 3). The Silhouette method defines the number of clusters that maximizes between cluster dissimilarity while minimizing within-cluster dissimilarity<sup>95,96</sup>.

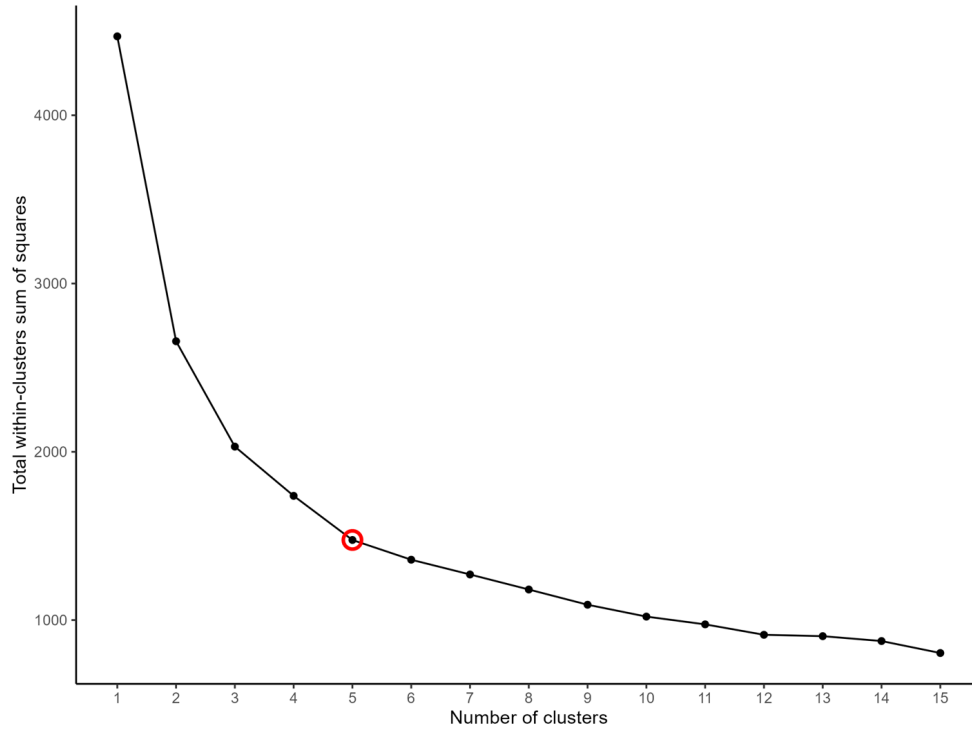


Figure 3: Elbow method to determine the optimal number of clusters based on the combined Norwegian and Danish registry data. The circled point is the “elbow” at which further increases in the number of clusters no longer significantly reduces the within-cluster variance.

### Model Output Evaluation

The output of unsupervised learning analysis is often complex and requires careful interpretation to determine how the model created the patient clusters. For this study, SHapley Additive exPlanations (SHAP) analysis was performed to help explain the defining characteristics of each cluster and minimize the “black-box” effect<sup>99</sup>. The “black-box” refers to the fact that the decision pathways, weighting of feature importance, and potential for bias is obscured with some complex machine learning methods<sup>100,101</sup>. This can lead to a lack of interpretability or trust in clinical models.

SHAP analysis involves the creation of a classification model aimed at predicting the clusters based on the input variables<sup>99</sup>. This classification model is then used to calculate SHAP values for each variable and cluster. SHAP values quantify the influence of each predictor variable on

the model's classification decisions. For example, SHAP values can be used to explain whether a specific variable, such as patient age at surgery, played a major or minor role in defining a specific cluster. By summarizing variable importance at both the cluster-level and patient-level, SHAP analysis can improve transparency, aiding in cluster interpretation and clinical application.

The distribution of variables within each cluster was reviewed by seven orthopaedic sports medicine surgeons involved in the study in light of the SHAP analysis. Together, consensus was reached on how best to define the clusters clinically and enable future patients to be assigned to one of the clusters.

The rate of subsequent revision ACLR was calculated along with Kaplan-Meier survival curves for each cluster. A tree diagram was then created to help classify future patients into one of the clusters including their respective approximate risk of experiencing subsequent revision ACLR.

## Results

### Paper I - Norwegian Revision Risk Prediction

There were 24,935 patients with a primary ACLR and known graft variable from the NKLR that were included in the study, of whom 1,219 (4.9%) underwent subsequent revision ACLR. The Cox Lasso model identified the most influential predictors for revision surgery as graft choice, femur fixation device, pre-surgery KOOS QoL score, time from injury to surgery, and age at surgery. Performance of the survival random forest and GBM algorithms dropped substantially when they were limited to the five variables selected by the Cox Lasso model which necessitated inclusion of all variables for these models. Imputation of missing data did not significantly improve the performance of any of the models. Therefore, the models were trained and tested using only patients with complete data for the required variables. This amounted to 18,887 patients (975 revisions; 5.2%) for the Cox Lasso and GAM, and 13,272 patients (619 revisions; 4.7%) for the survival random forest and GBM models.

The concordance ranged from 0.67-0.69 for all four machine learning models that were evaluated, and all were generally well-calibrated. Concordance was best for the Cox Lasso and GAM algorithms overall (0.68-0.69). Calibration was weakest when predicting two-year revision risk for all models. There was modest evidence of mis-calibration, defined as a calibration p-value 0.01-0.05, when predicting two-year revision risk for all but the GBM (Table 2 and Figure 4).

Table 2: Model Performance Measures – Norwegian Revision Risk Analysis

Revision Probability	Model	Concordance	Calibration statistic	Calibration p-value
1 year	Cox Lasso	0.686	4.89	0.18
	Survival Random forest	0.672	3.12	0.374
	Generalized additive model	0.687	4.79	0.188
	Gradient boosted regression	0.669	4.98	0.174
2 years	Cox Lasso	0.684	11.35	0.01
	Survival Random forest	0.670	11.66	0.009
	Generalized additive model	0.685	11.19	0.011
	Gradient boosted regression	0.666	3.76	0.288
5 years	Cox Lasso	0.683	6.19	0.103
	Survival Random forest	0.670	3.71	0.295
	Generalized additive model	0.684	6.98	0.073
	Gradient boosted regression	0.665	0.38	0.944



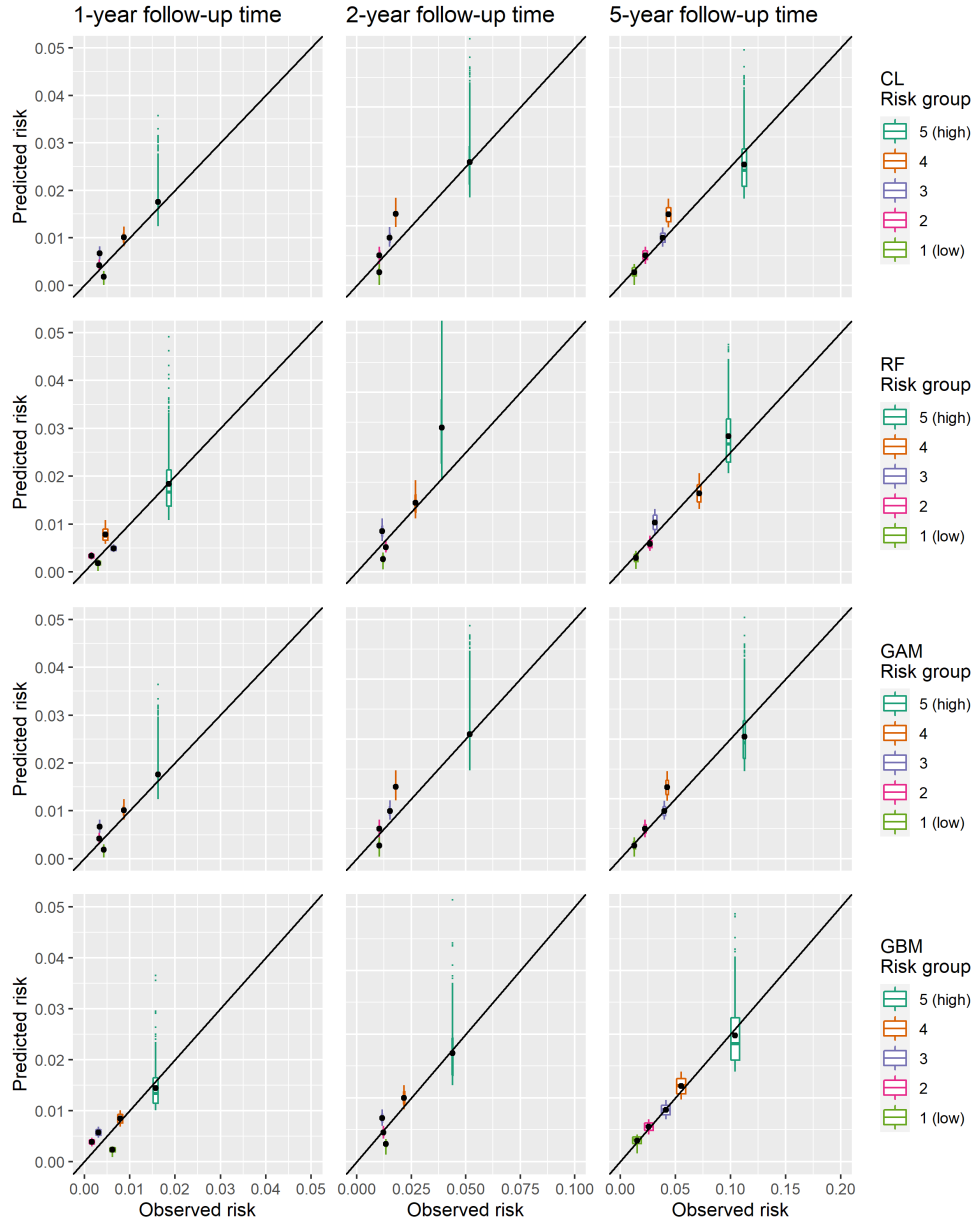


Figure 4: Calibration plots for the revision prediction models at each time point.

CL: cox lasso; RF: survival random forest; GAM: generalized additive model; GBM: gradient boosted regression.

Since the Cox Lasso model demonstrated similar performance and superior ease of use when compared with the other models, it was selected to create a user-friendly online calculator for predicting patient-specific risk of ACL revision ([https://swastvedt.shinyapps.io/calculator\\_rev/](https://swastvedt.shinyapps.io/calculator_rev/)). While the registry-wide revision risk was 4.9%, the calculator enables clinicians to estimate individualized risks, ranging from near 0% for low-risk patients to up to 20% at five years for high-risk individuals (Figure 5).



*Figure 5: QR Code for revision risk calculator*

## **Paper II - Norwegian Inferior Patient Reported Outcome Risk Prediction**

Of the nearly 25,000 patients in the NKLIR, there were 11,630 patients that met the inclusion criteria and had complete two-year follow-up data for the KOOS QoL score. A total of 4,122 patients were excluded due to an unknown graft variable, concomitant non-ACL injury, or follow-up time of less than two years. Of the 20,818 patients that remained, 9,188 were excluded due to missing two-year post-operative KOOS QoL variables. The primary outcome measure of inferior patient reported outcome was reported as subjective failure, which was defined as a KOOS QoL score below 44. This endpoint occurred in 2,556 (22%) patients. Inverse-probability weighting and imputation of missing data demonstrated similar findings between the complete case and full datasets.

Key predictors of inferior outcome identified by the lasso logistic regression model included:

- Pre-operative KOOS scores below the median on all subscales
- Presence of a cartilage injury
- Activity leading to the injury
- Previous surgery to the ipsilateral knee,
- Pre-operative KOOS Sports score
- Pre-operative KOOS QoL score
- BMI
- Age at injury

The random forest model identified additional variables such as age at surgery, graft choice, time between injury and surgery, and fixation devices. The GAM and GBM ranked features similarly to the other models.

All models except the random forest achieved AUC between 0.67 and 0.68. The GAM and GBM performed best with an AUC of 0.68. All models except the random forest were well-calibrated (Table 3). Although the GBM demonstrated similar prediction performance to the GAM, it required more variables for outcome prediction. Therefore, the GAM algorithm was selected to create an online calculator for predicting patient-specific risks of inferior patient reported outcome (KOOS QoL <44) two years post-surgery ([https://swastvedt.shinyapps.io/calculator\\_koosqol/](https://swastvedt.shinyapps.io/calculator_koosqol/)). While the overall risk of KOOS QoL <44 in the registry was 22%, the calculator allows clinicians to provide personalized risk estimates for individual patients (Figure 6).

Table 3: Model Performance Measures – Norwegian Inferior Patient Reported Outcome Analysis

Model	AUC	AUC Confidence	Calibration statistic	Calibration p-value
Logistic Regression (Lasso)	0.67	(0.64, 0.71)	4.57	0.206
Random forest	0.65	(0.62, 0.69)	26.83	< 0.001
Generalized additive model	0.68	(0.64, 0.71)	4.03	0.258
Gradient boosted regression	0.68	(0.64, 0.71)	4.74	0.192



Figure 6: QR Code for inferior patient reported outcome risk calculator

---

### **Paper III - Combined Norwegian and Danish Revision Risk Prediction**

The combined NKLR and DKRR produced a dataset comprised of 62,955 patients, of whom 3,205 (5%) underwent subsequent revision ACLR. Imputation of missing data yielded nearly identical results to those from complete case analyses, with similar concordance confidence intervals and observed calibration ratios.

Key predictors of revision surgery identified by the top-performing models (survival random forest, GBM, and Super Learner) included patient age at injury and surgery, the time between injury and surgery, graft choice, and pre-operative KOOS QoL and Sports scores.

All models except the Cox lasso produced concordance of 0.67 at all follow-up times in the complete case analysis. The Cox lasso model had lower concordance (0.58) and demonstrated moderate evidence of mis-calibration (p-values 0.01-0.05) at two-year and five-year follow-ups. The nonparametric models generally demonstrated better calibration, though some mis-calibration was observed for the Super Learner at one year and five years, and for the random survival forest and GBM at five years (Table 4).

Table 4: Model Performance Measures – Combined Registry Revision Risk Analysis

Revision Probability	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

## Paper IV - External Validation Using the Danish Registry

A total of 10,922 patients from the DKRR had all five variables required for revision risk prediction using the Cox Lasso model developed using the NKLR (Paper I). There were some notable differences in the patient populations between the two cohorts. In comparison to the

NKLR patients from Paper I, the DKRR patients had higher rates of HT use (81% versus 59%) and suspension/cortical femur fixation (72% versus 53%), with lower rates of concomitant meniscus (42% versus 53%) and chondral (14% versus 23%) injuries. Additionally, the revision rate was slightly higher in the Danish (6.9%) cohort compared to the Norwegian (5.2%) cohort. Patients with complete data in the DKRR group were broadly similar to those DKRR patients without complete data, particularly regarding the five variables required for the model.

The NKLR Cox Lasso model demonstrated discrimination (concordance) of 0.68 when applied to the DKRR cohort which was similar to the original NKLR internal validation concordance (0.68–0.69). However, calibration was less accurate for the DKRR population at one year and five years while being similar for the two-year predictions (Table 5).

*Table 5: Comparison of model performance of the revision risk algorithm between the original Norwegian internal validation and the Danish external validation cohorts.*

Revision Probability	Model	Concordance	Calibration statistic	Calibration p-value
1 year	Original Norwegian	0.69	4.89	0.18
	Danish Registry	0.68	22.24	<0.001
2 years	Original Norwegian	0.68	11.35	0.01
	Danish Registry	0.68	11.82	0.008
5 years	Original Norwegian	0.68	6.19	0.103
	Danish Registry	0.68	13.98	0.003

## Paper V - External Validation Using the STABILITY I Patients

A total of 591 patients from the STABILITY I RCT had all five variables required for revision risk prediction using the Cox Lasso model developed using the NKLR (Paper I). There were some notable differences in the patient populations between the two cohorts. Compared to the

NKLR cohort, the STABILITY I patients were younger with a narrower age range (14–25 years versus mean age 28), had shorter and more consistent time from injury to surgery, and uniformly received HT with suspensory femoral fixation.

The Cox Lasso revision prediction model performed best when patients in the STABILITY I cohort who received HT plus LET were coded as having undergone ACLR with BPTB. Concordance values for both one-year and two-year revision predictions were 0.71 which was higher than the concordance observed during model development in Paper I. However, the 95% confidence interval of the concordance among the STABILITY I patients was wider (0.63–0.79). The model was well-calibrated for one-year predictions but demonstrated mis-calibration regarding the two-year predictions, similar to the performance observed with the Norwegian patients in Paper I (Table 6).



Table 6: Comparison of model performance of the revision risk algorithm between the original Norwegian internal validation and the STABILITY I RCT external validation cohorts with patients randomized to hamstring tendon autograft plus lateral extra-articular tenodesis coded three different ways.

Revision Probability	Model	Concordance (95% CI)	Calibration statistic	Calibration p-value
1 year	Original Norwegian Algorithm	0.686 (0.652-0.721)	4.9	n.s.
	STABILITY data (HT + LET = BPTB)	0.713 (0.634-0.791)	2.6	n.s.
	STABILITY data (HT + LET = Other)	0.609 (0.528-0.691)	10.6	<0.01*
	STABILITY data (All patients = HT)	0.674 (0.597-0.751)	8.7	<0.01*
2 years	Original Norwegian Algorithm	0.684 (0.650-0.718)	11.3	0.01*
	STABILITY data (HT + LET = BPTB)	0.713 (0.637-0.789)	11.7	<0.01*
	STABILITY data (HT + LET = Other)	0.608 (0.530-0.688)	8.9	<0.01*
	STABILITY data (All patients = HT)	0.673 (0.598-0.747)	10.2	<0.01*

\*Statistical significance,  $p = <0.05$   
CI: confidence interval; HT: hamstring tendon autograft; LET: lateral extra-articular tenodesis; BPTB: bone-patellar tendon-bone autograft; n.s.: not statistically significant

## Paper VI - Combined Norwegian and Danish Unsupervised Machine Learning Analysis

Only patients with complete data were considered for unsupervised machine learning analysis, which resulted in a combined NKLR and DKRR database that included 28,631 patients. Of these, 1,770 (6.2%) patients underwent subsequent revision ACLR.

The optimal number of clusters was determined to be five. K-prototypes clustering was utilized to generate the five clusters due to the ability to consider both continuous and categorical variables in the analysis. SHAP analysis aided the surgeons' interpretation of the five clusters (Figure 7). Kaplan-Meier survival curves were created and demonstrated unique revision surgery rates among the five clusters (Figure 8). The surgeons' simplified interpretation of the unsupervised machine learning output was used to generate a tree diagram that enables approximate patient classification into one of the five clusters including the corresponding revision surgery rate (Figure 9).

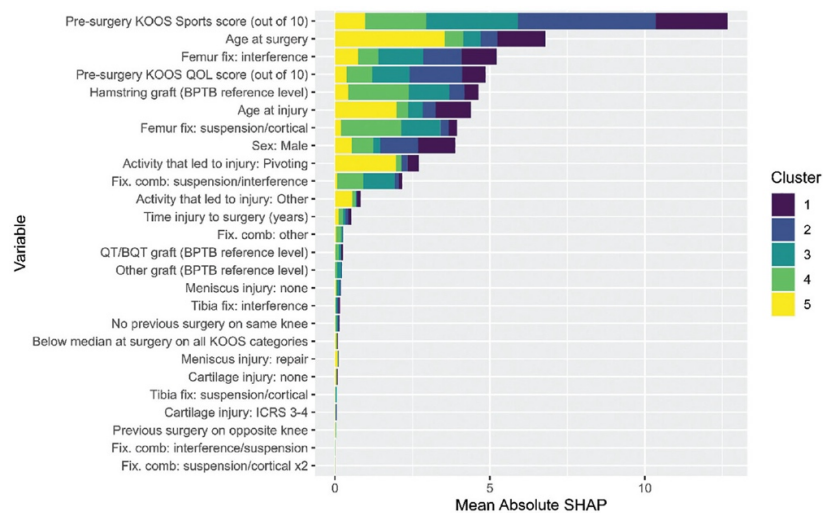


Figure 7: Mean absolute SHapley Additive exPlanations (SHAP) values by variable for each cluster. Colours represent the contributions of the variables assigned to each cluster.

BPTB: bone–patellar tendon–bone autograft; comb: combined; fix.: fixation; ICRS: International Cartilage Regeneration & Joint Preservation Society; KOOS: Knee injury and Osteoarthritis Outcome Score; QOL: Quality of Life subscale; QT/BQT: quadriceps tendon autograft (with or without bone); Sports: Sport and Recreation subscale.

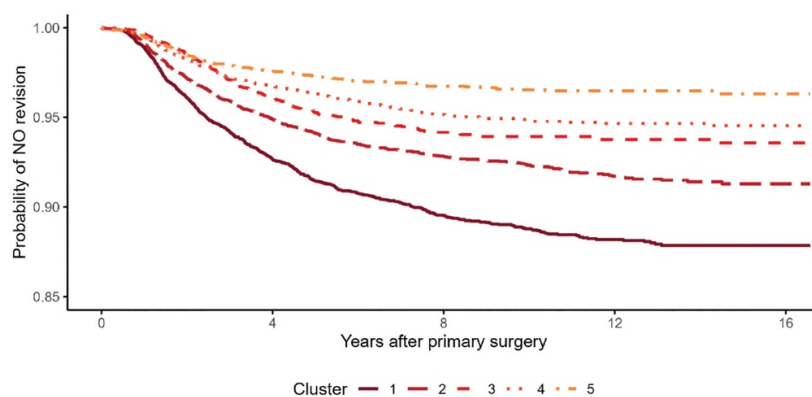


Figure 8: Kaplan-Meier survival curve for all 5 clusters.

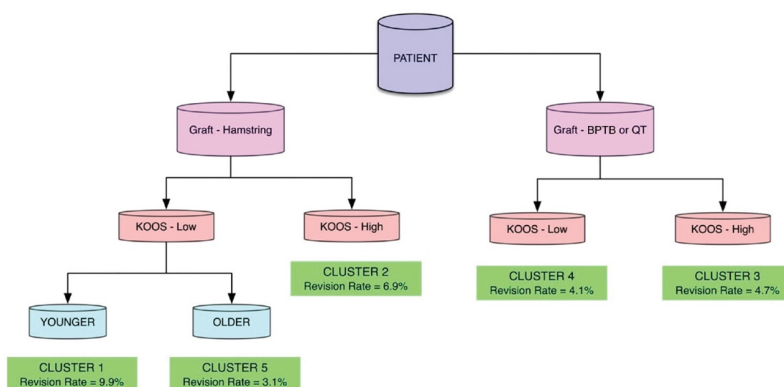


Figure 9: Tree diagram for approximate patient classification by cluster.

BPTB: bone–patellar tendon–bone autograft; KOOS: Knee injury and Osteoarthritis Outcome Score (Sports subscale); QT: quadriceps tendon autograft (with or without bone).

## Discussion

The most important findings from this thesis can be summarized as follows:

- a) machine learning analysis of the NKLRR and DKRR enabled the development and validation of prediction models that demonstrated moderate accuracy for predicting revision surgery and inferior outcome following ACLR, and identified the most important factors used to predict these outcomes
- b) a rigorous approach to clinical prediction modeling has been described, laying the foundation for future innovation
- c) more work is needed to evaluate the performance of the prediction models on patients from outside Scandinavia and to determine the threshold for clinical relevance regarding ACLR outcome prediction
- d) the development and validation of clinical prediction tools may be limited by both the quality and quantity of the available data and national knee ligament registries may benefit from expanded variable collection including additional factors that have been associated with outcome, such as pre-operative laxity, posterior tibial slope, and rehabilitation details.

As the body of literature related to clinical outcome predictions powered by artificial intelligence grows exponentially, it is anticipated that more models will be developed and refined with the intention of more accurate outcome predictions and hopefully, improved patient outcome. What follows is a more expansive analysis of the six studies, including how they fit together within the broader context of ACL outcome prediction. A discussion of the clinical relevance of this thesis will then be presented.

## Prediction Model Development (Papers I-III)

### Main Findings

Collectively, these three studies represent the first time that large ACL databases have been explored using supervised machine learning. The two main takeaways are:

- 1) Analysis of the NKLR enabled the development of online prediction calculators that could be used in the clinic to estimate a patient's risk of subsequent revision (Figure 10) or an inferior KOOS QoL score after primary ACLR (Figure 11).
- 2) Overall model performance was moderate and increasing the sample size from ~25,000 patients in the NKLR to ~63,000 patients in the combined NKLR/DKRR dataset did not improve the accuracy of the revision prediction algorithm, suggesting a need for the registries to increase variable collection to include more variables associated with ACLR outcome.

Several aspects of these studies necessitate further discussion including the performance of the prediction models, chosen end-points, factors associated with outcome, and a review of similar prediction models that have recently been developed.

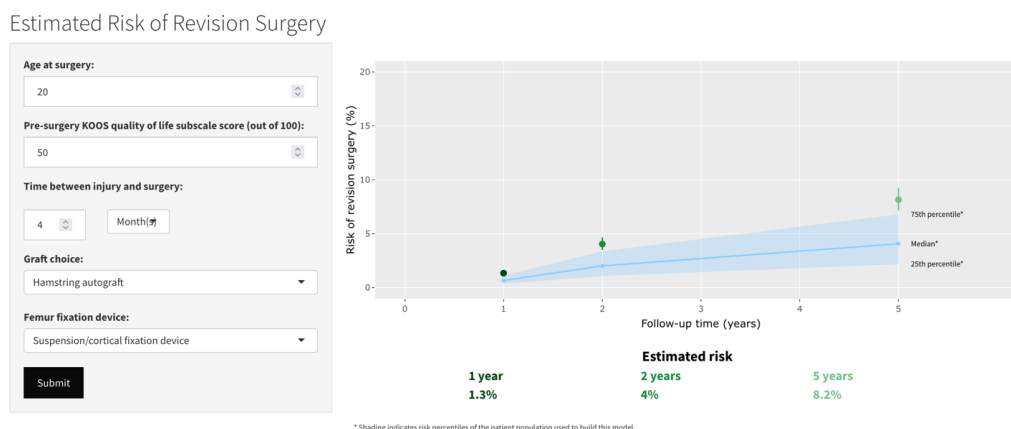


Figure 10: Example output of the online revision risk calculator. The patient is 20 years old with a pre-operative KOOS QoL score of 50 undergoing an ACL reconstruction with hamstring tendon autograft and suspension fixation on the femur four months after ACL injury. Patient-specific risk estimates are shown on the right, along with the median level of risk with 25<sup>th</sup> to 75<sup>th</sup> percentiles based on the Norwegian Knee Ligament Register patient population.

KOOS: Knee Injury and Osteoarthritis Outcome Score; QoL: Quality of Life subscale; ACL: anterior cruciate ligament

Estimated risk of subjective failure two years post-ACL reconstruction (KOOS QoL <44)

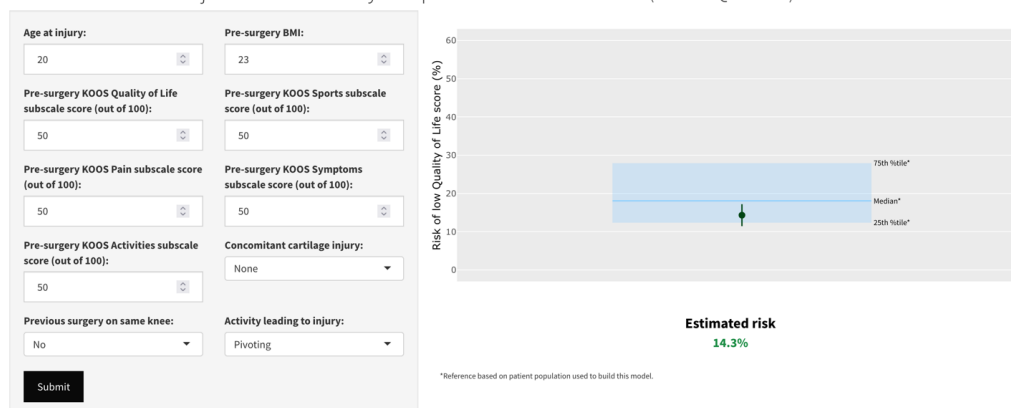


Figure 11: Example output of the online calculator to predict two-year post-operative KOOS QoL score less than 44 after ACL reconstruction. The patient is 20 years old, with a BMI of 23, and pre-operative KOOS score of 50 on all subscales. There is no history of previous ipsilateral knee surgery, no concomitant cartilage injury, and the injury occurred during a pivoting activity. Patient-specific risk estimate is shown on the right, along with the median level of risk with 25<sup>th</sup> to 75<sup>th</sup> percentiles based on the Norwegian Knee Ligament Register patient population.

KOOS: Knee Injury and Osteoarthritis Outcome Score; QoL: Quality of Life subscale; ACL: anterior cruciate ligament; BMI: Body Mass Index

## Model Performance

Model performance of all three studies was similar, producing discrimination values between 0.67-0.69 and a range of calibration values suggesting some well-calibrated models along with modest evidence of mis-calibration at certain time points. How to interpret these results and appropriately put the observed performance into context represents an important concept to review.

Regarding calibration, which measures how accurately the risk predictions reflect the agreement between the estimated and observed event rate, the models developed in Papers I-III were generally well-calibrated. The Hosmer-Lemeshow test produces a calibration p-value, and values above 0.05 are considered well-calibrated<sup>93</sup>. However, the two-year revision risk predictions demonstrated p-values below 0.05, suggesting statistical difference between the predicted and observed outcomes at this time point. There are several reasons why a model may demonstrate mis-calibration, such as a change in variables or techniques over time<sup>102</sup>. This outcome variability may lead to more difficulty achieving adequate calibration over the long-term.

Large sample sizes can also affect the assessment of calibration, as small discrepancies may produce significant p-values. In the case of the revision prediction model developed using the NKLRL patients, Figure 4 demonstrates how the deviation between predicted and observed revisions were within  $\sim 1\%$  for the cox lasso model. Due to the large sample size, this resulted in significant p-values for the two-year predictions. As is the case when interpreting any p-value, one must consider not only statistical significance, but also clinical significance. With an average prediction deviation of  $\sim 1\%$ , the predictions would be considered well-calibrated from a clinical perspective.

Discrimination, often quantified by concordance or AUC, assesses how well a model can rank individuals according to their likelihood of experiencing the outcome of interest. The two measures are similar, with some subtle differences. Survival analysis, such as the likelihood that a patient will undergo a revision surgery at some point in the future is typically evaluated using concordance, while AUC is used to predict risk for classification tasks, like inferior patient reported outcome at two years post-operatively<sup>89,93,103</sup>.

Concordance is calculated as the proportion of *all* possible pairs of patients (one who experiences the outcome and one who does not) where the model assigns a higher risk score to the patient who experiences the outcome<sup>93</sup>. In the context of the Cox Lasso model for predicting revision ACLR, a concordance of 0.68-0.69 means that in approximately 68-69% of patient pairs, the model correctly ranks the patient who underwent revision surgery as having a higher risk than the one who did not.

In contrast, AUC is calculated as the probability that a *randomly* chosen patient who experienced the outcome will have a higher predicted risk score than a *randomly* chosen patient who did not<sup>89,93</sup>. For the GAM algorithm for predicting inferior KOOS QoL two years post-operatively, an AUC of 0.68 means that 68% of the time, the model correctly assigns a higher risk score to a patient who had a KOOS QoL  $< 44$  compared to a patient who did not.

Concordance and AUC values range from 0.5 to 1.0 and can be interpreted similarly to one another<sup>89,93,103</sup>. A value of 0.5 indicates no discriminative ability, meaning the model is no better than random chance in ranking patients. In contrast, a value of 1.0 represents perfect discrimination, where the model ranks all patients correctly. When interpreting the performance of a clinical prediction model, one must consider what constitutes an acceptable level of discriminative accuracy.

Traditional teaching would suggest that model discrimination values  $>0.9$  should be considered excellent, while  $>0.8$  is good,  $>0.7$  is fair, and models with values  $<0.7$  are poor<sup>104–106</sup>. However, this convention does not always hold when applied to clinical prediction models, which may be limited by an imperfect outcome measure or the influence of random chance, and most clinically useful algorithms have values that fall in the 0.7–0.9 range<sup>107–109</sup>. In some cases, values below 0.7 are also relevant, such as when model performance has been demonstrated to be superior to human performance at the same task<sup>89,107</sup>. Further, values greater than 0.9 are often biased by model overfitting or mismanagement of the data<sup>110</sup>. For these reasons, combined with the fact that discrimination represents only one aspect of model performance, it has been advocated that discrimination values should be taken in context and presented without labels such as “excellent” or “good.”<sup>103,111</sup>

For Papers I–III, a concordance of 0.68–0.69 was observed. While this level of concordance is acceptable in many clinical prediction models, it also indicates room for improvement, as ideal models would achieve a concordance closer to 0.8 or higher, reflecting stronger discrimination. In reality, it is unlikely that a revision surgery or PROM-based prediction model for ACL patients would ever achieve discrimination close to 1.0 for two main reasons. First, the selected outcomes and measurement technique are flawed – not all patients who fail will go on to have revision surgery or low PROMs, and some patients may have experienced the outcome but were lost to follow-up in the registry. Secondly, random chance, such as a subsequent injury from a motor vehicle collision, may lead to revision surgery or inferior PROMs but cannot be accounted for in the prediction models. The observed values imply the ACL prediction models may be useful in identifying patients at higher risk of revision surgery and inferior KOOS QoL, but further work is needed to clarify the clinical utility, a topic that will be addressed later.



## Outcome Measures

The choice of outcome measure for a prediction model can impact the performance and clinical utility of the model. As alluded to previously, outcome measures such as revision surgery or inferior KOOS QoL score are prone to measurement error and may not completely reflect the desired outcome. Ultimately, the goal of the prediction models was to identify patients at an increased risk of a failure of ACLR, which is a challenging outcome to isolate. Failure of ACLR has been defined as graft rupture, persistent laxity, revision surgery, failure to return to sport, and several other criteria based on various PROMs<sup>19,112</sup>.

The primary focus of this thesis was on the prediction of subsequent revision as surrogate for clinical failure of primary ACLR. As mentioned, this endpoint is imperfect since some patients may have a failure but not undergo revision surgery, while other patients may have a revision surgery that is not accurately captured in the database<sup>113,114</sup>. Despite these limitations, revision surgery represents the most objective indicator of a failed ACLR and is reliably captured in the NKLR and DKRR<sup>64,70</sup>.

To account for the fact that some patients may experience ACLR failure but not undergo a subsequent revision surgery, Paper II sought to estimate the risk of an inferior patient reported outcome two years after surgery. The definition of an inferior outcome, which was termed “subjective failure” in Paper II, was a KOOS QoL score <44. This distinction was made based on two previous studies. The first, was a RCT that compared rehabilitation and early ACLR with rehabilitation and the option of a delayed ACLR if needed<sup>78</sup>. In that study, the authors defined an arbitrary cutoff value of 44 on the KOOS QoL based on the premise that it was “consistent with a report of more than moderately decreased knee-related quality of life.”<sup>78</sup> The second study was based on the NKLR and found a substantially higher rate of eventual revision ACLR among patients with two-year post-operative KOOS QoL scores below this threshold<sup>73</sup>. Since the publication of those papers, many other studies have adopted that threshold as a marker of a poor outcome or failure after knee surgery<sup>115–119</sup>.

Despite the relatively common practice of using a cutoff value of 44 on the KOOS QoL to denote a failure, there are some problems with this that should be discussed. First, the custom of

labelling patients who fall below this threshold as having a “subjective failure” is a misnomer, as this finding more accurately represents an inferior patient reported outcome on one of the five KOOS subscales. A more global assessment of failure would be needed to accurately classify a patient as having failed, and subjectivity should be avoided.

The second issue with the KOOS QoL <44 threshold as a prediction target is that it may not be the most clinically relevant outcome. The overall validity of the KOOS regarding the assessment of young patients after ACLR has been disputed in the recent literature<sup>19,120–122</sup>. More recently, the patient acceptable symptom state (PASS) has been identified as a useful marker of patient satisfaction after ACLR<sup>14,123–127</sup>. The PASS threshold can be determined statistically or via an anchor question and seeks to dichotomize patients into two groups: those who are satisfied with their outcome and those who are not<sup>123</sup>. An example of an anchor question designed to determine whether or not a patient has achieved PASS is: “Taking into account all the activities you have during your daily life, your level of pain, and also your functional impairment, do you consider that your current state is satisfactory?”<sup>128,129</sup> The PASS threshold for the KOOS after ACLR has been defined and this outcome measure is gaining popularity<sup>14,124–127</sup>. Future prediction models may offer more clinical relevance if they seek to predict PASS after ACLR.

### Factors Associated with Outcome

Multiple factors have been identified that place a patient at an increased risk of failure following ACLR<sup>11,20–24,28–31,33–37,45,130</sup>. Through a process known as feature selection, whereby the machine learning model sequentially excludes variables from the dataset that do not impact the prediction accuracy, the number of variables required for outcome prediction can be narrowed. In the NKLR prediction models, only five variables were required for revision prediction and eight were required for prediction of inferior patient reported outcome.

All five variables that were required for the estimation of revision risk have previously been associated with ACLR failure. It was also interesting to note that prediction of revision relied on some variables that are modifiable (time between injury and surgery, graft choice, and femoral fixation device) while the inferior patient reported outcome prediction model variables were non-modifiable by the surgeon. In fact, the two most important variables identified by the GAM

to predict a low post-operative KOOS QoL score were the patient's pre-operative KOOS scores. This finding is consistent with other studies that have identified pre-operative PROMs as strong predictors of post-operative PROMs<sup>131</sup>.

A general premise fundamental to machine learning is that model performance is only as good as the data available. Both data quality and quantity are required to develop strong and clinically useful models. An advantage to using national registry data is that there is sufficient data quantity, especially when registry data is pooled. However, the observed discrimination of 0.67-0.69 combined with the fact that the performance of the revision risk prediction model did not improve when the NKLR and DKRR were merged, suggests inherent problems with data quality relative to the complexity of the outcome.

This weakness of the national knee ligament registries that limits the predictive capacity largely relates to the variables that are collected. Although the registries collect multiple variables and have been able to identify several factors associated with outcome since their inception, they fail to capture some important variables that would likely aid in outcome prediction due to their apparent association with graft failure. These include radiographic measures like alignment<sup>37,132-136</sup>, physical examination findings such as degree of knee laxity<sup>21,45</sup>, rehabilitation information<sup>26,27,137</sup>, or surgical technique details such as tunnel position<sup>138,139</sup>. For the national knee ligament registries to produce prediction models with improved accuracy, they must evolve to capture more of these clinically relevant outcomes. This is no easy feat however, as the addition of new variables to the registries increases burden on the surgeons who are responsible for data collection in the current workflow. The result is potential survey fatigue and decreased compliance, which would negatively impact both data quality and quantity<sup>65</sup>. Potential ways to leverage artificial intelligence to overcome these obstacles will be presented in the *Future Opportunities and Next Steps* section of this thesis.

### Other ACLR Outcome Prediction Models

As of December 2024, five other ACL revision prediction models have been published. In 2020, the MOON group published an algorithm for predicting graft rupture following ACLR based on multivariable regression modeling of 770 patients followed for six years after surgery<sup>140</sup>. This study led to the development of an online calculator for risk prediction (<https://acltear.info/acl->

reinjury-risk/acl-autograft-retear-risk/) and underwent subsequent external validation using 618 patients from the STABILITY I RCT<sup>141</sup>. The discrimination of the model was 0.67 during development, and the authors report an improved discrimination of 0.73 during model external validation. However, the sample sizes were small in both the development and validation cohorts, and the discrimination confidence interval was wide, suggesting the true discriminative ability remains uncertain. Additionally, the model was limited to patients between the ages of 14–22 and only included BPTB or HT graft choices. Factors used to predict graft rupture using this model include patient age, height, weight, sex, sport, and activity level.

Four other ACLR outcome prediction models have been developed using machine learning techniques and demonstrate impressive model performance. Usami et al. developed a model to predict ipsilateral ACL graft rupture that only requires two factors: age at surgery and graft type<sup>142</sup>. The corresponding AUC was 0.81. Ye et al. published their machine learning approach to predict multiple ACLR outcomes including graft failure and reported an AUC of 0.94<sup>143</sup>. The most influential predictors of graft failure were found to be medial meniscus resection, participation in competitive sports, and high posterior tibial slope. Zhang et al. applied an ensemble model to the same cohort as the previous study by Ye et al.<sup>143</sup> to predict clinical ACLR failure (defined as graft rupture or rotational laxity) and reported an AUC of 0.91<sup>144</sup>. Eight variables were required for outcome prediction in this study: follow-up period, knee laxity grade, time from injury to surgery, participation in competitive sports, posterior tibial slope, graft diameter, age at surgery, and medial meniscus resection. Finally, Kunze et al. developed machine learning models to predict the achievement of minimal clinically important difference (MCID) on the International Knee Documentation Committee (IKDC) score at a minimum two years post-operatively<sup>145</sup>. The authors report a discrimination value of 0.82 and the top five most predictive features were BMI, MCL laxity grade, femoral fixation, history of contralateral knee surgery, and pre-operatively knee range of motion.

Although all four of these machine learning models demonstrate superior discriminative ability compared with the MOON<sup>140</sup> and national registry-based algorithms of this thesis, none have been externally validated and the sample sizes were 386 (Usami et al.)<sup>142</sup>, 432 (Ye et al. and Zhang et al.)<sup>143,144</sup>, and 442 (Kunze et al.)<sup>145</sup>. The problem with prediction algorithms produced from small

populations like these are that they are prone to overfitting. Model overfitting occurs when an algorithm learns patterns specific to the training data that do not generalize to new, unseen data<sup>146</sup>. This can result in excellent performance on the test set but poor accuracy or reliability on new data. This reinforces the concept that data quality and quantity are both crucial components of model performance and clinical utility.

## **External Validation (Papers IV and V)**

### **Main Findings**

The most important findings from the two external validation studies are that the NKLR revision prediction model demonstrated similar performance when applied to the different patient populations, though there were limitations of each study. This discussion will review the performance of the model during external validation, the finding that LET may influence how graft choice should be entered into the risk calculator, and the importance of external validation prior to clinical application of machine learning models.

### **Model Performance**

Paper IV demonstrated that when the NKLR revision prediction model was applied to patients from the DKRR, it maintained its discriminative ability with a concordance similar to that observed in the original dataset. This suggests that the algorithm's capacity to rank patients according to their risk of revision surgery generalizes across populations with differing surgical trends and injury characteristics. However, calibration was worse at one year and five years and similar at two years, leading to a model that was mis-calibrated overall for the DKRR patients. This observation was likely due to the large sample size and the variation in injury and surgical trends between the two populations, since the proportion of patients with HT ACLR and suspension/cortical femoral fixation was much higher in the DKRR cohort. This highlights the influence of population-specific factors on model performance and the need to account for these when applying prediction models to populations with different variable distributions.

Paper V further assessed the NKLR model's performance in a different context: a randomized trial with a narrowly defined patient population. While concordance values were similar to those in the NKLR and DKRR datasets when the patients who underwent ACLR with HT plus LET were coded as BPTB, the wide confidence intervals indicate uncertainty about the model's true performance for this group. This highlights the importance of adequate data quantity when evaluating model performance – a requirement that applies to not only the model development phase but also to external validation. The NKLR model was well calibrated for predicting one-year outcomes but less reliable at predicting risk at two years, consistent with the original NKLR model performance and reflective of the inherent challenges in longer-term prediction due to increased variability.

### The Effect of Lateral Extra-Articular Tenodesis

The other important outcome from Paper V is the fact that patients who had a LET with their HT ACLR behaved more like those receiving BPTB in terms of revision risk, a conclusion consistent with prior literature based on the MOON revision prediction calculator<sup>141</sup>. The LET is an adjunctive procedure to control the anterolateral instability that often accompanies ACL injury and the addition of a LET to ACLR is gaining in popularity as the role of the peripheral stabilizers in post-operative stability becomes more clearly understood<sup>76,147–150</sup>.

The logic underlying the decision to evaluate the effect of coding a HT plus LET as a graft choice other than HT in the revision risk model was based on two underlying assumptions. The first, was that failure rates are reportedly lower when a LET is performed concomitantly with a HT ACLR<sup>76</sup>. The second, is that in the Scandinavian ACL registries, HT ACLR have a higher revision rate than BPTB ACLR<sup>28,31</sup>. It therefore followed that the addition of a LET for patients having a HT ACLR would potentially result in revision surgery rates that more closely approximated those of a patient who received a BPTB.

The influence of LET on revision risk was not fully appreciated during the original model development using the NKLR owing to the low occurrence of this variable which was only added to in the registry in 2018. As the indications for LET evolve, it is anticipated that future analysis using the Scandinavian registries, both through conventional and novel statistical

approaches, will enable further evaluation of the impact this procedure has on patient outcomes and clarify the importance of this variable on risk prediction.

### **The Importance of External Validation**

The external validation of machine learning models is essential to fully understand their clinical relevance and applicability to patient populations that differ from the one used to train and internally validate the algorithm. However, despite the TRIPOD Statement's strong recommendation that external validation be performed prior to clinical deployment of prediction models, this crucial step is infrequently performed in orthopaedic research<sup>77,151</sup>. Overall, the lack of external validation is a major barrier to widespread adoption of published models – and rightly so. Clinicians must understand that available models may not perform as accurately for their patients as suggested by the developers and should therefore take caution when considering implementation.

One of the advantages of using national registry data to develop and validate clinical prediction algorithms is the fact that the results may be generalizable to a wide swathe of the population based on the diverse array of patients and surgeons supplying data. This is in contrast to smaller, more focused studies of individual surgeon, institution, or regional databases. This emphasizes somewhat of a catch-22 however, whereby models created for a more general population may not perform or apply as well to individual niche practices and vice versa. It is therefore important to consider these differences between the data used to develop the models and the intended population, and whenever possible, establish the external validity on populations that closely resemble the target population that will use the models.

### **Barriers to External Validation**

The external validation of clinical prediction models can pose several challenges for clinician scientists. First, a large volume of data is required for adequate model validation, as demonstrated by the wide confidence interval observed in Paper V, and it can be difficult to find a suitably large database containing all of the required predictor variables. Ideally, the validation patient population should be different enough to represent a new cohort, while being similar to the development population with respect to the nature of data collection and outcome tracking. As prospective and retrospective data collection becomes easier and more integrated into health

care records, this may become less of an issue. Another barrier to external validation is the potential for regulations limiting data transfer between countries or health regions due to local legislation and privacy concerns. The sharing of machine learning algorithms rather than patient data represents one strategy to overcome these challenges and requires collaborative efforts and trust between study groups.

## Unsupervised Learning Analysis (Paper VI)

### Main Findings

The most significant finding of this novel exploration of the combined NKLR and DKRR dataset is the generation of five distinct groups of patients that each had unique revision rates following primary ACLR. The clusters could be grouped according to revision rate, with Cluster 1 patients considered to be high-risk (9.9%), Cluster 2 patients to be moderate risk (6.9%), and patients in Clusters 3-5 considered low risk (3.1-4.7%). Aided by a SHAP analysis to overcome the black-box effect of the unsupervised learning method, the distinguishing characteristics of each cluster were defined, potentially enabling the assignment of future patients undergoing ACLR into one of the five clusters based on age, graft choice, and pre-operative KOOS Sports subscale score. The result was a tree diagram that would facilitate rapid risk stratification in the clinical setting should the model demonstrate acceptable classification performance during external validation (Figure 9). The distinguishing characteristics of each cluster were:

- Cluster 1: young patient with HT autograft and low baseline KOOS Sports score
- Cluster 2: patient with HT autograft and high baseline KOOS Sports score
- Cluster 3: patient with BPTB or QT autograft and high baseline KOOS Sports score
- Cluster 4: patient with BPTB or QT autograft and low baseline KOOS Sports score
- Cluster 5: older patient with HT autograft and low baseline KOOS Sports score

### The Challenge with Cluster Interpretation - The Black-Box Effect

This unsupervised approach to risk stratification highlights the potential of machine learning techniques to discover complex, previously unrecognized interactions between patient, surgical, and outcome-related variables in a dataset, but also presents some unique challenges. One such



challenge that may plague machine learning models, particularly unsupervised learning, is the "black-box" nature of the algorithm. The opaque decision-making process of machine learning models can obscure the understanding of how clusters are derived or what specific variable interactions contribute to the final output. Despite the advent of techniques like SHAP analysis which can enhance explainability, the underlying complexity of the machine learning algorithm still limits transparency.

The black-box effect may hinder clinical adoption by creating uncertainty among clinicians regarding the reliability of the model's predictions. For example, features such as graft choice, age, and pre-operative KOOS formed the basis of the clustering in Paper VI, but it remains unclear how subtle, multidimensional interactions between these variables were weighted. This lack of clarity may raise concerns about whether the model might miss or misinterpret critical factors not explicitly included in the clustering if used prospectively.

An additional concern relates to over-simplification of the clustering for clinical use. This simplification, while necessary for clinician interpretability and rapid risk stratification, can impact the accuracy of patient clustering in two ways. First, simplifying the multiple variables used for clustering down to only three, in this case age, graft choice, and pre-operative KOOS Sports subscale score, fails to consider the effect that other variables like the patient's pre-operative KOOS QoL score or the activity type that led to ACL injury may have on the assignment of a patient to a specific cluster. This is particularly evident when you consider that 10% of the patients in Cluster 4 had a HT ACLR, which would have unintentionally led to those patients being assigned to either Cluster 1 or Cluster 5 (depending on their age) by clinicians using the tree diagram (Figure 9). The second challenge with simplification is that the precise cut-off points for the continuous variables like age and pre-operative KOOS remain unclear, which may lead to problems when categorizing patients into "high" and "low" groups in the clinic. Overall, this simplification process serves to distance the model from its raw data-driven origins, leading to the potential for clinicians to assign patients to a different cluster than the model would have chosen.

As techniques to account for and mitigate the black-box nature of machine learning algorithms evolve, these challenges will hopefully become less impactful. In the meantime, those developing and using these models must proceed with caution to ensure that the outputs of machine learning models are not only accurate but also understandable and actionable within the clinical context.

## Putting it all Together

After review of these six studies involving machine learning analysis of the Norwegian and Danish knee ligament registries, the biggest question that remains is: “what does it all mean?” The section that follows will seek to address this important question by exploring how they fit together in a broader clinical context.

## Clinical Relevance

This project started with a goal that, on the surface, was relatively straightforward: to be able to accurately explain to a patient in the clinical setting what their individual and specific risk of ACLR failure or a poor outcome may be. These outcomes were chosen as they represent clinically relevant endpoints that are most often reported in the literature, but more importantly, are of extreme interest and importance to patients undergoing ACLR<sup>19</sup>. The underlying premise was that accurate outcome prediction is extremely difficult for clinicians to ascertain due to the overwhelming number of factors that not only influence a patient’s risk, but also interact with one another in complex ways. In contrast, if there were only one or two factors that drove risk, outcome prediction may be much simpler. Therefore, the task of this thesis was to address the problem of how to navigate the multitude of factors at play and develop a way to quantify and stratify patient risk. Complex problems often warrant complex solutions, and so the emerging field of machine learning was employed as a way forward.

The concept of accurate ACLR outcome prediction is clinically relevant for several reasons. First, a more informed surgical discussion can help patients and surgeons to have realistic expectations that are aligned with one another. These discussions can also help with surgical decision-making, since non-operative management remains a viable option for many patients with ACL deficiency<sup>78,152,153</sup>. These discussions are especially important when counseling

skeletally immature patients or patients with a lower functional demand. Skeletally immature patients have been found to have a significantly elevated rate of ACLR graft failure and run the additional risk of damage to the growth plates during surgical management<sup>154</sup>. An approach that utilizes early rehabilitation and close follow-up with delayed surgical management when necessary has been employed successfully and may be considered in these patients<sup>78,155–159</sup>. Similarly, for patients with lower functional demand or those who may not experience significant limitations or instability, non-operative management may be a viable alternative to surgical management<sup>78,158–160</sup>, especially in light of evidence that ACLR is associated with increased rates of knee osteoarthritis<sup>161–163</sup>. Since surgeons often make recommendations regarding the optimal treatment for their patients with an ACL injury based on their estimation of the eventual outcome, it is important to optimize the accuracy of these predictions. A better understanding of the accuracy of these predictions is also needed to either instill confidence in these recommendations or to identify areas for improvement in outcome prediction.

Accurate ACLR outcome prediction may also be useful for ultimately improving patient outcomes after surgery. In addition to closing the gap between expectations and reality, a better ability to predict outcomes may enable surgeons to modify treatment or rehabilitation plans for their patients. For example, if surgeons can accurately identify patients at higher risk of experiencing an inferior outcome, they may employ a different surgical technique (such as using a different graft choice or adding an adjunctive stabilization procedure like a LET), encourage those patients to be more diligent with their post-operative rehabilitation, or may delay their final clearance for return to sport. Meanwhile, for patients at lower risk of inferior outcomes, there may be potential for an accelerated rehabilitation approach or earlier return to activities. Additionally, psychological readiness for return to sport has been found to be an important factor affecting outcome after ACLR and surgeons may be able to influence the level of confidence in their patients through better understanding of the expected outcomes<sup>164–166</sup>.

The ideal output for a risk prediction model should clearly articulate the degree of risk to the user. A model that produces a risk score of 8% for example, without any further context, is suboptimal. Is an 8% probability of failure considered high-risk? Would a decrease to 6% given a change in one of the variables represent a significant improvement? The risk output must be put

into context, such that the probability of failure can be interpreted relative to the average or expected range of risk scores. It may also be advantageous to include categorizations along with the numerical value, such as low, medium, high, and very high risk. The provision of greater context will lead to a better understanding of expectations and far more meaningful discussions with patients.

Over five years has passed since work began on this project, and with that passage of time has come some important lessons, some of which were learned the hard way. The biggest lesson from this thesis is that, despite the clinical importance of the research question, the end result is that these studies possess no known clinical relevance themselves. That is, not in their current state.

These novel models identified complex interactions between variables in the NKLR and DKRR, enabling the creation of in-clinic tools designed to rapidly estimate patient-specific outcome probabilities. The revision prediction model also subsequently underwent external validation. However, additional information is needed to properly appraise and determine the true clinical value of these prediction models. The clinical utility of these machine learning models is unknown because the accuracy of their predictions has not been compared with predictions made by orthopaedic surgeons. Without knowing how well surgeons can estimate the risk of a poor outcome for their patients, it is not possible to evaluate whether these, and other, machine learning models outperform clinician judgment. Therefore, there is a critical need to quantify the predictive capability of orthopaedic surgeons to establish the benchmark with which to compare prior and future prediction models for ACLR.

Although the clinical relevance of this thesis remains indeterminate, it is proposed that the six studies within represent a valuable contribution to the literature and profession. At a minimum, they have highlighted some shortcomings of the national knee ligament registries, which have responded by taking steps to collect additional variables that may offer more insight into outcome prediction in the future. Together, this thesis also serves as a foundation upon which future efforts can build and learn from. The overall approach represents sound and rigorous methodology – starting with model development, followed by efforts to improve the model with

new data, and then pursuing external validation of the best performing model. The last step that remains is comparison of the models with human prediction accuracy. It is therefore still possible that the prediction models developed through this project may be clinically relevant if they are found to be superior to the prediction efforts of surgeons through a head-to-head comparison, a possibility that will be explored in the *Future Opportunities and Next Steps* section of this thesis.

In the pursuit of clinically relevant prediction models, a pyramid approach is proposed, designating only those studies that have climbed to the top as potentially holding clinical value (Figure 12).

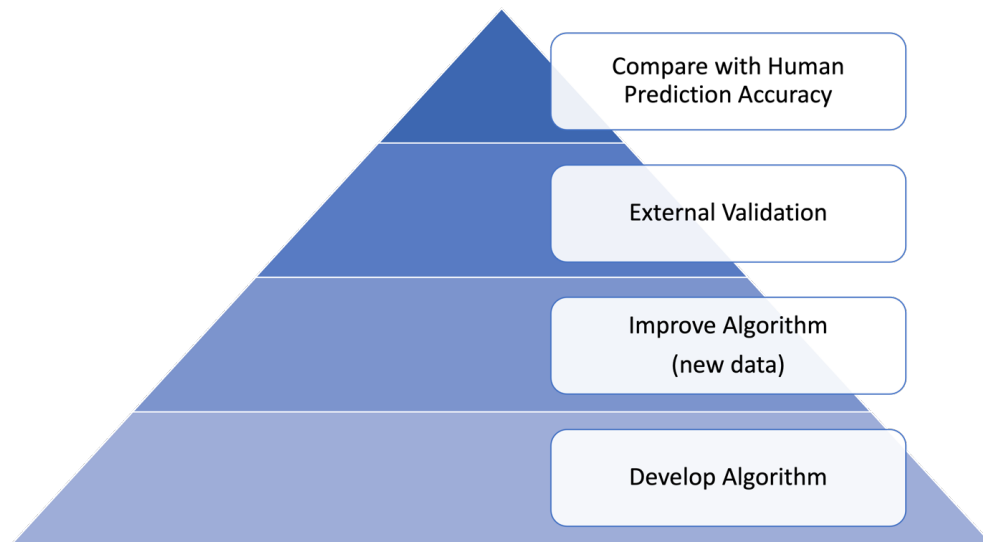


Figure 12: Clinical relevance pyramid for clinical prediction modelling

Although several prediction models have been developed in orthopaedics, most stall out at the bottom of the pyramid as very few have undergone external validation and only a select few have been validated against human performance. The order in which the steps up the pyramid are taken is less important than the climb itself. That is, a model could be found to be superior to human predictions prior to external validation. The important point is that each step is ultimately required to determine clinical relevance.

One final consideration is that as prediction models evolve in the future, new versions may only need to demonstrate superiority over the highest performing existing algorithm if the prevailing model has already been found to out-perform humans. In that situation, external validation remains crucial, along with careful consideration of whether repeat evaluation of performance versus humans is necessary for other reasons.

### **Other Limitations**

The specific models that were chosen for machine learning analysis and the influence of missing data on model performance represent two additional limitations besides the ones discussed previously. Each study that performed machine learning analysis utilized multiple different methods that represent a range of approaches to the problem of outcome prediction. The models were then compared for accuracy and the best performing models were identified. Despite this approach, it is possible that a different machine learning method would have demonstrated superior performance.

Regarding missing data in the registries, it is possible that the missing data was not random. For example, this may be the case if data collection improved over time which would lead to more complete data for patients who were enrolled in the registry more recently. Assessment of the impact of this missing data on model development suggested the patients with missing data were not meaningfully different from those with complete data and sample sizes were large for all studies except the external validation using the STABILITY I cohort. However, limiting the analysis to only those with complete data did result in reduced sample sizes which may have impacted the findings. Additionally, the data collection rates for the registry are high, but not perfect, leading to the likelihood that some patients who experienced revision surgery were not captured in the registry and therefore miscategorized as not having failed<sup>68</sup>.

## Ethical Considerations

The goal of clinical outcome prediction using machine learning is to quantify the likelihood of a particular outcome at a patient-specific level. This section will focus on the ethical considerations when applying retrospective population-based studies prospectively to an individual human being, especially when using new techniques and technology, for the purposes of guiding medical decisions.

The concept that retrospective analysis can be used to guide prospective medical care is not a new one, and in fact, represents level II evidence for prognostic studies<sup>167</sup>. However, information gleaned from these studies are often applied to patients to guide diagnostic and management decisions in a general sense. Put another way, these studies are used to identify factors or variables that may affect one's outcome, and results are often reported in terms of odds ratios, relative or absolute risk, or comparisons which suggest one treatment, intervention, or assessment tool is better, similar, or no worse than another. These studies are crucial to moving the profession forward as we strive to ultimately improve outcome.

What then, makes an outcome prediction tool created through machine learning different from the traditional research approach, and how do we ensure responsible and ethical deployment of these models in clinical practice? These are common questions among clinicians and patients as they grapple with the acceptance of a new approach to health-related data and consider the ethical framework regarding responsible dissemination of findings and clinical implementation.

This represents a rapidly evolving field and there are multiple ethical considerations related to the use of artificial intelligence in orthopaedics. One of the greatest ethical dilemmas facing researchers who seek to develop patient-specific prediction models, relates to the facilitation of appropriate implementation of these novel tools into clinical practice. There is also a responsibility to the entire profession – the researchers, clinicians, and patients alike – to provide appropriate context and honesty regarding the validity of the models in order to ensure that they are not misused or misinterpreted.

There are several differences between a machine learning approach to data compared with a more traditional approach<sup>168</sup> and a few of these are relevant to a discussion focused on research ethics. First, machine learning approaches are often able to elucidate non-linear and complex relationships between variables in a dataset that may not be realizable through more conventional statistical analysis. This is important because, when applied to medical databases containing multiple variables including outcomes, it can enable the creation of patient-specific clinical prediction tools. These may take the shape of an online calculator or other in-office tool that can quantify an individual's specific probability of achieving a specified outcome. There are several examples of this in the medical literature, and prediction performance of these models often match or exceed that of medical experts<sup>169–175</sup>.

Misuse of these clinical prediction models may come in many forms. First, clinicians may inappropriately apply the tool to a population that is inherently different than the one used to develop the model. In fact, this represents the main limitation to the widespread adoption of many machine learning models as most do not undergo the necessary step of external validation<sup>77</sup>. In the case of this thesis, the NKLR prediction tool was subsequently evaluated for external validity using patients from the DKRR. One could correctly argue that these are two very similar patient populations despite being from different countries. The relevance of the prediction model to clinicians and patients in other parts of the world such as in Asia or North America was not evaluated in Paper IV and therefore, it was important not to advocate for deployment of the model outside of Scandinavia until the validity of the model on these different populations has been established.

A second way clinical prediction models may be misused is for a clinician to over-rely on the output of the model and accept the results as truth or fact. This can be said for all artificial intelligence related tasks and is increasingly a topic of concern, for example with regards to large language models like Chat Generative Pre-Trained Transformer (ChatGPT) and its propensity to hallucinate<sup>176,177</sup>. The central issue that both researchers and clinicians must understand relates to the performance of these models, or more specifically, the imperfect nature of their performance.



The ability of a prediction model to estimate outcome following surgery is limited by three main factors: the quality of the data used to generate the model, the quantity of the data, and random chance. Researchers must take steps to minimize the impact of the first two through careful selection of the data source, but the impact of random chance is challenging to quantify and virtually impossible to control. Regarding efforts to predict ACLR failure, no prediction model will ever achieve perfect accuracy due to the inherent randomness associated with re-injury.

There is a duty for those involved in model development to mitigate the risk of over reliance on their prediction models. Researchers must be honest with respect to the performance of their models and not exhibit hyperbolic enthusiasm that may falsely elevate expectations among clinicians. At times this may be a delicate balance between not over-inflating the importance of the model while also advocating for the positive impact these models can have among colleagues who exhibit skepticism (which is occasionally extreme and unjustified). It also follows that, especially in the early stages of machine learning adoption within the field, those that are more familiar with the techniques and technology must take on leadership roles to guide responsible research. These leaders must not only advocate for ethical application, but also must engage in education with their peers to empower clinicians and researchers to critically appraise and properly interpret the findings of these studies.

The third way machine learning driven prediction models may be misused relates to the temporal nature of their development and deployment. This largely relates to machine learning models that do not apply reinforcement learning, which refers to the ability to continue to learn as new data is received. In this thesis, all models were developed using data collected between 2004 and 2020. Over that period alone, surgical techniques, implants, and instruments have advanced – as has the understanding of risk factors and outcomes. Considering the ongoing evolution of the field, it is unreasonable to expect prediction models to maintain relevance far into the future, and so in a way, the work is never done.

The challenge lies in identifying when the point has been reached that a model is no longer relevant. It is therefore imperative that researchers continually evaluate the effects that new data can have on prediction models. This can be done by evaluating the performance of the model

prospectively and watching for deterioration, or by re-calibrating the model with new data every few years. Given that data quantity is an integral component of model performance, there may also be a role for the establishment of international collaborative efforts aimed at aggregating large volumes of clinical data for this purpose. While this concept may have its own associated challenges associated with international data sharing principles and regulations, it nonetheless represents the best way to develop, evaluate, and deploy these prediction models on an ongoing basis.

In an era that is increasingly being shaped by artificial intelligence, the ethical development and deployment of these models in orthopaedics will remain an important topic for the foreseeable future. Clinician scientists must not only apply sound ethical principles to their own work, but to help guide the profession in a similar fashion. The principles of informed consent and data protection concerns will also be at the forefront over the coming years as patient data is collected and utilized at an exponential rate. A thorough understanding of these issues combined with those related to the ethical dissemination and application of results is crucial for scientists engaged in big data research and clinical care. Discussions related to these ethical principles must occur early and often to ensure that patients are not forgotten amongst the vast troves of data, and that they are protected at every step of the journey – from the data collection through to clinical care.

## Future Opportunities and Next Steps

As data collection, computing power, and technology evolve, there is almost no limit to the potential impact of artificial intelligence in orthopaedic sports medicine and the management of ACL injuries. The following section will highlight some exciting developments that are poised to impact the field greatly in the coming years.

### Automated Registry Data Collection

For the past 20 years in Scandinavia, the national knee ligament registries have been working to improve patient care for those with ACL injuries. However, the infrastructure has remained largely unchanged over this time period, with reliance on manual collection, validation, maintenance, and analysis of the registry data. With the recognition that additional variables are needed in the database, along with acknowledgement of the limitations of the current workflow, there is a defined need for evolution of the registries. Recent advancements in can now facilitate this vision.

The current model for the knee ligament registries involves the manual collection of data from individual patients and their surgeons. Patients are first educated on the registry and asked to provide informed consent and data including demographics, details about their injury, and PROMs<sup>63–65,67</sup>. After surgery, patients are contacted at standardized time-points to provide follow-up PROM data. Regarding the data collected from the surgeons, the details of the surgical findings and procedures are manually inputted after surgery. Data are collected at the centralized registry headquarters and manually reviewed for validity and completion, with queries made based on noted deficiencies.

However, the results of Paper III suggest that continued collection of the *same* data is unlikely to substantially increase knowledge and understanding of patient outcomes. To surpass this ceiling, an evolution of the registries is required and is indeed now possible thanks to the rapidly advancing field of artificial intelligence driven solutions. To evolve, the registries must be able to record and analyze additional information that cannot easily be obtained under the current infrastructure and data-collection processes. Patients and surgeons cannot simply be asked to provide more information due to the known limitations of survey fatigue and the subsequent

impact on compliance and data collection accuracy. To surpass this limitation, a shift from the current registry paradigm to one that leverages new technology to collect and interpret data from multiple sources is proposed. The ultimate goal of this initiative is the improvement in patient outcomes following ACL injury.

The current resource-demanding workflow of the registries limits their efficiency and effectiveness. An evolution powered by artificial intelligence should therefore be focused on the following key elements of the registries:

1. What data to collect?

Currently, the registries rely on manual pre-specification of which variables should be collected (structured data). By only collecting pre-specified structured data, newly suspected risk factors that may be associated with outcome cannot be easily evaluated in the registries until the additional variable is added to the prospective registry form and then collected over a period of time. In contrast, a system that supports the prospective collection of *unstructured* data (without pre-specification) would lend itself to a more rapid ability to evaluate these novel risk factors based on the large volume of data already contained within the registry. This can be accomplished through integration with the electronic health record (EHR) and imaging repositories to create a more comprehensive and adaptable surgical registry. Additionally, given the recognized association between several non-surgical factors and ACLR outcome, this approach will also facilitate the ability to collect data related to rehabilitation, return to play readiness, and psychological factors that may play an important role in outcome prediction and optimization.

One of the main reasons the knee ligament registries were created was the ability to identify implants or techniques associated with failure much earlier than would be possible through other surveillance or research methods<sup>63</sup>. Another advantage to a shift toward more automated and comprehensive data collection is the ability to include different diagnoses and surgical procedures in the registry. Examples include patellar instability, cartilage injuries, and meniscus pathology along with their associated surgical treatments. Incorporating more diagnoses and

their corresponding treatments into the registries, along with newly developed devices or implants, will help expand the ability to detect early failures and impact patient outcome.

## 2. How to collect the data?

The number of variables that can be manually collected from surgeons and patients is limited by survey fatigue, and adding more questions results in lower compliance rates. This requires registries to find the balance between the ideal number of variables to collect and the realistic number that can be obtained without sacrificing response rates and accuracy. This presents a significant barrier to registry evolution as improved data quantity (more variables) and quality (high compliance and accuracy) are mutually exclusive in the present registry model.

However, much of the information collected by the registries is already contained within the patient EHR. While this data cannot be directly pulled into the registry database in the current model, this can be alleviated through the automated extraction of these data elements aided by large language models<sup>178</sup>, computer vision<sup>179</sup>, and a direct pipeline from the EHR to the registry. Valuable data sources may include consultation and follow-up clinical notes, operative reports, physiotherapy notes, plain radiographs, magnetic resonance imaging (MRI), computed tomography (CT), and radiology reports. Although this would not completely eliminate the need for data entry from the surgeons since some information is not easily captured in these formats, it would enable more purposeful and efficient manual data collection. Additionally, with less time spent completing duplicative tasks, surgeons will have more time to spend on direct patient care and other obligations.

This innovative approach also has implications for patient-driven data collection using patient-facing apps, modules, and fitness trackers. Together, these tools can help to increase patient engagement, monitor rehabilitation goals and progress, and streamline the collection of PROMs.

## 3. How to validate the data?

To be useful, the registry data must be accurate and valid. In the current model, data collection has been of high quality but relies on manual data entry and validation – steps that may be prone

to human error and require substantial human resources<sup>70</sup>. Data collection from multiple sources can be automatically screened for inaccuracies, inconsistencies, or missing data and flagged for manual review. This guided approach can optimize registry validation in real-time and lessen the administrative burden for the registry.

#### 4. How to use the data?

The possibilities related to a system of ACL registries enriched with artificial intelligence are seemingly limitless. With vast troves of high-quality information from multiple sources, the ability to retrieve and analyze the data will benefit clinicians, researchers, and patients alike. Through the application of advanced algorithms, it may be possible to uncover previously unrecognized patterns and relationships in the registry and overcome the current predictive ceiling of the registries. Additionally, a transformative aspect of this approach is the ability to create digital twins for our patients that can enable surgeons to personalize treatment plans and more accurately predict treatment outcomes. The goal of this approach is to have more informed decision-making and improved patient outcomes.

Another advantage to the incorporation of automated registry data collection is the ability to apply these innovations both prospectively and retrospectively. Standardizing tools to extract data from multiple sources can be used to facilitate the creation of new knee ligament registries for institutions or nations without existing registry infrastructure. Further, the potential to use historical documentation or imaging for data collection means that data, including newly recognized variables of interest, can be extracted in a retrospective manner to supplement the new or existing prospective registries. Expanding the ability to collect and analyze large troves of relevant data around the world holds immense potential for meaningful collaborative initiatives.

Overall, the integration of artificial intelligence into the knee ligament registries presents a unique opportunity to transcend the limitations of traditional data collection and analysis methods. By leveraging new technology, these registries can expand their scope, enhance data accuracy, and unlock new insights that were previously unattainable. This evolution will not only streamline the workflow for surgeons and registry administrators but may also significantly enrich the quality of patient care. Embracing these technological advancements will ensure that the registries continue

to lead the way in musculoskeletal healthcare, providing valuable information that can shape the future of injury prevention, treatment, and ultimately improve patient outcomes.

### **Prospective External Validation**

The prediction models in this thesis were developed and tested using data that date back to 2004, and the application of these models has not been validated prospectively. An important next step is to apply the algorithms developed in Papers I, II, and VI to a cohort of patients from the national knee ligament registries in Denmark, Norway, and Sweden to evaluate the prospective validity of the prior models on patients from 2019 and onwards. Additionally, it has been suggested that the addition of a LET to patients undergoing an ACLR with HT may behave more similarly to a patient receiving a BPTB with respect to revision rates. This study will therefore also evaluate the effect of anterolateral stabilization with respect to the revision prediction model. The results of this study will help answer the question of whether the performance of the ACLR outcome prediction models developed using retrospective data is retained when applied prospectively in Scandinavia.

### **Testing ACL Reconstruction Outcome Predictions (TAROT) Study**

Additional information is needed to properly appraise and determine the true clinical value of the existing ACLR outcome prediction models, as their performance has not been evaluated against orthopaedic surgeons. The Testing ACL Reconstruction Outcome Predictions (TAROT) study (Figure 13) proposes to bridge this knowledge gap through assessment of the accuracy of orthopaedic surgeons at predicting the likelihood of a patient experiencing three clinically relevant outcomes following primary ACLR:

- 1) revision surgery
- 2) functional limitations
- 3) achievement of PASS

Once established, the accuracy of the human predictions will be directly compared with previously published revision prediction algorithms<sup>140,142–144</sup>, including the model developed in Paper I. Additional outcome measures include the change in surgeon predictions and accuracy before and after surgery, and the association between surgeon experience level and prediction accuracy.



*Figure 13: Testing ACL Reconstruction Outcome Predictions (TAROT) Study*

The results of this study will impact the ACLR literature by establishing the benchmark level of accuracy among orthopaedic surgeons regarding outcome prediction. If the surgeons outperform or perform similarly to the previously published prediction models, it suggests that more work is needed to improve clinical prediction models in the future. Once the benchmark accuracy of orthopaedic surgeons has been established, it will open the door to the development of models that seek to outperform, and therefore improve, the ability to predict surgical outcomes. However, if the surgeons underperform relative to one or more of the published models, it affirms the clinical utility of the algorithm while still establishing the minimum clinically relevant level of accuracy for future models.

When tasked with predicting patient outcomes, the information available to inform the prediction is different for the machine learning algorithms and surgeons. Machine learning models are adept at calibrating risk through the interpretation of complex interactions between variables in large datasets. However, they are limited to variables that are easily measured and stored within the EHR or data repositories. In contrast, human risk stratification may be better at integrating unquantifiable information into the decision-making process, such as a conversation with the patient regarding their future goals and ambitions. Fundamentally, the question is not centered around whether surgeons can aggregate the same information better than machine learning algorithms, but whether they are able to use all of the information available to them (some of which is not accessible to machine learning models) to make more accurate predictions.



This large prospective study is a joint project between the University of Minnesota and Fowler Kennedy Sports Medicine Clinic in London, Canada. The aim is to record surgeons' outcome predictions for 2,500 patients undergoing ACLR, involving 25 study sites from nine different countries. A pilot study is currently underway at three sites in Minnesota while funding for the larger study is being pursued. Once completed, this study will answer the question of *how well* surgeons can predict outcome for their patients undergoing ACLR.

## Conclusions

The overall objective of this thesis was to apply machine learning to the NKLR and DKRR to create and validate machine learning algorithms capable of predicting outcome following ACLR with particular emphasis on ease of use and clinical applicability.

AIM 1: To identify the most important risk factors associated with subsequent revision following primary ACLR using supervised machine learning analysis of the NKLR (Paper I)

The most important factors required for estimation of revision risk using the Cox Lasso model were: age at primary ACLR, pre-operative KOOS QoL score, graft choice, graft fixation method on the femur, and time between injury and primary ACLR. Each of these has previously been identified as risk factors associated with failure of ACLR.

AIM 2: To develop a clinically useful prediction model to estimate patient-specific risk of subsequent revision following primary ACLR using supervised machine learning analysis of the NKLR (Paper I)

A novel approach using supervised machine learning was employed to address the problem of outcome prediction. Analysis was carried out using data from the NKLR and several prediction models were developed. Overall, the prediction models were well-calibrated, with discrimination in the 0.67-0.69 range. The Cox Lasso model was selected for the creation of an online risk estimation calculator.

AIM 3: To identify the most important risk factors associated with inferior patient reported outcome two years after primary ACLR using supervised machine learning analysis of the NKLR (Paper II)

Factors required for prediction of two-year post-operative inferior patient reported outcome were: pre-operative KOOS scores below the median on all subscales, presence of a cartilage injury, activity leading to the injury, previous surgery to the ipsilateral knee, pre-operative KOOS Sports score, pre-operative KOOS QoL score, BMI, and age at injury. Unlike the revision prediction model, the factors associated with KOOS QoL <44 two years after surgery were non-modifiable by the surgeon. The most important predictors of post-operative PROM were pre-operative PROMs, a finding consistent with other studies.

AIM 4: To develop a clinically useful prediction model to estimate patient-specific risk of inferior patient reported outcome following primary ACLR using supervised machine learning analysis of the NKLR (Paper II)

A novel approach using supervised machine learning was employed to address the problem of outcome prediction. Analysis was carried out using data from the NKLR and several prediction models were developed. Overall, the prediction models were well-calibrated, with discrimination in the 0.67-0.68 range, not including the random forest which performed more poorly. The GAM was selected for the creation of an online risk estimation calculator.

AIM 5: To improve the accuracy of the revision prediction model through amalgamation of the NKLR and DKRR databases (Paper III)

Merging the data from the NKLR and DKRR to create a larger sample size did not result in improved accuracy of the revision prediction model.

AIM 6: To evaluate the external validity of the NKLR revision prediction model when applied to patients from the DKRR (Paper IV)

The external validity of the Norwegian revision prediction model was subsequently assessed using the patients from the DKRR. Model concordance was similar, but calibration was worse at

predicting one-year and five-year outcomes when compared with the original model development study. This suggested that the revision prediction model may be valid for use outside of the original patient population, but the performance of the model on patients from outside of Scandinavia remains unclear.

AIM 7: To evaluate the external validity of the NKLR revision prediction model when applied to patients from the STABILITY I randomized clinical trial (Paper V)

The external validity of the Norwegian revision prediction model was subsequently assessed using the patients from the STABILITY I cohort. The revision prediction model performed best when patients who had a HT plus LET were coded as having received BPTB. Overall, the model performed similarly with the STABILITY I patients, but true assessment was limited by a small sample size, which produced a wide confidence interval.

AIM 8: To identify distinct subgroups (clusters) of patients within the NKLR and DKRR with similar characteristics using an unsupervised learning technique, and determine how the rate of subsequent revision ACLR differs between them (Paper VI)

Unsupervised learning analysis generated five clusters of patients with unique risk profiles. The five clusters could be divided into high-risk for Cluster 1 (9.9% revision rate), medium-risk for Cluster 2 (6.9% revision rate), and low-risk for Clusters 3-5 (3.1-4.7% revision rate).

AIM 9: To develop a clinically relevant rapid risk-stratification algorithm based on the unsupervised learning clusters (Paper VI)

Using SHAP analysis to guide interpretation of the clusters, a tree diagram was created to facilitate the assignment of future patients to a specific cluster based on age, graft choice, and pre-operative KOOS Sports subscale score. This may enable rapid risk stratification if validated prospectively.

**Key Points**

There were several important takeaways from these studies. First, they highlight both the advantages and shortcomings of the knee ligament registries and have prompted changes to variable collection with the goal of improving data quality and quantity. Second, these studies lay out a systematic approach to the problem of outcome prediction and may be used to guide similar efforts in the future. Finally, although several prediction tools were developed, their clinical utility is uncertain as none have been compared with predictions made by surgeons. This missing step is crucial when determining the clinical relevance of prediction models and represents the next phase in the quest for improved outcome prediction.

Collectively, these studies suggest optimism regarding the future of ACLR outcome prediction. Efforts to expand variable collection and facilitate international collaboration has the potential to build on this foundation and improve the accuracy of outcome prediction models. The knowledge gained from this thesis will be used to further refine ACLR outcome prediction, leading to more informed discussions with patients and, hopefully, improved patient care.

## References

1. Pareek A, Ro DH, Karlsson J, Martin RK. Machine learning/artificial intelligence in sports medicine: state of the art and future directions. *Journal of ISAKOS*. 2024;9(4):635-644. doi:10.1016/j.jisako.2024.01.013
2. Pruneski JA, Williams RJ, Nwachukwu BU, et al. The development and deployment of machine learning models. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(12):3917-3923. doi:10.1007/s00167-022-07155-4
3. Deacon A, Bennell K, Kiss ZS, Crossley K, Brukner P. Osteoarthritis of the knee in retired, elite Australian Rules footballers. *Med J Aust*. 1997;166(4):187-190.
4. Webster KE, Hewett TE. Anterior Cruciate Ligament Injury and Knee Osteoarthritis: An Umbrella Systematic Review and Meta-analysis. *Clin J Sport Med*. 2022;32(2):145-152. doi:10.1097/JSM.0000000000000894
5. Gornitzky AL, Lott A, Yellin JL, Fabricant PD, Lawrence JT, Ganley TJ. Sport-Specific Yearly Risk and Incidence of Anterior Cruciate Ligament Tears in High School Athletes: A Systematic Review and Meta-analysis. *Am J Sports Med*. 2016;44(10):2716-2723. doi:10.1177/0363546515617742
6. Zhang Y, McCammon J, Martin RK, Prior HJ, Leiter J, MacDonald PB. Epidemiological Trends of Anterior Cruciate Ligament Reconstruction in a Canadian Province. *Clin J Sport Med*. Published online October 10, 2018. doi:10.1097/JSM.0000000000000676
7. Leathers MP, Merz A, Wong J, Scott T, Wang JC, Hame SL. Trends and Demographics in Anterior Cruciate Ligament Reconstruction in the United States. *J Knee Surg*. 2015;28(5):390-394. doi:10.1055/s-0035-1544193
8. Mall NA, Chalmers PN, Moric M, et al. Incidence and trends of anterior cruciate ligament reconstruction in the United States. *Am J Sports Med*. 2014;42(10):2363-2370. doi:10.1177/0363546514542796
9. Buller LT, Best MJ, Baraga MG, Kaplan LD. Trends in Anterior Cruciate Ligament Reconstruction in the United States. *Orthopaedic Journal of Sports Medicine*. 2015;3(1):232596711456366. doi:10.1177/2325967114563664
10. Mather RC, Koenig L, Kocher MS, et al. Societal and economic impact of anterior cruciate ligament tears. *J Bone Joint Surg Am*. 2013;95(19):1751-1759. doi:10.2106/JBJS.L.01705
11. 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. doi:10.1177/0363546515578836
12. Wiggins AJ, Grandhi RK, Schneider DK, Stanfield D, Webster KE, Myer GD. Risk of Secondary Injury in Younger Athletes After Anterior Cruciate Ligament Reconstruction: A

- Systematic Review and Meta-analysis. *Am J Sports Med.* 2016;44(7):1861-1876. doi:10.1177/0363546515621554
13. Webster KE, Feller JA. Exploring the High Reinjury Rate in Younger Patients Undergoing Anterior Cruciate Ligament Reconstruction. *Am J Sports Med.* 2016;44(11):2827-2832. doi:10.1177/0363546516651845
  14. Herman ZJ, Kaarre J, Grassi A, Senorski EH, Musahl V, Samuelsson K. Registry-based cohort study comparing percentages of patients reaching PASS for knee function outcomes after revision ACLR compared to primary ACLR. *BMJ Open.* 2024;14(8):e081688. doi:10.1136/bmjopen-2023-081688
  15. Marx JS, Plantz MA, Gerlach EB, et al. Revision ACL reconstruction has higher incidence of 30-day hospital readmission, reoperation, and surgical complications relative to primary procedures. *Knee Surg Sports Traumatol Arthrosc.* 2022;30(5):1605-1610. doi:10.1007/s00167-021-06646-0
  16. Carolan D, King E, Richter C, Franklyn-Miller A, Moran R, Jackson M. Differences in Strength, Patient-Reported Outcomes, and Return-to-Play Rates Between Athletes With Primary Versus Revision ACL Reconstruction at 9 Months After Surgery. *Orthop J Sports Med.* 2020;8(9):2325967120950037. doi:10.1177/2325967120950037
  17. Wright R, Spindler K, Huston L, et al. Revision ACL reconstruction outcomes: MOON cohort. *J Knee Surg.* 2011;24(4):289-294. doi:10.1055/s-0031-1292650
  18. Meena A, Farinelli L, Hoser C, et al. Primary Versus Revision ACL Reconstruction Using Quadriceps Autograft: A Matched-Control Cohort Study. *Orthop J Sports Med.* 2024;12(2):23259671231224501. doi:10.1177/23259671231224501
  19. Svantesson E, Hamrin Senorski E, Webster KE, et al. Clinical outcomes after anterior cruciate ligament injury: Panther Symposium ACL Injury Clinical Outcomes Consensus Group. *Journal of ISAKOS.* 2020;5(5):281-294. doi:10.1136/jisakos-2020-000494
  20. Hewett TE, Myer GD, Ford KR, Paterno MV, Quatman CE. Mechanisms, prediction, and prevention of ACL injuries: Cut risk with three sharpened and validated tools. *J Orthop Res.* 2016;34(11):1843-1855. doi:10.1002/jor.23414
  21. Zhao D, Pan J ke, Lin F zheng, et al. Risk Factors for Revision or Rerupture After Anterior Cruciate Ligament Reconstruction: A Systematic Review and Meta-analysis. *Am J Sports Med.* 2023;51(11):3053-3075. doi:10.1177/03635465221119787
  22. Snaebjörnsson T, Svantesson E, Sundemo D, et al. Young age and high BMI are predictors of early revision surgery after primary anterior cruciate ligament reconstruction: a cohort study from the Swedish and Norwegian knee ligament registries based on 30,747 patients. *Knee Surg Sports Traumatol Arthrosc.* Published online March 16, 2019. doi:10.1007/s00167-019-05487-2
  23. 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. doi:10.2106/JBJS.18.01467
24. 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. doi:10.1016/j.arthro.2011.11.024
  25. Grindem H, Engebretsen L, Axe M, Snyder-Mackler L, Risberg MA. Activity and functional readiness, not age, are the critical factors for second anterior cruciate ligament injury - the Delaware-Oslo ACL cohort study. *Br J Sports Med.* 2020;54(18):1099-1102. doi:10.1136/bjsports-2019-100623
  26. Grindem H, Snyder-Mackler L, Moksnes H, Engebretsen L, Risberg MA. Simple decision rules can reduce reinjury risk by 84% after ACL reconstruction: the Delaware-Oslo ACL cohort study. *Br J Sports Med.* 2016;50(13):804-808. doi:10.1136/bjsports-2016-096031
  27. Kyritsis P, Bahr R, Landreau P, Miladi R, Witvrouw E. Likelihood of ACL graft rupture: not meeting six clinical discharge criteria before return to sport is associated with a four times greater risk of rupture. *Br J Sports Med.* 2016;50(15):946-951. doi:10.1136/bjsports-2015-095908
  28. Persson A, Fjeldsgaard K, Gjertsen JE, et al. Increased risk of revision with hamstring tendon grafts compared with patellar tendon grafts after anterior cruciate ligament reconstruction: a study of 12,643 patients from the Norwegian Cruciate Ligament Registry, 2004-2012. *Am J Sports Med.* 2014;42(2):285-291. doi:10.1177/0363546513511419
  29. Persson A, Kjellsen AB, Fjeldsgaard K, Engebretsen L, Espehaug B, Fevang JM. Registry data highlight increased revision rates for endobutton/biosure HA in ACL reconstruction with hamstring tendon autograft: a nationwide cohort study from the Norwegian Knee Ligament Registry, 2004-2013. *Am J Sports Med.* 2015;43(9):2182-2188. doi:10.1177/0363546515584757
  30. Persson A, Gifstad T, Lind M, et al. Graft fixation influences revision risk after ACL reconstruction with hamstring tendon autografts. *Acta Orthop.* 2018;89(2):204-210. doi:10.1080/17453674.2017.1406243
  31. Gifstad T, Foss OA, Engebretsen L, et al. Lower risk of revision with patellar tendon autografts compared with hamstring autografts: a registry study based on 45,998 primary ACL reconstructions in Scandinavia. *Am J Sports Med.* 2014;42(10):2319-2328. doi:10.1177/0363546514548164
  32. *Norwegian Knee Ligament Register Annual Report.*; 2017.
  33. Hashemi J, Chandrashekar N, Mansouri H, et al. Shallow medial tibial plateau and steep medial and lateral tibial slopes: new risk factors for anterior cruciate ligament injuries. *Am J Sports Med.* 2010;38(1):54-62. doi:10.1177/0363546509349055
  34. Jaeger 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. doi:10.1007/s00402-018-2968-z
35. Salmon LJ, Heath E, Akrawi H, Roe JP, Linklater J, Pinczewski LA. 20-Year Outcomes of Anterior Cruciate Ligament Reconstruction With Hamstring Tendon Autograft: The Catastrophic Effect of Age and Posterior Tibial Slope. *The American Journal of Sports Medicine.* 2018;46(3):531-543. doi:10.1177/0363546517741497
  36. 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. doi:10.1177/0363546513503288
  37. Duerr R, Ormseth B, Adelstein J, et al. Elevated Posterior Tibial Slope Is Associated With Anterior Cruciate Ligament Reconstruction Failures: A Systematic Review and Meta-analysis. *Arthroscopy.* 2023;39(5):1299-1309.e6. doi:10.1016/j.arthro.2022.12.034
  38. Dæhlin L, Inderhaug E, Strand T, Parkar AP, Solheim E. The Effect of Posterior Tibial Slope on the Risk of Revision Surgery After Anterior Cruciate Ligament Reconstruction. *Am J Sports Med.* 2022;50(1):103-110. doi:10.1177/03635465211054100
  39. Cooper JD, Wang W, Prentice HA, Funahashi TT, Maletis GB. The Association Between Tibial Slope and Revision Anterior Cruciate Ligament Reconstruction in Patients  $\leq 21$  Years Old: A Matched Case-Control Study Including 317 Revisions. *Am J Sports Med.* 2019;47(14):3330-3338. doi:10.1177/0363546519878436
  40. Su AW, Bogunovic L, Smith MV, et al. Medial Tibial Slope Determined by Plain Radiography Is Not Associated with Primary or Recurrent Anterior Cruciate Ligament Tears. *J Knee Surg.* 2020;33(1):22-28. doi:10.1055/s-0038-1676456
  41. Lund-Hanssen H, Gannon J, Engebretsen L, Holen KJ, Anda S, Vatten L. Intercondylar notch width and the risk for anterior cruciate ligament rupture. A case-control study in 46 female handball players. *Acta Orthop Scand.* 1994;65(5):529-532. doi:10.3109/17453679409000907
  42. Zeng C, Gao S, Guang, Wei J, et al. The influence of the intercondylar notch dimensions on injury of the anterior cruciate ligament: a meta-analysis. *Knee Surg Sports Traumatol Arthrosc.* 2013;21(4):804-815. doi:10.1007/s00167-012-2166-4
  43. Li Z, Li C, Li L, Wang P. Correlation between notch width index assessed via magnetic resonance imaging and risk of anterior cruciate ligament injury: an updated meta-analysis. *Surg Radiol Anat.* 2020;42(10):1209-1217. doi:10.1007/s00276-020-02496-6
  44. Hughes JD, Boden SA, Belayneh R, et al. Association of Smaller Intercondylar Notch Size With Graft Failure After Anterior Cruciate Ligament Reconstruction. *Orthop J Sports Med.* 2024;12(8):23259671241263883. doi:10.1177/23259671241263883
  45. Guimarães TM, Giglio PN, Sobrado MF, et al. Knee Hyperextension Greater Than 5° Is a Risk Factor for Failure in ACL Reconstruction Using Hamstring Graft. *Orthop J Sports Med.* 2021;9(11):23259671211056325. doi:10.1177/23259671211056325

46. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ*. 2013;346(feb05 1):e5595-e5595. doi:10.1136/bmj.e5595
47. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. *IBM J Res & Dev*. 1959;3(3):210-229. doi:10.1147/rd.33.0210
48. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. *Bull Math Biol*. 1990;52(1-2):99-115; discussion 73-97.
49. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*. 1958;65(6):386-408. doi:10.1037/h0042519
50. Khorsandi SE, Hardgrave HJ, Osborn T, et al. Artificial Intelligence in Liver Transplantation. *Transplantation Proceedings*. 2021;53(10):2939-2944. doi:10.1016/j.transproceed.2021.09.045
51. Muthukrishnan N, Maleki F, Ovens K, Reinhold C, Forghani B, Forghani R. Brief History of Artificial Intelligence. *Neuroimaging Clinics of North America*. 2020;30(4):393-399. doi:10.1016/j.nic.2020.07.004
52. Pruneski JA, Pareek A, Kunze KN, et al. Supervised machine learning and associated algorithms: applications in orthopedic surgery. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(4):1196-1202. doi:10.1007/s00167-022-07181-2
53. Eckhardt CM, Madjarova SJ, Williams RJ, et al. Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surg Sports Traumatol Arthrosc*. Published online November 15, 2022. doi:10.1007/s00167-022-07233-7
54. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550(7676):354-359. doi:10.1038/nature24270
55. Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*. 2019;23(5):408-422. doi:10.1016/j.tics.2019.02.006
56. Ardila D, Kiraly AP, Bharadwaj S, et al. Author Correction: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25(8):1319. doi:10.1038/s41591-019-0536-x
57. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Published online 2017. doi:10.48550/ARXIV.1711.05225
58. Yong L, Zhenzhou L. Deep learning-based prediction of in-hospital mortality for sepsis. *Sci Rep*. 2024;14(1):372. doi:10.1038/s41598-023-49890-9
59. El-Sofany H, Bouallegue B, El-Latif YMA. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Sci Rep*. 2024;14(1):23277. doi:10.1038/s41598-024-74656-2

60. Ramkumar PN, Pang M, Polisetty T, Helm JM, Karnuta JM. Meaningless Applications and Misguided Methodologies in Artificial Intelligence–Related Orthopaedic Research Propagates Hype Over Hope. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*. 2022;38(9):2761-2766. doi:10.1016/j.arthro.2022.04.014
61. Martin RK, Pareek A, Krych AJ, Maradit Kremers H, Engebretsen L. Machine learning in sports medicine: need for improvement. *J ISAKOS*. 2021;6(1):1-2. doi:10.1136/jisakos-2020-000572
62. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z
63. 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. doi:10.1177/0363546507308939
64. Lind M, Menhert F, Pedersen AB. The first results from the Danish ACL reconstruction registry: epidemiologic and 2 year follow-up results from 5,818 knee ligament reconstructions. *Knee Surg Sports Traumatol Arthrosc*. 2009;17(2):117-124. doi:10.1007/s00167-008-0654-3
65. 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. doi:10.1007/978-3-662-58254-1\_39
66. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. *J Orthop Sports Phys Ther*. 1998;28(2):88-96. doi:10.2519/jospt.1998.28.2.88
67. Granan LP, Forssblad M, Lind M, Engebretsen L. The Scandinavian ACL registries 2004-2007: baseline epidemiology. *Acta Orthop*. 2009;80(5):563-567. doi:10.3109/17453670903350107
68. Hamrin Senorski E, Svantesson E, Engebretsen L, et al. 15 years of the Scandinavian knee ligament registries: lessons, limitations and likely prospects. *British Journal of Sports Medicine*. Published online April 9, 2019:bjsports-2018-100024. doi:10.1136/bjsports-2018-100024
69. Roos EM, Lohmander LS. The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis. *Health Qual Life Outcomes*. 2003;1:64. doi:10.1186/1477-7525-1-64
70. 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. doi:10.1186/s12891-022-05183-2
71. *Norwegian Knee Ligament Register: Annual Report*. Helse Bergen; 2024. www.helse-bergen.no/nrl

72. Aga C, Kartus JT, Lind M, Lygre SHL, Granan LP, Engebretsen L. Risk of Revision Was Not Reduced by a Double-bundle ACL Reconstruction Technique: Results From the Scandinavian Registers. *Clin Orthop Relat Res.* 2017;475(10):2503-2512. doi:10.1007/s11999-017-5409-3
73. Granan LP, Baste V, Engebretsen L, Inacio MCS. Associations between inadequate knee function detected by KOOS and prospective graft failure in an anterior cruciate ligament-reconstructed knee. *Knee Surg Sports Traumatol Arthrosc.* 2015;23(4):1135-1140. doi:10.1007/s00167-014-2925-5
74. LaPrade CM, Dornan GJ, Granan LP, LaPrade RF, Engebretsen L. Outcomes After Anterior Cruciate Ligament Reconstruction Using the Norwegian Knee Ligament Registry of 4691 Patients: How Does Meniscal Repair or Resection Affect Short-term Outcomes? *Am J Sports Med.* 2015;43(7):1591-1597. doi:10.1177/0363546515577364
75. Ulstein S, Årøen A, Engebretsen L, Forssblad M, Lygre SHL, Røtterud JH. Effect of Concomitant Cartilage Lesions on Patient-Reported Outcomes After Anterior Cruciate Ligament Reconstruction: A Nationwide Cohort Study From Norway and Sweden of 8470 Patients With 5-Year Follow-up. *Orthop J Sports Med.* 2018;6(7):2325967118786219. doi:10.1177/2325967118786219
76. 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. doi:10.1177/0363546519896333
77. 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. doi:10.7326/M14-0697
78. Frobell RB, Roos EM, Roos HP, Ranstam J, Lohmander LS. A Randomized Trial of Treatment for Acute Anterior Cruciate Ligament Tears. *N Engl J Med.* 2010;363(4):331-342. doi:10.1056/NEJMoa0907797
79. Gravesteyn BY, Nieboer D, Ercole A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol.* 2020;122:95-107. doi:10.1016/j.jclinepi.2020.03.005
80. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
81. 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. doi:10.1016/j.jbi.2016.03.009
82. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Soft.* 2011;39(5). doi:10.18637/jss.v039.i05

83. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat.* 2008;2(3):841-860. doi:10.1214/08-AOAS169
84. Wood SN. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman and Hall/CRC; 2017. doi:10.1201/9781315370279
85. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Statist.* 2001;29(5). doi:10.1214/aos/1013203451
86. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis.* 2002;38(4):367-378. doi:10.1016/S0167-9473(01)00065-2
87. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Statistical Applications in Genetics and Molecular Biology.* 2007;6(1). doi:10.2202/1544-6115.1309
88. Buuren S van, Groothuis-Oudshoorn K. **mice** : Multivariate Imputation by Chained Equations in R. *J Stat Soft.* 2011;45(3). doi:10.18637/jss.v045.i03
89. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2
90. Leisman DE, Harhay MO, Lederer DJ, et al. Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. *Critical Care Medicine.* 2020;48(5):623-633. doi:10.1097/CCM.0000000000004246
91. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 5. [Dr.]. Springer; 2010.
92. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Comm in Stats - Theory & Methods.* 1980;9(10):1043-1069. doi:10.1080/03610928008827941
93. Harrell FE. Evaluating the Yield of Medical Tests. *JAMA.* 1982;247(18):2543. doi:10.1001/jama.1982.03320430047030
94. Wolfson J, Vock DM, Bandyopadhyay S, et al. Use and Customization of Risk Scores for Predicting Cardiovascular Events Using Electronic Health Record Data. *J Am Heart Assoc.* 2017;6(4). doi:10.1161/JAHA.116.003670
95. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics.* 1979;28(1):100. doi:10.2307/2346830
96. University of Cincinnati. K-means Cluster Analysis. *UC Business Analytics R Programming Guide*. Accessed December 19, 2022. [https://uc-r.github.io/kmeans\\_clustering](https://uc-r.github.io/kmeans_clustering)
97. Müllner D. **fastcluster** : Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *J Stat Soft.* 2013;53(9). doi:10.18637/jss.v053.i09
98. Szepannek G. clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *The R Journal.* 2019;10(2):200. doi:10.32614/RJ-2018-048

- 
99. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. Published online November 24, 2017. Accessed June 7, 2023. <http://arxiv.org/abs/1705.07874>
  100. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x
  101. Bullock GS, Hughes T, Arundale AH, Ward P, Collins GS, Kluzek S. Black Box Prediction Methods in Sports Medicine Deserve a Red Card for Reckless Practice: A Change of Tactics is Needed to Advance Athlete Care. *Sports Med*. 2022;52(8):1729-1735. doi:10.1007/s40279-022-01655-6
  102. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
  103. De Hond AAH, Steyerberg EW, Van Calster B. Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*. 2022;4(12):e853-e855. doi:10.1016/S2589-7500(22)00188-1
  104. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240(4857):1285-1293. doi:10.1126/science.3287615
  105. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean J Anesthesiol*. 2022;75(1):25-36. doi:10.4097/kja.21209
  106. Muller MP, Tomlinson G, Marrie TJ, et al. Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia? *Clin Infect Dis*. 2005;40(8):1079-1086. doi:10.1086/428577
  107. Maiga A, Farjah F, Blume J, et al. Risk Prediction in Clinical Practice: A Practical Guide for Cardiothoracic Surgeons. *Ann Thorac Surg*. 2019;108(5):1573-1582. doi:10.1016/j.athoracsur.2019.04.126
  108. 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. doi:10.1093/jpepsy/jst062
  109. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; 2003.
  110. Kernbach JM, Staartjes VE. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II-Generalization and Overfitting. *Acta Neurochir Suppl*. 2022;134:15-21. doi:10.1007/978-3-030-85292-4\_3
  111. White N, Parsons R, Collins G, Barnett A. Evidence of questionable research practices in clinical prediction models. *BMC Med*. 2023;21(1):339. doi:10.1186/s12916-023-03048-6

112. Berk AN, Piasecki DP, Fleischli JE, Trofa DP, Saltzman BM. Trends in Patient-Reported Outcomes After Anterior Cruciate Ligament Reconstruction: A Systematic Review. *Orthop J Sports Med.* 2023;11(5):23259671231174472. doi:10.1177/23259671231174472
113. Vemulapalli KV, Sunil Kumar KH, Khanduja V. Registry Studies Use Inconsistent Methods to Account for Patients Lost to Follow-up, and Rates of Patients LTFU Are High. *Arthroscopy, Sports Medicine, and Rehabilitation.* 2021;3(6):e1607-e1619. doi:10.1016/j.asmr.2021.07.016
114. Ueland TE, Carreira DS, Martin RL. Substantial Loss to Follow-Up and Missing Data in National Arthroscopy Registries: A Systematic Review. *Arthroscopy: The Journal of Arthroscopic & Related Surgery.* 2021;37(2):761-770.e3. doi:10.1016/j.arthro.2020.08.007
115. Visnes H, Gifstad T, Persson A, et al. ACL Reconstruction Patients Have Increased Risk of Knee Arthroplasty at 15 Years of Follow-up: Data from the Norwegian Knee Ligament Register and the Norwegian Arthroplasty Register from 2004 to 2020. *JBJS Open Access.* 2022;7(2). doi:10.2106/JBJS.OA.22.00023
116. Moatshe G, LaPrade CM, Fenstad AM, et al. Rates of Subjective Failure After Both Isolated and Combined Posterior Cruciate Ligament Reconstruction: A Study From the Norwegian Knee Ligament Registry 2004-2021. *Am J Sports Med.* 2024;52(6):1491-1497. doi:10.1177/03635465241238461
117. Aga C, Risberg MA, Fagerland MW, et al. No Difference in the KOOS Quality of Life Subscore Between Anatomic Double-Bundle and Anatomic Single-Bundle Anterior Cruciate Ligament Reconstruction of the Knee: A Prospective Randomized Controlled Trial With 2 Years' Follow-up. *Am J Sports Med.* 2018;46(10):2341-2354. doi:10.1177/0363546518782454
118. Soreide E, Granan LP, Hjorthaug GA, Espehaug B, Dimmen S, Nordsletten L. The Effect of Limited Perioperative Nonsteroidal Anti-inflammatory Drugs on Patients Undergoing Anterior Cruciate Ligament Reconstruction. *Am J Sports Med.* 2016;44(12):3111-3118. doi:10.1177/0363546516657539
119. Ingelsrud LH, Granan LP, Terwee CB, Engebretsen L, Roos EM. Proportion of Patients Reporting Acceptable Symptoms or Treatment Failure and Their Associated KOOS Values at 6 to 24 Months After Anterior Cruciate Ligament Reconstruction: A Study From the Norwegian Knee Ligament Registry. *Am J Sports Med.* 2015;43(8):1902-1907. doi:10.1177/0363546515584041
120. Zsidai B, Narup E, Senorski EH, et al. The Knee Injury and Osteoarthritis Outcome Score: shortcomings in evaluating knee function in persons undergoing ACL reconstruction. *Knee Surg Sports Traumatol Arthrosc.* 2022;30(11):3594-3598. doi:10.1007/s00167-022-06990-9
121. 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 Relat Res.* 2022;480(7):1342-1350. doi:10.1097/CORR.0000000000002158

- 
122. Hansen CF, Jensen J, Odgaard A, et al. Four of five frequently used orthopedic PROMs possess inadequate content validity: a COSMIN evaluation of the mHHS, HAGOS, IKDC-SKF, KOOS and KNEES-ACL. *Knee Surg Sports Traumatol Arthrosc.* 2022;30(11):3602-3615. doi:10.1007/s00167-021-06761-y
123. Mabrouk A, Nwachukwu B, Pareek A, et al. MCID and PASS in Knee Surgeries. Theoretical Aspects and Clinical Relevance References. *Knee Surg Sports Traumatol Arthrosc.* 2023;31(6):2060-2067. doi:10.1007/s00167-023-07359-2
124. Urhausen AP, Grindem H, Engebretsen L, et al. The Delaware-Oslo ACL Cohort treatment algorithm yields superior outcomes to usual care 9-12 years after anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc.* 2024;32(2):214-222. doi:10.1002/ksa.12039
125. Vega JF, Jacobs CA, Strnad GJ, et al. Prospective Evaluation of the Patient Acceptable Symptom State to Identify Clinically Successful Anterior Cruciate Ligament Reconstruction. *Am J Sports Med.* 2019;47(5):1159-1167. doi:10.1177/0363546519831008
126. Roos EM, Boyle E, Frobell RB, Lohmander LS, Ingelsrud LH. It is good to feel better, but better to feel good: whether a patient finds treatment “successful” or not depends on the questions researchers ask. *Br J Sports Med.* 2019;53(23):1474-1478. doi:10.1136/bjsports-2018-100260
127. Muller B, Yabroudi MA, Lynch A, et al. Defining Thresholds for the Patient Acceptable Symptom State for the IKDC Subjective Knee Form and KOOS for Patients Who Underwent ACL Reconstruction. *Am J Sports Med.* 2016;44(11):2820-2826. doi:10.1177/0363546516652888
128. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. *Ann Rheum Dis.* 2005;64(1):34-37. doi:10.1136/ard.2004.023028
129. Tubach F, Dougados M, Falissard B, Baron G, Logeart I, Ravaud P. Feeling good rather than feeling better matters more to patients. *Arthritis & Rheumatism.* 2006;55(4):526-530. doi:10.1002/art.22110
130. Cristiani R, Forssblad M, Edman G, Eriksson K, Ståhlman A. Age, time from injury to surgery and quadriceps strength affect the risk of revision surgery after primary ACL reconstruction. *Knee Surg Sports Traumatol Arthrosc.* 2021;29(12):4154-4162. doi:10.1007/s00167-021-06517-8
131. 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. doi:10.1016/j.arthro.2021.12.030
132. Bernholt DL, Dornan GJ, DePhillipo NN, Aman ZS, Kennedy MI, LaPrade RF. High-Grade Posterolateral Tibial Plateau Impaction Fractures in the Setting of a Primary Anterior Cruciate Ligament Tear Are Correlated With an Increased Preoperative Pivot



- Shift and Inferior Postoperative Outcomes After Anterior Cruciate Ligament Reconstruction. *Am J Sports Med.* 2020;48(9):2185-2194. doi:10.1177/0363546520932912
133. Bayer S, Meredith SJ, Wilson KW, et al. Knee Morphological Risk Factors for Anterior Cruciate Ligament Injury: A Systematic Review. *Journal of Bone and Joint Surgery.* 2020;102(8):703-718. doi:10.2106/JBJS.19.00535
  134. Li Y, Hong L, Feng H, et al. Posterior Tibial Slope Influences Static Anterior Tibial Translation in Anterior Cruciate Ligament Reconstruction: A Minimum 2-Year Follow-up Study. *Am J Sports Med.* 2014;42(4):927-933. doi:10.1177/0363546514521770
  135. Bernhardtson AS, Aman ZS, Dornan GJ, et al. Tibial Slope and Its Effect on Force in Anterior Cruciate Ligament Grafts: Anterior Cruciate Ligament Force Increases Linearly as Posterior Tibial Slope Increases. *Am J Sports Med.* 2019;47(2):296-302. doi:10.1177/0363546518820302
  136. 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. doi:10.1007/s00167-019-05770-2
  137. Roe C, Jacobs C, Kline P, et al. Correlations of Single-Leg Performance Tests to Patient-Reported Outcomes After Primary Anterior Cruciate Ligament Reconstruction. *Clin J Sport Med.* 2021;31(5):e265-e270. doi:10.1097/JSM.0000000000000780
  138. Liu A, Sun M, Ma C, et al. Clinical outcomes of transtibial versus anteromedial drilling techniques to prepare the femoral tunnel during anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc.* 2017;25(9):2751-2759. doi:10.1007/s00167-015-3672-y
  139. Inderhaug E, Raknes S, Østvold T, Solheim E, Strand T. Increased revision rate with posterior tibial tunnel placement after using the 70-degree tibial guide in ACL reconstruction. *Knee Surg Sports Traumatol Arthrosc.* 2017;25(1):152-158. doi:10.1007/s00167-016-4341-5
  140. MOON Knee Group, Spindler KP, Huston LJ, et al. Anterior Cruciate Ligament Reconstruction in High School and College-Aged Athletes: Does Autograft Choice Influence Anterior Cruciate Ligament Revision Rates? *Am J Sports Med.* 2020;48(2):298-309. doi:10.1177/0363546519892991
  141. Marmura H, Getgood AMJ, Spindler KP, Kattan MW, Briskin I, Bryant DM. Validation of a Risk Calculator to Personalize Graft Choice and Reduce Rupture Rates for Anterior Cruciate Ligament Reconstruction. *Am J Sports Med.* 2021;49(7):1777-1785. doi:10.1177/03635465211010798
  142. Usami S, Kimura Y, Matsuzaka M, et al. A new neural network model that detects graft ruptures and contralateral anterior cruciate ligament injuries. *Knee Surg Sports Traumatol Arthrosc.* 2024;32(4):872-880. doi:10.1002/ksa.12123
  143. Ye Z, Zhang T, Wu C, et al. Predicting the Objective and Subjective Clinical Outcomes of Anterior Cruciate Ligament Reconstruction: A Machine Learning Analysis of 432 Patients. *Am J Sports Med.* 2022;50(14):3786-3795. doi:10.1177/03635465221129870

144. Zhang T, Ye Z, Cai J, et al. Ensemble Algorithm for Risk Prediction of Clinical Failure After Anterior Cruciate Ligament Reconstruction. *Orthop J Sports Med*. 2024;12(8):23259671241261695. doi:10.1177/23259671241261695
145. Kunze KN, Polce EM, Ranawat AS, et al. Application of Machine Learning Algorithms to Predict Clinically Meaningful Improvement After Arthroscopic Anterior Cruciate Ligament Reconstruction. *Orthopaedic Journal of Sports Medicine*. 2021;9(10):23259671211046575. doi:10.1177/23259671211046575
146. Oetl FC, Pareek A, Winkler PW, et al. A practical guide to the implementation of AI in orthopaedic research, Part 6: How to evaluate the performance of AI research? *J exp orthop*. 2024;11(3):e12039. doi:10.1002/jeo2.12039
147. Spencer L, Burkhart TA, Tran MN, et al. Biomechanical analysis of simulated clinical testing and reconstruction of the anterolateral ligament of the knee. *Am J Sports Med*. 2015;43(9):2189-2197. doi:10.1177/0363546515589166
148. Hewison CE, Tran MN, Kaniki N, Remtulla A, Bryant D, Getgood AM. Lateral Extra-articular Tenodesis Reduces Rotational Laxity When Combined With Anterior Cruciate Ligament Reconstruction: A Systematic Review of the Literature. *Arthroscopy*. 2015;31(10):2022-2034. doi:10.1016/j.arthro.2015.04.089
149. Getgood A, Brown C, Lording T, et al. The anterolateral complex of the knee: results from the International ALC Consensus Group Meeting. *Knee Surg Sports Traumatol Arthrosc*. 2019;27(1):166-176. doi:10.1007/s00167-018-5072-6
150. Getgood A, Hewison C, Bryant D, et al. No Difference in Functional Outcomes When Lateral Extra-Articular Tenodesis Is Added to Anterior Cruciate Ligament Reconstruction in Young Active Patients: The Stability Study. *Arthroscopy*. 2020;36(6):1690-1701. doi:10.1016/j.arthro.2020.02.015
151. Groot OQ, Bindels BJJ, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop*. 2021;92(4):385-393. doi:10.1080/17453674.2021.1910448
152. Diermeier TA, Rothrauff BB, Engebretsen L, et al. Treatment after ACL injury: Panther Symposium ACL Treatment Consensus Group. *Br J Sports Med*. 2021;55(1):14-22. doi:10.1136/bjsports-2020-102200
153. Meunier A, Odensten M, Good L. Long-term results after primary repair or non-surgical treatment of anterior cruciate ligament rupture: a randomized study with a 15-year follow-up. *Scandinavian Med Sci Sports*. 2007;17(3):230-237. doi:10.1111/j.1600-0838.2006.00547.x
154. International Olympic Committee Pediatric ACL Injury Consensus Group, Ardern CL, Ekås G, et al. 2018 International Olympic Committee Consensus Statement on Prevention, Diagnosis, and Management of Pediatric Anterior Cruciate Ligament Injuries. *Orthopaedic Journal of Sports Medicine*. 2018;6(3):232596711875995. doi:10.1177/2325967118759953
155. Ekås GR, Laane MM, Larmo A, et al. Knee Pathology in Young Adults After Pediatric Anterior Cruciate Ligament Injury: A Prospective Case Series of 47 Patients With a Mean

- 9.5-Year Follow-up. *Am J Sports Med.* 2019;47(7):1557-1566. doi:10.1177/0363546519837935
156. Ekås GR, Moksnes H, Grindem H, Risberg MA, Engebretsen L. Coping With Anterior Cruciate Ligament Injury From Childhood to Maturation: A Prospective Case Series of 44 Patients With Mean 8 Years' Follow-up. *Am J Sports Med.* Published online November 26, 2018;363546518810750. doi:10.1177/0363546518810750
  157. Moksnes H, Engebretsen L, Eitzen I, Risberg MA. Functional outcomes following a non-operative treatment algorithm for anterior cruciate ligament injuries in skeletally immature children 12 years and younger. A prospective cohort with 2 years follow-up. *Br J Sports Med.* 2013;47(8):488-494. doi:10.1136/bjsports-2012-092066
  158. Lohmander LS, Roemer FW, Frobell RB, Roos EM. Treatment for Acute Anterior Cruciate Ligament Tear in Young Active Adults. *NEJM Evid.* 2023;2(8):EVIDoa2200287. doi:10.1056/EVIDoa2200287
  159. Frobell RB, Roos HP, Roos EM, Roemer FW, Ranstam J, Lohmander LS. Treatment for acute anterior cruciate ligament tear: five year outcome of randomised trial. *Br J Sports Med.* 2015;49(10):700. doi:10.1136/bjsports-2014-f232rep
  160. De Jonge R, Máté M, Kovács N, et al. Nonoperative Treatment as an Option for Isolated Anterior Cruciate Ligament Injury: A Systematic Review and Meta-analysis. *Orthopaedic Journal of Sports Medicine.* 2024;12(4):23259671241239665. doi:10.1177/23259671241239665
  161. Cinque ME, Dornan GJ, Chahla J, Moatshe G, LaPrade RF. High Rates of Osteoarthritis Develop After Anterior Cruciate Ligament Surgery: An Analysis of 4108 Patients. *Am J Sports Med.* 2018;46(8):2011-2019. doi:10.1177/0363546517730072
  162. Ferrero S, Louvois M, Barnetche T, Breuil V, Roux C. Impact of anterior cruciate ligament surgery on the development of knee osteoarthritis: A systematic literature review and meta-analysis comparing non-surgical and surgical treatments. *Osteoarthritis and Cartilage Open.* 2023;5(3):100366. doi:10.1016/j.ocarto.2023.100366
  163. Von Porat A. High prevalence of osteoarthritis 14 years after an anterior cruciate ligament tear in male soccer players: a study of radiographic and patient relevant outcomes. *Annals of the Rheumatic Diseases.* 2004;63(3):269-273. doi:10.1136/ard.2003.008136
  164. 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. doi:10.1177/2325967116638341
  165. Webster KE, Feller JA. Psychological Readiness to Return to Sport After Anterior Cruciate Ligament Reconstruction in the Adolescent Athlete. *Journal of Athletic Training.* 2022;57(9-10):955-960. doi:10.4085/1062-6050-0543.21
  166. Webster KE, Nagelli CV, Hewett TE, Feller JA. Factors Associated With Psychological Readiness to Return to Sport After Anterior Cruciate Ligament Reconstruction Surgery. *Am J Sports Med.* 2018;46(7):1545-1550. doi:10.1177/0363546518773757

167. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg*. 2011;128(1):305-310. doi:10.1097/PRS.0b013e318219c171
168. Ley C, Martin RK, Pareek A, Groll A, Seil R, Tischler T. Machine learning and conventional statistics: making sense of the differences. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(3):753-757. doi:10.1007/s00167-022-06896-6
169. 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. doi:10.1097/RLI.0000000000000615
170. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056
171. 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. doi:10.1016/j.arth.2021.02.026
172. 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. *Radiology: Artificial Intelligence*. 2021;3(2):e200198. doi:10.1148/ryai.2020200198
173. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577(7792):706-710. doi:10.1038/s41586-019-1923-7
174. 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. doi:10.1007/s00256-018-3016-3
175. 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. doi:10.1080/17453674.2020.1803664
176. Goddard J. Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers. *Am J Med*. Published online June 25, 2023:S0002-9343(23)00401-1. doi:10.1016/j.amjmed.2023.06.012
177. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel)*. 2023;11(6):887. doi:10.3390/healthcare11060887
178. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. *J Med Syst*. 2024;48(1):22. doi:10.1007/s10916-024-02045-3

179. 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. doi:10.1007/s00167-021-06838-8



## Papers I-VI





## Paper I

Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Engebretsen L. Predicting Anterior Cruciate Ligament Reconstruction Revision: A Machine Learning Analysis Utilizing the Norwegian Knee Ligament Register. *J Bone Joint Surg Am.* 2022;104(2):145-153. doi:10.2106/JBJS.21.00113



# Predicting Anterior Cruciate Ligament Reconstruction Revision

## A Machine Learning Analysis Utilizing the Norwegian Knee Ligament Register

R. Kyle Martin, MD, FRCSC, Solvejg Wastvedt, BA, Ayoosh Pareek, MD, Andreas Persson, MD, PhD, Håvard Visnes, MD, PhD, Anne Marie Fenstad, MS, Gilbert Moatshe, MD, PhD, Julian Wolfson, PhD, and Lars Engebretsen, MD, PhD

*Investigation performed at the University of Minnesota, Minneapolis, Minnesota*

**Background:** Several factors are associated with an increased risk of anterior cruciate ligament (ACL) reconstruction revision. However, the ability to accurately translate these factors into a quantifiable risk of revision at a patient-specific level has remained elusive. We sought to determine if machine learning analysis of the Norwegian Knee Ligament Register (NKLK) can identify the most important risk factors associated with subsequent revision of primary ACL reconstruction and develop a clinically meaningful calculator for predicting revision of primary ACL reconstruction.

**Methods:** Machine learning analysis was performed on the NKLK data set. The primary outcome was the probability of revision ACL reconstruction within 1, 2, and/or 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, generalized additive model, and gradient boosted regression. Concordance and calibration were calculated for all 4 models.

**Results:** The data set included 24,935 patients, and 4.9% underwent a revision surgical procedure during a mean follow-up (and standard deviation) of  $8.1 \pm 4.1$  years. All 4 models were well-calibrated, with moderate concordance (0.67 to 0.69). The Cox Lasso model required only 5 variables for outcome prediction. The other models either used more variables without an appreciable improvement in accuracy or had slightly lower accuracy overall. An in-clinic calculator was developed that can estimate the risk of ACL revision (Revision Risk Calculator). This calculator can quantify risk at a patient-specific level, with a plausible range from near 0% for low-risk patients to 20% for high-risk patients at 5 years.

**Conclusions:** Machine learning analysis of a national knee ligament registry can predict the risk of ACL reconstruction revision with moderate accuracy. This algorithm supports the creation of an in-clinic calculator for point-of-care risk stratification based on the input of only 5 variables. Similar analysis using a larger or more comprehensive data set may improve the accuracy of risk prediction, and future studies incorporating patients who have experienced failure of ACL reconstruction but have not undergone subsequent revision may better predict the true risk of ACL reconstruction failure.

**Level of Evidence:** Prognostic Level III. See Instructions for Authors for a complete description of levels of evidence.

The anterior cruciate ligament (ACL) is one of the main knee stabilizers, and its rupture can lead to pain, instability, and functional limitation<sup>1</sup>. Injury rates have been rising globally, and surgical reconstruction of the ACL is often performed to restore normal biomechanics and to improve knee stability<sup>2-5</sup>. Recent studies have associated several

factors with an increased risk of failed surgical reconstruction<sup>6-14</sup>. However, due to the complex relationships between these various factors, accurate prediction and quantification of patient-specific risk are challenging.

A novel approach to health-care research, machine learning, has the potential to improve our predictive capability.

**Disclosure:** The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (<http://links.lww.com/JBJS/G758>).

Copyright © 2021 The Authors. Published by the Journal of Bone and Joint Surgery, Incorporated. All rights reserved. This is an open-access article distributed under the terms of the [Creative Commons Attribution-Non Commercial-No Derivatives License 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/) (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Machine learning refers to a set of techniques that model complex relationships between variables in order to predict an outcome. Relationships can be more complex than those assessed with traditional statistical techniques. Although applications of machine learning in sports medicine have been slow to develop, machine learning has broadly impacted the medical field, including within orthopaedic surgery<sup>13,16</sup>. Established in 2004, the Norwegian Knee Ligament Register (NKLK) contains demographic, injury, surgical, and outcome data on >25,000 patients. The NKLK has produced many studies since its inception that have impacted our understanding of ACL injuries<sup>11,12,17,18</sup>, and the application of machine learning presents the opportunity to further evaluate factors associated with outcome.

Previous studies into the risk factors for ACL reconstruction failure have assessed the strength of association (effect measure) and the probability of seeing results at least as strong as those that were observed if there is no true association between the independent and the dependent variables. This has resulted in the identification of numerous factors associated with outcome such as age, sex, graft choice, fixation method, body mass index (BMI), and return to pivoting sports<sup>11,12,19-21</sup>. Although traditional statistical models require human selection of variables thought to be of importance, machine learning allows a computer to consider all possible combinations and interactions of variables contained in a data set and their relationships to the outcome of interest. The machine learning analysis can identify which factors from this much larger pool are focal in predicting the outcome. As with traditional methods, machine learning can develop an algorithm to predict the outcome for future patients. However, more complex interactions and relationships can be used in machine learning predictive algorithms, which may yield more accurate and patient-specific predictive capability.

An accurate predictive model for clinical outcome following ACL reconstruction would be beneficial for both the orthopaedic surgeon and the patient. This would allow patient and surgical information to guide shared clinical decision-making with regard to patient-specific management. There are currently no machine learning-driven models to predict outcome after ACL reconstruction based on national knee ligament registry data. The purpose of this study was therefore to use machine learning analysis of the NKLK to identify the most important risk factors associated with subsequent revision of primary ACL reconstruction and develop a clinically meaningful model for predicting primary ACL reconstruction revision. The hypothesis was that machine learning analysis would enable accurate prediction of revision risk for a patient undergoing a primary ACL reconstruction.

## Materials and Methods

### Data Preparation

Patients contained within the NKLK with primary ACL reconstruction surgery dates from January 2004 through December 2018 were included. Those with missing values for graft choice were excluded. All variables captured by the reg-

ister were considered for the analysis. We recoded or defined new variables for the following: years between the injury and the surgical procedure, meniscus injury identified at the surgical procedure, any additional injury identified at the surgical procedure, choice of graft (patellar tendon autograft, hamstring tendon autograft, other), and height and weight variables that combined data from patient and surgeon-reported variables. Time to revision was calculated as the number of years from the primary surgical procedure to revision. For assessing concordance at specific follow-up times, we considered patients with a revision at or prior to the time point as experiencing the event. We also created a predictor indicating if a patient was below the median score in all 4 Knee Injury and Osteoarthritis Outcome Score (KOOS) categories at the time of the primary surgical procedure and scaled predictors for KOOS Quality of Life (QoL) and Sports measures to a score of 10. The final list of predictor variables included for analysis is presented in Table I.

### Model Creation

The primary outcome was the probability of revision ACL reconstruction within 1, 2, and/or 5 years. We randomly split the cleaned data into training sets (75%) that were used to fit the models and test sets (25%) that were used to evaluate the models. We used R (version 3.6.1; The R Foundation for Statistical Computing) to fit several machine learning models to the training data<sup>22</sup>. All models and their performance measures described below account for censoring of our time-to-event outcome. "Censoring" means that, at any given follow-up time, we do not have complete information on the outcome for all patients. This is because some patients have not been in the registry for the requisite number of years, and others have not yet experienced revision and it is unknown when or if they ultimately will. Four models intended for this type of data were tested: Cox Lasso, survival random forest, generalized additive model (GAM), and gradient boosted regression model (GBM). These models are among the most commonly used in machine learning. The Cox Lasso model is a semiparametric, penalized regression model that selects a subset of variables for inclusion<sup>23</sup>. The survival random forest model is a tree-based, nonparametric method adapted for right-censored data such as ours<sup>24</sup>. GBMs are also nonparametric, meaning that they do not require prespecification of a model structure, and iteratively improve the model fit using all available variables<sup>25,26</sup>. GAMs allow for machine-selected nonlinear relationships among a prespecified group of variables<sup>27</sup>. Further details on each model are included in Appendix A.

We applied the L1-regularized Cox model ("Cox Lasso," package *glmnet*; lambda value selected via cross-validation) to select variables and retained those with nonzero coefficients, shown in the top left of Figure 1. We trained a survival random forest (function *rfsrc* from package *randomForestSRC*) with node size 200, 10 variables tried per split, 100 trees, and the full set of predictors (Table I). We trained a GAM (function *gam* from package *mgcv*) with those variables selected in the Cox Lasso, using a smooth term for the years from injury to surgery predictor. Finally, we trained a GBM (functions *gbm* and *basehaz.gbm*

**TABLE I Characteristics of the Registry Population and Variables Considered for Machine Learning Analysis**

Characteristic or Variable*	Values (N = 24,935)
Age	
At surgery† (yr)	28 ± 11
At injury† (yr)	27 ± 10
Missing data‡	1,251 (5%)
Sex‡	
Male	14,019 (56%)
Female	10,916 (44%)
BMI† (kg/m <sup>2</sup> )	25.0 ± 3.8
Missing data‡	7,920 (32%)
KOOS QoL at primary surgery†	3.49 ± 1.86
Missing data‡	5,149 (21%)
KOOS Sports at primary surgery†	4.28 ± 2.73
Missing data‡	5,324 (21%)
Below median on all KOOS subscales‡	
Yes	3,972 (16%)
No	15,982 (64%)
Missing data	4,981 (20%)
Hospital geographic region‡	
Southeast	9,335 (37%)
West	3,974 (16%)
Central	2,162 (8.7%)
North	958 (3.8%)
Missing data	8,506 (34%)
Hospital type‡	
Public	16,429 (66%)
Private	8,506 (34%)
Injury‡	
Meniscus	13,145 (53%)
Cartilage	5,801 (23%)
Any	171 (0.7%)
Posterior cruciate ligament	398 (1.6%)
Medial collateral ligament	1,993 (8.0%)
Lateral collateral ligament	464 (1.9%)
Posterolateral corner	243 (1.0%)
Missing data	2,720 (10.9%)
Graft choice‡	
Bone-patellar tendon-bone autograft	9,891 (40%)
Hamstring autograft	14,481 (58%)
Unknown or other	563 (2.3%)
Tibial fixation device‡	
Interference screw	19,283 (77%)
Suspension or cortical device	2,367 (9.5%)
Unknown or other	3,285 (13%)

*continued***TABLE I (continued)**

Characteristic or Variable*	Values (N = 24,935)
Femoral fixation device‡	
Interference screw	8,287 (33%)
Suspension or cortical device	13,072 (52%)
Unknown or other	3,576 (14%)
Fixation device combination‡	
2 interference screws	8,086 (32%)
Interference or suspension	154 (0.6%)
2 suspension or cortical devices	1,809 (7.3%)
Suspension or interference	9,725 (39%)
Unknown or other	5,161 (21%)
Injured side‡	
Right	12,675 (51%)
Left	12,260 (49%)
Previous surgical procedure‡	
On contralateral knee	1,804 (7.2%)
On ipsilateral knee	4,213 (17%)
Time from injury to primary surgery† (yr)	1.63 ± 3.26
Missing data‡	1,255 (5%)
Systemic antibiotic prophylaxis‡	
Yes	24,769 (99%)
No	108 (0.4%)
Missing data	58 (0.2%)

\*All variables represent patient demographic characteristics, injury, patient-reported outcome scores, or surgical details at the time of the primary ACL reconstruction. †The values are given as the mean and the standard deviation. ‡The values are given as the number of patients, with the percentage in parentheses.

from package *gbm*), using the full set of predictors, a shrinkage parameter of 0.001, and 6,550 trees (number of trees selected via cross-validation). To maximize accuracy for the tree-based methods, we used a finer grouping for fixation device variables (Supplementary Tables 1a, 1b, and 1c). To achieve a more direct comparison between the models using variable selection and those using the full set of predictors, we also trained the random forest and GBM using only predictors selected in the Cox Lasso. All 4 models were restricted to patients with complete data for the predictors used (see Table II and Missing Data section below).

#### Model Evaluation

We evaluated model performance by calculating predicted survival probabilities for the held-out test data using the trained models. Model calibration was assessed using a version of the Hosmer-Lemeshow statistic that accounts for censoring<sup>28</sup>. Calibration refers to the accuracy of the risk estimates, comparing the expected outcomes with the actual observed outcomes. This statistic sums the average misclassification in each predicted risk

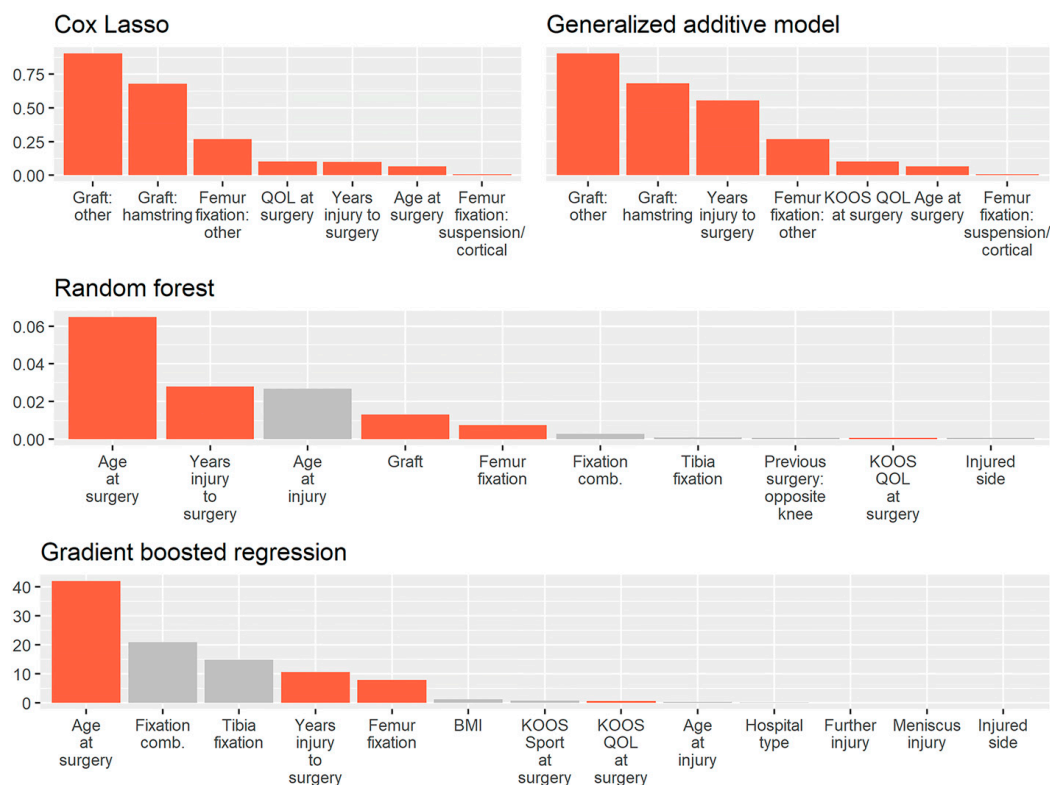


Fig. 1

Feature importance. The 4 plots show relative feature importance in each of the machine learning models. The highlighted bars indicate features selected into the Cox Lasso model. The random forest plot shows variables with importance of  $>0.0005$  and the gradient boosted plot shows variables with importance of  $>0$ , for readability. The orange bars represent variables selected as important in the Cox Lasso model, and the gray bars represent the other variables used in the models.

quintile and converts the sum into a chi-square statistic. Larger calibration statistics correspond to smaller p values, and significance means that the null hypothesis of perfect calibration is rejected. Concordance was calculated using the Harrell C-index<sup>29</sup> at 1, 2, and 5-year follow-up times. The C-index measures the proportion of ranked pairs of observations in which the predicted ranking corresponds with true outcomes. It is a generalization of the area under the curve (AUC) appropriate for censored data, where not all patients have completed the follow-up time. As with the AUC, the C-index ranges from 0 to 1, with 1 indicating perfect concordance.

#### Missing Data

To assess the impact of restricting data to complete cases, we retrained the models using multiple imputation. This common statistical technique fills in a patient's missing data values based on characteristics of other patients in the population. Because

our population had nontrivial missing data on several variables, multiple imputation allowed us to gauge the reasonableness of excluding these incomplete observations. We conducted multivariate imputation by chained equations (MICE) with 5 imputations on both training and test data (function *mice* from package *mice*). Using the variables with nonzero coefficients in the complete-case Cox Lasso, we refit the Cox model on

TABLE II Proportion of Complete Cases by Model

Model	Total Cases	Incomplete Cases	Complete Cases
Cox Lasso and GAM	24,935	6,048	76%
Random forest and GBM	24,935	11,663	53%

TABLE III Description of Censoring

Follow-up Time	Patients with Revision*	Patients with Complete Follow-up and No Revision*	Patients with Incomplete Follow-up and No Revision*†
1 year	190 (0.8%)	22,908 (91.9%)	1,837 (7.4%)
2 years	529 (2.1%)	20,703 (83.0%)	3,703 (14.9%)
5 years	999 (4.0%)	15,107 (60.6%)	8,829 (35.4%)

\*The values are given as the number of patients, with the percentage in parentheses. †This category represents patients who have not yet reached the specified end point.

imputed data, averaging predictions over the 5 imputations. We similarly refit the GAM and the GBM. For the random forest model, imputation was done using the adaptive tree imputation algorithm of Ishwaran et al.<sup>24</sup>, as implemented in the *rfsrc* function from the *randomForestSRC* R package. We maintained the default of 1 iteration of the algorithm for imputing training data. Supplementary Tables 2a through 2d show model performance with imputation on training data only and training and test data.

#### Source of Funding

This study was funded by the Norwegian Arthroplasty & Knee Ligament Register, the University of Oslo School of Medicine, and a Norwegian Centennial Chair seed grant. Funding supported the machine learning analysis and interpretation. The funding agencies had no direct role in the investigation.

## Results

#### Data Characteristics

Table I describes characteristics of the registry population at the time of the primary surgical procedure and the varia-

bles included for analysis. After data cleaning (5 patients were excluded for missing graft choice), 24,935 patients met the inclusion criteria; of these patients, 1,219 (4.9%) underwent a revision surgical procedure during a mean follow-up period (and standard deviation) of  $8.1 \pm 4.1$  years. Table III presents the proportion of patients with complete follow-up at each of the 3 time points. The population was predominantly male (56%), with a mean age of  $27 \pm 10$  years at the time of the primary injury and  $28 \pm 11$  years at the time of the surgical procedure.

To assess the potential impact of missing data on our results, we compared covariate distributions between complete cases and the full data set (Supplementary Tables 1a, 1b, and 1c). Although the large sample size results in the complete cases and the full data set being significantly different ( $p < 0.05$ ) on multiple variables, the magnitudes of the between-group differences were generally small and not clinically meaningful.

#### Model Performance

All 4 models were generally well-calibrated, with concordance in the moderate range (0.67 to 0.69). Only the 2-year

TABLE IV Model Performance Measures

Model	Concordance	Calibration Statistic	Calibration P Value
Probability of revision: 1 year			
Cox Lasso	0.686	4.89	0.18
Random forest	0.672	3.12	0.374
GAM	0.687	4.79	0.188
GBM	0.669	4.98	0.174
Probability of revision: 2 years			
Cox Lasso	0.684	11.35	0.01
Random forest	0.67	11.66	0.009
GAM	0.685	11.19	0.011
GBM	0.666	3.76	0.288
Probability of revision: 5 years			
Cox Lasso	0.683	6.19	0.103
Random forest	0.67	3.71	0.295
GAM	0.684	6.98	0.073
GBM	0.665	0.38	0.944

TABLE V Randomly Selected Example Patients from 3 Predicted 5-Year Risk Groups\*

Variable	Low-Risk Patients	Medium-Risk Patients	High-Risk Patients
Age (yr)	39	15	15
KOOS QoL at primary surgery (points)	25	25	6
Graft choice	Hamstring autograft	Bone-patellar tendon-bone autograft	Hamstring autograft
Femoral fixation device	Suspension or cortical device	Interference screw	Suspension or cortical device
Time between injury and primary surgery (mo)	14	9	8
Risk of revision			
At 1 year	0.5%	1.2%	2.8%
At 2 years	1.4%	3.6%	8.5%
At 5 years	2.8%	7.2%	17.2%

\*Low (<5%), medium (between 5% and 15%), and high (>15%). The patients' values for each variable used in the Cox model are given, along with the Cox model-predicted risk of revision at 1, 2, and 5 years.

Cox Lasso model, random forest model, and GAM had calibration p values between 0.01 and 0.05, suggesting modest evidence of miscalibration (Table IV). The GBM had a small edge in calibration for 2-year and 5-year follow-up times. However, concordance was slightly lower for the GBM and the random forest model at all follow-up times (0.67 compared with 0.68).

Imputing missing data did not significantly improve performance for any of the models (Supplementary Tables 2a through 2d). When the random forest model and the GBM were restricted to the Cox Lasso predictors, calibration worsened substantially when limited to complete cases and stayed about the same under imputation. Concordance was virtually unchanged (Supplementary Tables 3a and 3b).

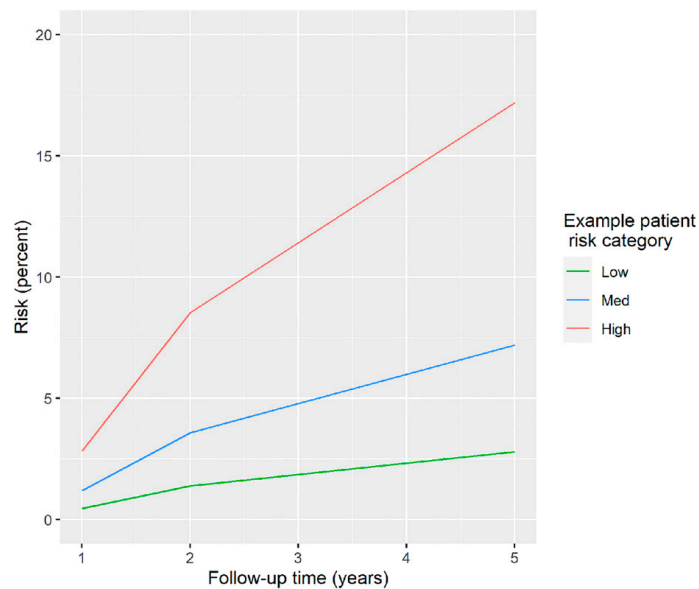


Fig. 2

Risk of revision ACL reconstruction in 3 randomly selected example patients corresponding with Table V.



### Factors Predicting Outcome

The most important predictors for revision in the Cox Lasso model, in order, were graft choice, femoral fixation device, KOOS QoL at the time of the surgical procedure, years from the injury to the surgical procedure, and age at the time of the surgical procedure. In the random forest model, predictors in the top third by variable importance score also included age at the time of the injury, tibial fixation device, and fixation device combination. The most important features in both the GAM and the GBM were relatively similar to those in the Cox Lasso model. The Cox Lasso model and the GAM quantify feature importance in terms of effect size associated with the variable. The other models use the difference in the model error rate that results if the feature is removed (Fig. 1).

### Risk-Prediction Calculator

The Cox Lasso model was selected to create an easy-to-use in-clinic calculator to predict the risk of ACL reconstruction revision (Revision Risk Calculator). Whereas the overall risk of revision in the registry was 4.9%, this calculator can quantify the risk at a patient-specific level, with a plausible range from near 0% for low-risk patients to 20% for high-risk patients at 5 years. Table V, Figure 2, and Video 1 demonstrate examples of the calculator's risk prediction using 3 sample patients.

### Discussion

The most important finding of this study was that machine learning analysis of a knee ligament register allows the creation of a validated algorithm to predict a patient's risk of ACL reconstruction revision with moderate accuracy. Additionally, despite having 24 possible prognostic variables contained within the NKLR, the algorithm required only 5 factors for prediction: age and KOOS QoL at the time of the primary surgical procedure, graft choice, femoral fixation device, and the number of years between the injury and the primary surgical procedure. Using this algorithm, an in-clinic calculator was developed that can estimate revision risk.

This study represents the first machine learning-driven model for predicting the outcome of ACL reconstruction at a patient-specific level. Currently, the risk of a patient undergoing a revision ACL reconstruction is estimated on the basis of clinical experience and subjective consideration of the known risk factors. Although it is generally accepted that these factors influence the outcome, the ability to accurately quantify this risk has remained elusive. For the clinician, the introduction of an easy-to-use calculator can guide the patient-specific discussion surrounding the surgical options and realistic outcome goals.

Machine learning is a relatively new tool in the health-care research realm. In this study, 4 models were used to analyze the data and create algorithms predicting the risk of undergoing a revision ACL reconstruction. All models first identified which factors were predictive of a revision surgical procedure and then calculated the relative

weight of their influence on the risk of this outcome. Of all of the various factors contained within the registry, the Cox Lasso model identified only 5 variables necessary to predict outcome, and the other 3 models either used more variables without an appreciable improvement in accuracy or had slightly lower accuracy overall. For this reason, the Cox Lasso model was selected for creation of the in-clinic calculator.

It is interesting to note that several variables that have previously been considered important for predicting ACL reconstruction failure were not necessary for inclusion in the Cox Lasso machine learning model. Some examples include sex<sup>19</sup>, tibial fixation<sup>12</sup>, and increased BMI<sup>20</sup>. Variables were excluded from this model using the Lasso technique, which retains only those predictors adding significantly to the model's accuracy. Although these previously identified risk factors are no doubt associated with outcome, the Lasso method suggests that they are either less important than the factors selected by the Lasso or somehow represented in those factors. In comparison with the Cox Lasso model, the random forest model and the GBM included more variables. However, this inclusion did not significantly improve performance. The reason for this is similar: the information offered by these added variables is already contained within the few most important predictors, so adding the extra variables does not improve performance. All 5 of the variables that were found to be important for outcome prediction have previously been identified as being associated with an increased risk of revision ACL reconstruction<sup>11,12,14,17,20,21,30</sup>.

Revision ACL reconstruction was selected as the primary outcome measure for this study because of the long follow-up and completeness of the data provided for this end point. This is in contrast to a study designed to predict ACL reconstruction failure based on revision surgical procedures and/or inferior patient-reported outcomes. Although this wider outcome would also capture patients who experience a failure but do not undergo a subsequent revision surgical procedure, the number of patients within the register with patient-reported outcome measures substantially drops over time. In contrast, the overall compliance with data entry in the register is 86%<sup>4</sup>. Machine learning analysis requires a large volume of robust data and we therefore chose this narrower outcome measure to maximize patient inclusion and model accuracy.


There were limitations to the current study. First, although we considered a variety of machine learning methods in this analysis, it is possible that a model not considered might have performed better. Second, there were substantial missing data in some predictors such as BMI (32%) and preoperative KOOS (21%), and we could not rule out that data were not missing at random. We noted that observations with complete data for all variables included in the random forest model and the GBM tended to be newer to the registry than incomplete observations, possibly reflecting improvement in data collection over time. Additionally, revision was a relatively rare outcome in these data (<5% of individuals), and most patients

were predicted as being at low risk for revision. For this large majority of low-risk patients, functional scores might have offered more clinical insight.

There were also limitations with regard to the clinical application of this analysis. Especially in the case of the random forest model and the GBM, our models used variables that may not have been readily available in a clinical setting. Clinical utility was greatest with the Cox Lasso model, which required only 5 variables and showed no significant difference in performance from the more complex models. Further, the results of this study may not be applicable to populations in other countries as they represented data from a single national register. Although national registers offer generalizability and real-world applicability<sup>21</sup>, the large number of surgeons included in the data collection may also have produced wide variability in surgical decision-making, skill, and technique. Finally, although the machine learning algorithm was well-calibrated, the concordance was moderate. The accuracy of the model would presumably be improved if a larger data set, such as one composed of combined data from multiple registries or one that included additional variables, was assessed. Potentially important variables may include coronal or sagittal alignment (tibial slope), physical examination findings, rehabilitation information, or surgical technique details such as tunnel position or graft size.

In conclusion, machine learning analysis of a national knee ligament register can predict the risk of ACL reconstruction revision with moderate accuracy. This supports the creation of an in-clinic calculator for point-of-care risk stratification based on the input of only 5 variables. Similar analysis using larger or more comprehensive data may improve the accuracy of risk prediction, and future studies incorporating patients who have experienced a failure of ACL reconstruction but have not undergone subsequent revision may better predict the true risk of ACL reconstruction failure.

## Appendix

 Supporting material provided by the authors is posted with the online version of this article as a data supplement at [jbjs.org \(http://links.lww.com/JBJS/G759\)](http://links.lww.com/JBJS/G759). ■

R. Kyle Martin, MD, FRCSC<sup>1,2</sup>  
Solveig Wastvedt, BA<sup>3</sup>  
Ayoosh Pareek, MD<sup>4</sup>  
Andreas Persson, MD, PhD<sup>5,6,7</sup>  
Håvard Visnes, MD, PhD<sup>6</sup>  
Anne Marie Fenstad, MS<sup>6</sup>  
Gilbert Moatshe, MD, PhD<sup>7,8</sup>  
Julian Wolfson, PhD<sup>3</sup>  
Lars Engebretsen, MD, PhD<sup>7,8</sup>

<sup>1</sup>Department of Orthopedic Surgery, University of Minnesota, Minneapolis, Minnesota

<sup>2</sup>Department of Orthopedic Surgery, CentraCare, Saint Cloud, Minnesota

<sup>3</sup>Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota

<sup>4</sup>Department of Orthopedic Surgery, Mayo Clinic, Rochester, Minnesota

<sup>5</sup>Department of Orthopedic Surgery, Martina Hansens Hospital, Bærum, Norway

<sup>6</sup>Norwegian Knee Ligament Register, Haukeland University Hospital, Bergen, Norway

<sup>7</sup>Oslo Sport Trauma Research Center, Norwegian School of Sports Science, Oslo, Norway

<sup>8</sup>Orthopaedic Clinic, Oslo University Hospital Ullevål, Oslo, Norway

Email for corresponding author: [rkylemmartin@gmail.com](mailto:rkylemmartin@gmail.com)

## References

1. Deacon A, Bennell K, Kiss ZS, Crossley K, Brukner P. Osteoarthritis of the knee in retired, elite Australian Rules footballers. *Med J Aust.* 1997 Feb 17;166(4):187-90.
2. Gombitzy AL, Lott A, Yellin JL, Fabricant PD, Lawrence JT, Ganley TJ. Sport-specific yearly risk and incidence of anterior cruciate ligament tears in high school athletes: a systematic review and meta-analysis. *Am J Sports Med.* 2016 Oct;44(10):2716-23.
3. Granan LP, Forssblad M, Lind M, Engebretsen L. The Scandinavian ACL registries 2004-2007: baseline epidemiology. *Acta Orthop.* 2009 Oct;80(5):563-7.
4. Norwegian Arthroplasty Register, Norwegian Cruciate Ligament Register, Norwegian Hip Fracture Register, Norwegian Paediatric Hip Register. 2020 Annual Report. Norwegian National Advisory Unit on Arthroplasty and Hip Fractures; 2020. Accessed 2020 Jul 9. <http://nrlweb.helse.net/Rapporter/Rapport2020.pdf>
5. Zhang Y, McCammon J, Martin RK, Prior HJ, Leiter J, MacDonald PB. Epidemiological trends of anterior cruciate ligament reconstruction in a Canadian province. *Clin J Sport Med.* 2020 Nov;30(6):e207-13.
6. Davey AP, Vacek PM, Caldwell RA, Slaughterbeck JR, Gardner-Morse MG, Tourville TW, Beynon BD. Risk factors associated with a noncontact anterior cruciate ligament injury to the contralateral knee after unilateral anterior cruciate ligament injury in high school and college female athletes: a prospective study. *Am J Sports Med.* 2019 Dec;47(14):3347-55.
7. Kızılgöz V, Sivrioğlu AK, Ulusoy GR, Aydın H, Karayol SS, Menderes U. Analysis of the risk factors for anterior cruciate ligament injury: an investigation of structural tendencies. *Clin Imaging.* 2018 Jul-Aug;50:20-30.
8. Ma Y, Ao YF, Yu JK, Dai LH, Shao ZX. Failed anterior cruciate ligament reconstruction: analysis of factors leading to instability after primary surgery. *Chin Med J (Engl).* 2013 Jan;126(2):280-5.
9. Montalvo AM, Schneider DK, Webster KE, Yut L, Galloway MT, Heidt RS Jr, Kaeding CC, Kremcheck TE, Magnussen RA, Parikh SN, Stanfield DT, Wall EJ, Myer GD. Anterior cruciate ligament injury risk in sport: a systematic review and meta-analysis of injury incidence by sex and sport classification. *J Athl Train.* 2019 May;54(5):472-82.
10. Montalvo AM, Schneider DK, Yut L, Webster KE, Beynon B, Kocher MS, Myer GD. "What's my risk of sustaining an ACL injury while playing sports?" A systematic review with meta-analysis. *Br J Sports Med.* 2019 Aug;53(16):1003-12.
11. Persson A, Fjeldsgaard K, Gjertsen JE, Kjellsen AB, Engebretsen L, Hole RM, Fevang JM. Increased risk of revision with hamstring tendon grafts compared with patellar tendon grafts after anterior cruciate ligament reconstruction: a study of 12,643 patients from the Norwegian Cruciate Ligament Registry, 2004-2012. *Am J Sports Med.* 2014 Feb;42(2):285-91.
12. Persson A, Kjellsen AB, Fjeldsgaard K, Engebretsen L, Espehaug B, Fevang JM. Registry data highlight increased revision rates for Endobutton/Biosure HA in ACL reconstruction with hamstring tendon autograft: a nationwide cohort study from the Norwegian Knee Ligament Registry, 2004-2013. *Am J Sports Med.* 2015 Sep;43(9):2182-8.

13. Shen X, Xiao J, Yang Y, Liu T, Chen S, Gao Z, Zuo J. Multivariable analysis of anatomic risk factors for anterior cruciate ligament injury in active individuals. *Arch Orthop Trauma Surg.* 2019 Sep;139(9):1277-85.
14. Kaeding CC, Pedroza AD, Reinke EK, Huston LJ, Spindler KP; MOON Consortium. 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 Jul;43(7):1583-90.
15. Fontana MA, Lyman S, Sarker GK, Padgett DE, MacLean CH. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clin Orthop Relat Res.* 2019 Jun;477(6):1267-79.
16. Shohat N, Goswami K, Tan TL, Yayac M, Soriano A, Sousa R, Wouthuyzen-Bakker M, Parvizi J; ESCMID Study Group of Implant Associated Infections (ES-GIAI) and the Northern Infection Network of Joint Arthroplasty (NINJA). 2020 Frank Stinchfield Award: Identifying who will fail following irrigation and debridement for prosthetic joint infection. *Bone Joint J.* 2020 Jul;102-B(7\_Supple\_B)(Supple\_B):11-9.
17. Granan LP, Baste V, Engebretsen L, Inacio MCS. Associations between inadequate knee function detected by KOOS and prospective graft failure in an anterior cruciate ligament-reconstructed knee. *Knee Surg Sports Traumatol Arthrosc.* 2015 Apr;23(4):1135-40.
18. LaPrade CM, Dornan GJ, Granan LP, LaPrade RF, Engebretsen L. Outcomes after anterior cruciate ligament reconstruction using the Norwegian Knee Ligament Registry of 4691 patients: how does meniscal repair or resection affect short-term outcomes? *Am J Sports Med.* 2015 Jul;43(7):1591-7.
19. Hewett TE, Myer GD, Ford KR, Paterno MV, Quatman CE. Mechanisms, prediction, and prevention of ACL injuries: cut risk with three sharpened and validated tools. *J Orthop Res.* 2016 Nov;34(11):1843-55.
20. Snaebjörnsson T, Svantesson E, Sundemo D, Westin O, Sansone M, Engebretsen L, Hamrin-Senorski E. Young age and high BMI are predictors of early revision surgery after primary anterior cruciate ligament reconstruction: a cohort study from the Swedish and Norwegian knee ligament registries based on 30,747 patients. *Knee Surg Sports Traumatol Arthrosc.* 2019 Nov;27(11):3583-91.
21. Webster KE, Feller JA, Leigh WB, Richmond AK. Younger patients are at increased risk for graft rupture and contralateral injury after anterior cruciate ligament reconstruction. *Am J Sports Med.* 2014 Mar;42(3):641-7.
22. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2019. Accessed 2020 May 19. <https://www.R-project.org/>
23. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw.* 2011 Mar;39(5):1-13.
24. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat.* 2008;2(3):841-60.
25. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001;29(5).
26. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal.* 2002; 38(4):367-78.
27. Wood SN. Generalized Additive Models: An Introduction with R. 2nd ed. Chapman and Hall/CRC; 2017.
28. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Johnson PE, Vazquez-Benitez G, O'Connor PJ. 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 Jun;61:119-31.
29. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA.* 1982 May 14;247(18):2543-6.
30. Snaebjörnsson T, Hamrin-Senorski E, Svantesson E, Westin O, Persson A, Karlsson J, Samuelsson K. Graft fixation and timing of surgery are predictors of early anterior cruciate ligament revision: a cohort study from the Swedish and Norwegian Knee Ligament Registries based on 18,425 patients. *JB JS Open Access.* 2019 Dec 12;4(4):e0037.
31. Naylor CD, Guyatt GH; The Evidence-Based Medicine Working Group. Users' guides to the medical literature. X. How to use an article reporting variations in the outcomes of health services. *JAMA.* 1996 Feb 21;275(7):554-8.

**The following content was supplied by the authors as supporting material and has not been copy-edited or verified by JBJS.**

## **Appendix A: Machine Learning Models**

### **Cox Lasso**

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.<sup>1</sup>

### **Survival Random Forest**

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.<sup>2</sup>

### **Generalized additive model**

A generalized additive model (GAM) is a regression model that allows for non-linear relationships between predictors and the outcome. In the R package *mgcv*, which we used for our model, smooth terms are fit using penalized regression splines. The generalized additive model accommodates right-censored time-to-event data by fitting a Cox proportional hazards model with the smooth terms incorporated in the partial likelihood.<sup>3</sup>

### **Gradient boosted regression**

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.<sup>4,5</sup>

## 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. Wood SN. *Generalized Additive Models: An Introduction with R.* 2nd ed. Chapman and Hall/CRC; 2017. doi:10.1201/9781315370279
4. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat.* 2001;29(5). doi:10.1214/aos/1013203451
5. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal.* 2002;38(4):367-378. doi:10.1016/S0167-9473(01)00065-2

## **Supplementary Tables**

Supplementary Table 1a: Complete/incomplete case comparison

<b>Variable*</b>	<b>Full data N = 24935</b>	<b>Cox Lasso/GAM complete cases N = 18887</b>	<b>Random forest/GBM complete cases** N = 13272</b>
Years: surgery to present (1/2020)	8.1 (4.1)	8.4 (4.1)	6.5 (3.1)
Revision	1219 (4.9%)	975 (5.2%)	619 (4.7%)
Follow-up time/Time to revision	6.7 (4.2)	7.1 (4.2)	5.2 (3.1)
Age at surgery	28 (11)	28 (10)	28 (11)
Age at injury	27 (10)	26 (10)	26 (10)
Missing	1251	0	0
Sex			
Male	14019 (56%)	10452 (55%)	7302 (55%)
Female	10916 (44%)	8435 (45%)	5970 (45%)
BMI	25.0 (3.8)	25.0 (3.8)	25.0 (3.8)
Missing	7920	5462	0
QOL score at surgery	3.49 (1.86)	3.49 (1.86)	3.52 (1.88)
Missing	5149	0	0
Sports score at surgery	4.28 (2.73)	4.28 (2.73)	4.34 (2.74)
Missing	5324	192	0
Below median on all KOOS	3972 (20%)	3698 (20%)	2541 (19%)
Missing	4981	0	0
Hospital type			
Southeast	9335 (37%)	6853 (36%)	4621 (35%)
West	3974 (16%)	3080 (16%)	2112 (16%)
Central	2162 (8.7%)	1616 (8.6%)	1013 (7.6%)
North	958 (3.8%)	547 (2.9%)	308 (2.3%)
Private	8506 (34%)	6791 (36%)	5218 (39%)
Meniscus injury	13145 (53%)	9957 (53%)	7219 (54%)
Cartilage injury	5801 (23%)	4464 (24%)	3008 (23%)
Any further injury	171 (0.7%)	92 (0.5%)	59 (0.4%)
PCL injury	398 (1.6%)	213 (1.1%)	127 (1.0%)
MCL injury	1993 (8.0%)	1458 (7.7%)	1125 (8.5%)
LCL injury	464 (1.9%)	302 (1.6%)	241 (1.8%)
PLC injury	243 (1.0%)	134 (0.7%)	93 (0.7%)
Graft choice			
BPTB	9891 (40%)	7393 (39%)	5363 (40%)
Hamstring	14481 (58%)	11142 (59%)	7591 (57%)
Unknown/Other	563 (2.3%)	352 (1.9%)	318 (2.4%)
Damaged side.			
Right	12675 (51%)	9598 (51%)	6733 (51%)
Left	12260 (49%)	9289 (49%)	6539 (49%)
Missing	0 (0%)	0 (0%)	0 (0%)
Previous surgery on opposite knee	1804 (7.2%)	1340 (7.1%)	975 (7.3%)
Previous surgery on same knee	4213 (17%)	3167 (17%)	1852 (14%)

Time injury to surgery (years)	1.63 (3.26)	1.63 (3.27)	1.54 (3.10)
Missing	1255	0	0
Systemic Antibiotic Prophylaxis	24769 (99%)	18784 (99%)	13231 (100%)
Missing	58 (0.2%)	39 (0.2%)	28 (0.2%)

\*Statistics presented: Mean (SD); n (%)

\*\*Fixation device variables (used in random forest and gradient boosted regression models) are omitted from this table for readability (see supplement Table 2c).

Supplementary Table 1b: Cox Lasso/generalized additive model complete/incomplete case comparison

Variable*	Incomplete N = 6048	Complete N = 18887	Total N = 24935	P-value**
Years: surgery to present (1/2020)	7.0 (4.0)	8.4 (4.1)	8.1 (4.1)	<0.001
Revision	244 (4.0%)	975 (5.2%)	1219 (4.9%)	<0.001
Follow-up time/Time to revision	5.7 (4.0)	7.1 (4.2)	6.7 (4.2)	<0.001
Age at surgery	30 (11)	28 (10)	28 (11)	<0.001
QOL score at surgery	3.43 (1.86)	3.49 (1.86)	3.49 (1.86)	0.33
Missing	5149	0	5149	
Graft choice				<0.001
BPTB	2498 (41%)	7393 (39%)	9891 (40%)	
Hamstring	3339 (55%)	11142 (59%)	14481 (58%)	
Unknown/Other	211 (3.5%)	352 (1.9%)	563 (2.3%)	
Femur fixation device				<0.001
Interference screw	1942 (32%)	6345 (34%)	8287 (33%)	
Suspension/cortical device	3065 (51%)	10007 (53%)	13072 (52%)	
Unknown/Other	1041 (17%)	2535 (13%)	3576 (14%)	
Time injury to surgery (years)	1.61 (3.21)	1.63 (3.27)	1.63 (3.26)	0.76
Missing	1255	0	1255	

\*Statistics presented: Mean (SD); n (%)

\*\*Statistical tests performed: t-test, chi-square test

Supplementary Table 1c: Random forest/gradient boosted regression complete/incomplete case comparison

Variable*	Incomplete N = 11663	Complete N = 13272	Total N = 24935	P-value**
Years: surgery to present (1/2020)	9.9 (4.4)	6.5 (3.1)	8.1 (4.1)	<0.001
Revision	600 (5.1%)	619 (4.7%)	1219 (4.9%)	0.084
Follow-up time/Time to revision	8.4 (4.6)	5.2 (3.1)	6.7 (4.2)	<0.001
Age at surgery	29 (11)	28 (11)	28 (11)	<0.001
Age at injury	27 (10)	26 (10)	27 (10)	<0.001
Missing	1251	0	1251	
Sex				<0.001
Male	6717 (58%)	7302 (55%)	14019 (56%)	
Female	4946 (42%)	5970 (45%)	10916 (44%)	
BMI	25.2 (3.8)	25.0 (3.8)	25.0 (3.8)	<0.001
Missing	7920	0	7920	
QOL score at surgery	3.43 (1.82)	3.52 (1.88)	3.49 (1.86)	0.002
Missing	5149	0	5149	

Sports score at surgery	4.16 (2.70)	4.34 (2.74)	4.28 (2.73)	<0.001
Missing	5324	0	5324	
Below median on all KOOS	1431 (21%)	2541 (19%)	3972 (20%)	<0.001
Missing	4981	0	4981	
Hospital type				<0.001
Southeast	4714 (40%)	4621 (35%)	9335 (37%)	
West	1862 (16%)	2112 (16%)	3974 (16%)	
Central	1149 (9.9%)	1013 (7.6%)	2162 (8.7%)	
North	650 (5.6%)	308 (2.3%)	958 (3.8%)	
Private	3288 (28%)	5218 (39%)	8506 (34%)	
Meniscus injury	5926 (51%)	7219 (54%)	13145 (53%)	<0.001
Cartilage injury	2793 (24%)	3008 (23%)	5801 (23%)	0.017
Any further injury	112 (1.0%)	59 (0.4%)	171 (0.7%)	<0.001
PCL injury	271 (2.3%)	127 (1.0%)	398 (1.6%)	<0.001
MCL injury	868 (7.4%)	1125 (8.5%)	1993 (8.0%)	0.003
LCL injury	223 (1.9%)	241 (1.8%)	464 (1.9%)	0.61
PLC injury	150 (1.3%)	93 (0.7%)	243 (1.0%)	<0.001
Graft choice				0.006
BPTB	4528 (39%)	5363 (40%)	9891 (40%)	
Hamstring	6890 (59%)	7591 (57%)	14481 (58%)	
Unknown/Other	245 (2.1%)	318 (2.4%)	563 (2.3%)	
Damaged side.				0.74
Right	5942 (51%)	6733 (51%)	12675 (51%)	
Left	5721 (49%)	6539 (49%)	12260 (49%)	
Missing	0 (0%)	0 (0%)	0 (0%)	
Previous surgery on opposite knee	829 (7.1%)	975 (7.3%)	1804 (7.2%)	0.48
Previous surgery on same knee	2361 (20%)	1852 (14%)	4213 (17%)	<0.001
Time injury to surgery (years)	1.74 (3.44)	1.54 (3.10)	1.63 (3.26)	<0.001
Missing	1255	0	1255	
Systemic Antibiotic Prophylaxis	11538 (99%)	13231 (100%)	24769 (99%)	<0.001
Missing	30 (0.3%)	28 (0.2%)	58 (0.2%)	
Femur fixation device				<0.001
ACL TightRope	28 (0.2%)	16 (0.1%)	44 (0.2%)	
Aesculap Position ACL	27 (0.2%)	27 (0.2%)	54 (0.2%)	
BioComposite SwiveLock C	1 (<0.1%)	0 (0%)	1 (<0.1%)	
Biodegr screw	50 (0.4%)	53 (0.4%)	103 (0.4%)	
BioRCI	4 (<0.1%)	3 (<0.1%)	7 (<0.1%)	
BioRCI-HA	2 (<0.1%)	0 (0%)	2 (<0.1%)	
Biosure HA	4 (<0.1%)	31 (0.2%)	35 (0.1%)	
Biosure HA Interference screw	0 (0%)	1 (<0.1%)	1 (<0.1%)	
Biosure PK	0 (0%)	2 (<0.1%)	2 (<0.1%)	
BioTenodesis Screw System	1 (<0.1%)	0 (0%)	1 (<0.1%)	
Bone Mulch	483 (4.2%)	135 (1.0%)	618 (2.5%)	
Bone Mulch Screw	1 (<0.1%)	0 (0%)	1 (<0.1%)	
BTB TightRope	87 (0.8%)	45 (0.3%)	132 (0.5%)	
Comp non-degr	139 (1.2%)	185 (1.4%)	324 (1.3%)	
Cortical button	78 (0.7%)	76 (0.6%)	154 (0.6%)	



Endobutton	3260 (28%)	5349 (41%)	8609 (35%)
EndoButton CL	2 (<0.1%)	0 (0%)	2 (<0.1%)
Endobutton CL BTB	465 (4.1%)	811 (6.2%)	1276 (5.2%)
Endobutton CL Ultra	16 (0.1%)	43 (0.3%)	59 (0.2%)
EzLoc	1152 (10%)	594 (4.5%)	1746 (7.1%)
EZLoc	3 (<0.1%)	0 (0%)	3 (<0.1%)
Full Thread Interference screw	2 (<0.1%)	2 (<0.1%)	4 (<0.1%)
Guardsman Femoral	1 (<0.1%)	1 (<0.1%)	2 (<0.1%)
Linvatec Cannulated	1 (<0.1%)	0 (0%)	1 (<0.1%)
Metal int screw	635 (5.5%)	866 (6.6%)	1501 (6.1%)
Other suspension devices/cortical	9 (<0.1%)	13 (<0.1%)	22 (<0.1%)
Other Suspension devices/cortical	226 (2.0%)	305 (2.3%)	531 (2.2%)
Other transfemoral devices	2 (<0.1%)	0 (0%)	2 (<0.1%)
Peek Interference Screw	14 (0.1%)	5 (<0.1%)	19 (<0.1%)
Profile interference screw	86 (0.8%)	333 (2.5%)	419 (1.7%)
Profile Interference Screw	0 (0%)	1 (<0.1%)	1 (<0.1%)
Propel Cannulated	0 (0%)	2 (<0.1%)	2 (<0.1%)
Propel cannulated int. screw	188 (1.6%)	33 (0.3%)	221 (0.9%)
RCI screw	431 (3.8%)	316 (2.4%)	747 (3.0%)
RCI Screw	11 (<0.1%)	7 (<0.1%)	18 (<0.1%)
Rigidfix	508 (4.4%)	100 (0.8%)	608 (2.5%)
Rigidfix BTB cross-pin	205 (1.8%)	182 (1.4%)	387 (1.6%)
Rigidfix BTB cross pin	0 (0%)	2 (<0.1%)	2 (<0.1%)
Rigidfix ST cross pin Kit	3 (<0.1%)	0 (0%)	3 (<0.1%)
Sheated Cannulated Interference Screw	6 (<0.1%)	14 (0.1%)	20 (<0.1%)
Soft screw	12 (0.1%)	3 (<0.1%)	15 (<0.1%)
Soft Screw	10 (<0.1%)	16 (0.1%)	26 (0.1%)
SoftSilk	1615 (14%)	1828 (14%)	3443 (14%)
TendonSoft	0 (0%)	1 (<0.1%)	1 (<0.1%)
Tightrope ABS	18 (0.2%)	18 (0.1%)	36 (0.1%)
ToggleLoc	144 (1.3%)	591 (4.5%)	735 (3.0%)
Transfix II	852 (7.4%)	256 (1.9%)	1108 (4.5%)
TunneLoc	462 (4.0%)	469 (3.6%)	931 (3.8%)
UltraButton	0 (0%)	1 (<0.1%)	1 (<0.1%)
Universal Wedge	212 (1.9%)	433 (3.3%)	645 (2.6%)
Missing	207	103	310
Tibia fixation device			<0.001
ACL TightRope	5 (<0.1%)	4 (<0.1%)	9 (<0.1%)
Aesculap Position ACL	15 (0.1%)	25 (0.2%)	40 (0.2%)
AO Screw	2 (<0.1%)	0 (0%)	2 (<0.1%)
Bio-Intrafix Screw	1 (<0.1%)	1 (<0.1%)	2 (<0.1%)
Bio Composite Interference Screw	1 (<0.1%)	5 (<0.1%)	6 (<0.1%)
Bio Intrafix	371 (3.2%)	351 (2.7%)	722 (2.9%)
BioComposite SwiveLock C	22 (0.2%)	2 (<0.1%)	24 (<0.1%)
Biodegr screw	675 (5.9%)	712 (5.4%)	1387 (5.6%)
BioRCI	183 (1.6%)	486 (3.7%)	669 (2.7%)

BioRCI-HA	5 (<0.1%)	9 (<0.1%)	14 (<0.1%)
BIORCI Screw	1 (<0.1%)	3 (<0.1%)	4 (<0.1%)
Biosure HA	294 (2.6%)	1768 (13%)	2062 (8.4%)
Biosure HA Interference screw	23 (0.2%)	32 (0.2%)	55 (0.2%)
Biosure PK	47 (0.4%)	119 (0.9%)	166 (0.7%)
BioTenodesis Screw System	0 (0%)	1 (<0.1%)	1 (<0.1%)
BTB TightRope	2 (<0.1%)	1 (<0.1%)	3 (<0.1%)
Comp non-degr	445 (3.9%)	813 (6.2%)	1258 (5.1%)
ComposiTCP 60	0 (0%)	4 (<0.1%)	4 (<0.1%)
Cortical button	0 (0%)	2 (<0.1%)	2 (<0.1%)
Cramp	1 (<0.1%)	0 (0%)	1 (<0.1%)
Delta Tapered Bio-Interference screw	1 (<0.1%)	0 (0%)	1 (<0.1%)
Endobutton	14 (0.1%)	44 (0.3%)	58 (0.2%)
Endobutton CL BTB	6 (<0.1%)	4 (<0.1%)	10 (<0.1%)
Full Thread Interference screw	2 (<0.1%)	1 (<0.1%)	3 (<0.1%)
Intrafix	954 (8.3%)	696 (5.3%)	1650 (6.7%)
Intrafix Screw	1 (<0.1%)	1 (<0.1%)	2 (<0.1%)
Lintratec Cannulated	2 (<0.1%)	1 (<0.1%)	3 (<0.1%)
Low Profile Cancelless	4 (<0.1%)	12 (<0.1%)	16 (<0.1%)
Metal int screw	733 (6.4%)	875 (6.7%)	1608 (6.5%)
Milagro	0 (0%)	1 (<0.1%)	1 (<0.1%)
Other suspension devices/cortical	16 (0.1%)	14 (0.1%)	30 (0.1%)
Other Suspension devices/cortical	114 (1.0%)	168 (1.3%)	282 (1.1%)
Other transtibial devices	2 (<0.1%)	0 (0%)	2 (<0.1%)
Peek Interference Screw	14 (0.1%)	11 (<0.1%)	25 (0.1%)
Profile interference screw	83 (0.7%)	333 (2.5%)	416 (1.7%)
Profile Interference Screw	0 (0%)	1 (<0.1%)	1 (<0.1%)
Propel Cannulated	1 (<0.1%)	2 (<0.1%)	3 (<0.1%)
Propel cannulated int. screw	516 (4.5%)	461 (3.5%)	977 (4.0%)
RCI screw	2355 (21%)	2050 (16%)	4405 (18%)
RCI Screw	48 (0.4%)	44 (0.3%)	92 (0.4%)
Rigidfix	1 (<0.1%)	0 (0%)	1 (<0.1%)
Rigidfix BTB cross-pin	7 (<0.1%)	6 (<0.1%)	13 (<0.1%)
Sheated Cannulated Interference Screw	1 (<0.1%)	1 (<0.1%)	2 (<0.1%)
Soft screw	523 (4.6%)	395 (3.0%)	918 (3.7%)
Soft Screw	13 (0.1%)	19 (0.1%)	32 (0.1%)
SoftSilk	1948 (17%)	2232 (17%)	4180 (17%)
SoftSilk 2	0 (0%)	1 (<0.1%)	1 (<0.1%)
Staple	56 (0.5%)	53 (0.4%)	109 (0.4%)
Suture washer star. Box of 1	1 (<0.1%)	4 (<0.1%)	5 (<0.1%)
TendonSoft	0 (0%)	1 (<0.1%)	1 (<0.1%)
Tightrope ABS	7 (<0.1%)	7 (<0.1%)	14 (<0.1%)
TunneLoc	456 (4.0%)	477 (3.6%)	933 (3.8%)
Universal Wedge	62 (0.5%)	415 (3.2%)	477 (1.9%)
WasherLoc	1395 (12%)	473 (3.6%)	1868 (7.6%)
WasherLoc Screw	5 (<0.1%)	0 (0%)	5 (<0.1%)

Missing	229	131	360	
Fixation device combination				<0.001
Bone Mulch/Intrafix	103 (0.9%)	118 (0.9%)	221 (0.9%)	
Bone Mulch/WasherLoc	376 (3.2%)	16 (0.1%)	392 (1.6%)	
Endobutton/Biodegr. int. screw	87 (0.7%)	292 (2.2%)	379 (1.5%)	
Endobutton/BioIntrafix	92 (0.8%)	204 (1.5%)	296 (1.2%)	
Endobutton/BioRCI	159 (1.4%)	453 (3.4%)	612 (2.5%)	
Endobutton/Biosure HA	283 (2.4%)	1722 (13%)	2005 (8.0%)	
Endobutton/Comp non-degr.	171 (1.5%)	324 (2.4%)	495 (2.0%)	
Endobutton/Intrafix	488 (4.2%)	400 (3.0%)	888 (3.6%)	
Endobutton/Met. int. screw	91 (0.8%)	172 (1.3%)	263 (1.1%)	
Endobutton/RCI	1791 (15%)	1606 (12%)	3397 (14%)	
EzLoc/WasherLoc	1004 (8.6%)	437 (3.3%)	1441 (5.8%)	
Metal int screw x 2	336 (2.9%)	523 (3.9%)	859 (3.4%)	
Other combination	3024 (26%)	3646 (27%)	6670 (27%)	
RCI/RCI	284 (2.4%)	279 (2.1%)	563 (2.3%)	
RCI/Softsilk	138 (1.2%)	23 (0.2%)	161 (0.6%)	
Rigidfix BTB/Met. int. screw	77 (0.7%)	52 (0.4%)	129 (0.5%)	
Rigidfix BTB/Prop. cannulated screw	119 (1.0%)	127 (1.0%)	246 (1.0%)	
Rigidfix/Bio-Intrafix	173 (1.5%)	22 (0.2%)	195 (0.8%)	
Rigidfix/Intrafix	285 (2.4%)	76 (0.6%)	361 (1.4%)	
Softsilk x 2	1415 (12%)	1586 (12%)	3001 (12%)	
Softsilk/RCI	98 (0.8%)	90 (0.7%)	188 (0.8%)	
ToggleLoc/Bio-screw	55 (0.5%)	209 (1.6%)	264 (1.1%)	
Transfix/Biodegr int. screw	249 (2.1%)	24 (0.2%)	273 (1.1%)	
Transfix/Metal int. screw incl RCI	101 (0.9%)	4 (<0.1%)	105 (0.4%)	
TunneLoc/TunneLoc	445 (3.8%)	447 (3.4%)	892 (3.6%)	
Universal Wedge x 2	62 (0.5%)	414 (3.1%)	476 (1.9%)	
Universal Wedge/Bio-screw	137 (1.2%)	6 (<0.1%)	143 (0.6%)	
Missing	20	0	20	

\*Statistics presented: Mean (SD); n (%)

\*\*Statistical tests performed: t-test, chi-square test

Supplementary Table 2a: Cox Lasso performance with imputation

Year	Training data imputed (predictions averaged)			Training and test data imputed (predictions averaged)		
	Concordance	Calibration statistic	P-value	Concordance	Calibration statistic	P-value
1	0.681	4.89	0.18	0.685	4.74	0.192
2	0.679	10.21	0.017	0.681	17.87	< 0.001
5	0.678	3.24	0.357	0.678	1.57	0.667

Supplementary Table 2b: Random forest performance with imputation

	Training data imputed			Training and test data imputed		
Year	Concordance	Calibration statistic	P-value	Concordance	Calibration statistic	P-value
1	0.683	1.9	0.593	0.69	1.76	0.624
2	0.68	8.94	0.03	0.689	10.08	0.018
5	0.677	2.96	0.399	0.69	3.64	0.303

Supplementary Table 2c: Generalized additive model performance with imputation

	Training data imputed (predictions averaged)			Training and test data imputed (predictions averaged)		
Year	Concordance	Calibration statistic	P-value	Concordance	Calibration statistic	P-value
1	0.686	4.93	0.177	0.689	9.32	0.025
2	0.684	10.52	0.015	0.685	17.17	< 0.001
5	0.682	5.3	0.151	0.682	4.78	0.189

Supplementary Table 2d: Gradient boosted regression performance with imputation

	Training data imputed (predictions averaged)			Training and test data imputed (predictions averaged)		
Year	Concordance	Calibration statistic	P-value	Concordance	Calibration statistic	P-value
1	0.675	0.42	0.936	0.685	1.37	0.713
2	0.672	1.99	0.575	0.682	4.53	0.21
5	0.668	4.22	0.239	0.681	11.67	0.009

Supplementary Table 3a: Random forest restricted to Lasso-selected variables

	Complete cases			Training and test data imputed		
Year	Concordance	Calibration statistic	P-value	Concordance	Calibration statistic	P-value
1	0.671	5.95	0.114	0.669	7.22	0.065
2	0.673	38.28	< 0.001	0.669	12.29	0.006
5	0.677	137.74	< 0.001	0.669	5.15	0.161

Supplementary Table 3b: Gradient boosted regression restricted to Lasso-selected variables

	Complete cases	Training and test data imputed
--	----------------	--------------------------------

Year	Concordance	Calibration statistic	P-value	Concordance	Calibration statistic	P-value
1	0.683	2535.36	< 0.001	0.684	6.07	0.108
2	0.683	5731.62	< 0.001	0.682	10.27	0.016
5	0.685	10008.69	< 0.001	0.68	8.62	0.035



## Paper II

Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Engebretsen L. Predicting subjective failure of ACL reconstruction: a machine learning analysis of the Norwegian Knee Ligament Register and patient reported outcomes. *J ISAKOS*. 2022;7(3):1-9. doi:10.1016/j.jisako.2021.12.005







Contents lists available at ScienceDirect

Journal of ISAKOS

journal homepage: [www.elsevier.com/locate/jisakos](http://www.elsevier.com/locate/jisakos)

## Original Research

## Predicting subjective failure of ACL reconstruction: a machine learning analysis of the Norwegian Knee Ligament Register and patient reported outcomes



R. Kyle Martin, MD, FRCSC<sup>a,b,\*</sup>, Solvejg Wastvedt, BA<sup>c</sup>, Ayoosh Pareek, MD<sup>d</sup>, Andreas Persson, MD, PhD<sup>e,f,g</sup>, Håvard Visnes, MD, PhD<sup>f</sup>, Anne Marie Fenstad, MS<sup>f</sup>, Gilbert Moatshe, MD, PhD<sup>g,h</sup>, Julian Wolfson, PhD<sup>c</sup>, Lars Engebretsen, MD, PhD<sup>g,h</sup>

<sup>a</sup> Department of Orthopaedic Surgery, University of Minnesota, Minneapolis, MN, USA

<sup>b</sup> Department of Orthopaedic Surgery, CentraCare, Saint Cloud, MN, USA

<sup>c</sup> Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA

<sup>d</sup> Department of Orthopaedic Surgery, Mayo Clinic, Rochester, MN, USA

<sup>e</sup> Department of Orthopaedic Surgery, Martina Hansens Hospital, Berum, Norway

<sup>f</sup> Norwegian Knee Ligament Register, Haukeland University Hospital, Norway

<sup>g</sup> Oslo Sport Trauma Research Center, Norwegian School of Sports Science, Oslo, Norway

<sup>h</sup> Orthopaedic Clinic, Oslo University Hospital Ullevål, Oslo, Norway

## ARTICLE INFO

## Keywords:

Machine learning  
Artificial intelligence  
ACL reconstruction  
Subjective outcome

## ABSTRACT

**Objectives:** Accurate prediction of outcome following anterior cruciate ligament (ACL) reconstruction is challenging, and machine learning has the potential to improve our predictive capability. The purpose of this study was to determine if machine learning analysis of the Norwegian Knee Ligament Register (NKLK) can (1) identify the most important risk factors associated with subjective failure of ACL reconstruction and (2) develop a clinically meaningful calculator for predicting the probability of subjective failure following ACL reconstruction.

**Methods:** Machine learning analysis was performed on the NKLK. All patients with 2-year follow-up data were included. The primary outcome was the probability of subjective failure 2 years following primary surgery, defined as a Knee Injury and Osteoarthritis Outcome Score (KOOS) Quality of Life (QoL) subscale score of <44. Data were split randomly into training (75%) and test (25%) sets. Four models intended for this type of data were tested: Lasso logistic regression, random forest, generalized additive model (GAM), and gradient boosted regression (GBM). These four models represent a range of approaches to statistical details like variable selection and model complexity. Model performance was assessed by calculating calibration and area under the curve (AUC).

**Results:** Of the 20,818 patients who met the inclusion criteria, 11,630 (56%) completed the 2-year follow-up KOOS QoL questionnaire. Of those with complete KOOS data, 22% reported subjective failure. The lasso logistic regression, GBM, and GAM all demonstrated AUC in the moderate range (0.67–0.68), with the GAM performing best (0.68; 95% CI 0.64–0.71). Lasso logistic regression, GBM, and the GAM were well-calibrated, while the random forest showed evidence of mis-calibration. The GAM was selected to create an in-clinic calculator to predict subjective failure risk at a patient-specific level ([https://swastvedt.shinyapps.io/calculator\\_koosqol/](https://swastvedt.shinyapps.io/calculator_koosqol/)).

**Conclusion:** Machine learning analysis of the NKLK can predict subjective failure risk following ACL reconstruction with fair accuracy. This algorithm supports the creation of an easy-to-use in-clinic calculator for point-of-care risk stratification. Clinicians can use this calculator to estimate subjective failure risk at a patient-specific level when discussing outcome expectations preoperatively.

**Level of evidence:** Level-III Retrospective review of a prospective national register.

\* Corresponding author. Department of Orthopaedic Surgery, University of Minnesota, 2512 South 7th Street, Suite R200 Minneapolis, MN 55455, USA. Tel.: +1 612 273 1177.

E-mail address: [rkylemmartin@gmail.com](mailto:rkylemmartin@gmail.com) (R.K. Martin).

<https://doi.org/10.1016/j.jisako.2021.12.005>

Received 23 September 2021; Accepted 30 December 2021

Available online 11 January 2022

2059-7754/© 2022 The Author(s). Published by Elsevier Inc. on behalf of International Society of Arthroscopy, Knee Surgery and Orthopaedic Sports Medicine. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

### What are the new findings?

- Machine learning analysis can be performed on a national knee ligament register to predict the risk of subjective failure following anterior cruciate ligament reconstruction
- An in-clinic calculator can guide clinical discussion and expectations at a patient-specific level
- Variables for predicting subjective failure following anterior cruciate ligament reconstruction are patient-related and non-modifiable by the surgeon

## Introduction

Anterior cruciate ligament (ACL) reconstruction is a common orthopaedic procedure aimed at restoring function and stability following injury. Literature regarding the surgical outcome is often reported in relation to patient-reported outcome measures (PROM), and several risk factors for a poor outcome have been suggested [1–4]. Currently, however, the ability to use these predictors at the time of surgery to accurately predict which patients are at risk of experiencing a poor outcome is poor [1].

Recently, there has been an increased focus on the use of artificial intelligence and machine learning to improve predictive capability within several fields of medicine, including orthopaedic surgery [5–9]. These advanced statistical techniques utilise computer algorithms to model complex interactions between variables and may lead to improved capacity to predict the outcome. The “advanced” nature of these techniques is derived from the fact that the interactions can be more complex than with traditional statistics. Machine learning analyses can consider all possible interactions between variables in a database and determine the relationships to the desired outcome measure. The factors important for predicting outcomes can then be identified and used to develop the predictive algorithm. Often, minimal explicit and direct human computer programming is required, and the resulting algorithms can be used to prospectively predict the patient-specific outcome.

The Norwegian Knee Ligament Register (NKLK) has been prospectively collecting demographic, injury, surgical, and outcome data since 2004. It now includes over 25,000 patients who have undergone ACL reconstruction with high compliance across the country [10]. Several studies that have improved our understanding of ACL injuries have been based on the NKLK [11–14], and machine learning analysis allows deeper evaluation of factors associated with outcome [9]. There are currently no machine learning models to predict subjective outcomes following primary ACL reconstruction, and the development of such a tool could impact clinical practice by informing shared decision-making and outcome expectations.

The purpose of this study was to use machine learning analysis of the NKLK to (1) identify the most important risk factors associated with subjective failure of primary ACL reconstruction and (2) develop a clinically meaningful model for predicting subjective failure of primary ACL reconstruction. Subjective failure was defined as a Knee Injury and Osteoarthritis Outcome Score (KOOS) Quality of Life (QoL) subscale score of <44. This endpoint has been clinically validated as a marker of failure following ACL reconstruction [11]. The hypothesis was that machine learning analysis would facilitate accurate prediction of subjective failure for a patient undergoing primary ACL reconstruction.

## Materials and methods

This manuscript was written in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis statement [15].

### Data source

The NKLK is a nationwide register aiming to collect all reconstructive surgery on cruciate ligament injuries in Norway. Reporting has been mandatory since 2017, and the compliance of reporting to the register was 86% in 2017 to 2018 [10]. The patients are registered with their personal social security number, which allows them to be followed in case of later surgery independent of service provider. Patient-specific and intraoperative data are submitted to the NKLK by the surgeons (through an article or web-based form directly after surgery). The patients are to report KOOS preoperatively and at 2, 5, and 10 years of follow-up.

### Ethics

Informed consent is obtained from all patients at time of enrolment in the NKLK. Based on this consent, the Norwegian Data Inspectorate provides permission for the NKLK to collect, analyse, and publish on health data. The registration of data was performed confidentially and according to Norwegian and European Union data protection rules, with all data de-identified prior to retrieval from the NKLK. The Regional Ethics Committee has previously determined that it is not necessary to obtain further ethical approval for Norwegian register-based studies [16].

### Data preparation

This level-III retrospective review of a prospective national register included all patients contained within the NKLK with primary ACL reconstruction surgery dates from January 2004 through December 2018. Those with values for graft choice recorded as “direct suture,” “other,” or missing were excluded. Patients with other ligamentous injuries at the time of primary surgery or <2 years of follow-up were also excluded. Variables considered in the analysis are presented in Table 1. Variables were re-coded or newly defined for the following: years between injury and primary surgery; cartilage injury identified at surgery (none, ICRS 1–2, ICRS 3–4); meniscus injury identified at surgery (yes/no); graft choice (patellar tendon autograft, hamstring tendon autograft, other); fixation choice (interference screw, suspension/cortical device, other); and height and weight variables that combined data from the patient- and surgeon-reported variables. A predictor indicating if a patient was below the median score in all five KOOS categories at the time of primary surgery was also created, and predictors for KOOS QoL and Sports measures were scaled to a score out of 10.

### Model creation

The primary outcome was the probability of subjective failure at 2 years following primary ACL reconstruction, as defined as a KOOS QoL score of <44. Cleaned data were randomly split into training (75%) and test (25%) sets that were used to fit and evaluate the models, respectively. The program R (version: 3.6.1, R Core Team 2019) was used to fit four machine learning models to the training data: lasso logistic regression, random forest, gradient boosted regression model (GBM), and generalized additive model (GAM) [17]. These four models are among the most commonly used for machine learning classification tasks and offer a range of approaches in terms of variable selection, optimisation technique, and complexity. Lasso logistic regression is a parametric, penalised regression model that selects a subset of variables for inclusion [18]. The random forest is a tree-based, nonparametric method [19]. GBMs are also nonparametric, meaning that they do not require pre-specification of a model structure and iteratively improve the model fit using all available variables [20,21]. GAM allow for machine-selected nonlinear relationships among a pre-specified group of variables [22]. Further description of each of the four machine learning models can be found in Appendix A.

An L1-regularised logistic regression model (“lasso logistic regression,” package *glmnet*; lambda value selected via cross-validation) was

**Table 1**  
Characteristics of patients.

Variable <sup>a</sup>	All N = 20,818	Complete 2-year Outcome Data N = 11,630
Follow-up time or time to revision	7.3 (3.9)	7.9 (3.6)
KOOS QoL <44 at 2 years	2,556 (22%)	2,556 (22%)
Missing	9,188	0
Age at surgery	28 (10)	29 (11)
Age at injury	26 (10)	27 (11)
Missing	1072	544
Sex		
Male	11,669 (56%)	5,836 (50%)
Female	9,149 (44%)	5,794 (50%)
Pre-surgery BMI	25.0 (3.7)	24.8 (3.7)
Missing	7,244	4,365
Pre-surgery KOOS QoL score (out of 10)	3.50 (1.83)	3.52 (1.83)
Missing	4,022	2,008
Pre-surgery KOOS Sports score (out of 10)	4.33 (2.71)	4.37 (2.69)
Missing	4,162	2,087
Below median on all pre-surgery KOOS	3,285 (19%)	1,806 (19%)
Missing	3,893	1,942
Activity that led to injury		
Non-pivoting	4,109 (25%)	2,392 (26%)
Pivoting	12,007 (75%)	6,716 (74%)
Other/Unknown	0 (0%)	0 (0%)
Missing	4,702	2,522
Meniscus injury	10,942 (53%)	5,927 (51%)
Cartilage injury		
ICRS 1-2	3,625 (17%)	2,016 (17%)
ICRS 3-4	993 (4.8%)	577 (5.0%)
None	16,200 (78%)	9,037 (78%)
Graft choice		
BPTB autograft	7,334 (35%)	3,782 (33%)
Hamstring autograft	13,197 (63%)	7,740 (67%)
Other	287 (1.4%)	108 (0.9%)
Tibia fixation device		
Interference screw	17,893 (89%)	9,905 (88%)
Suspension/cortical device	2,073 (10%)	1,245 (11%)
Other	152 (0.8%)	88 (0.8%)
Missing	700	392
Femur fixation device		
Interference screw	6,325 (31%)	3,314 (29%)
Suspension/cortical device	11,629 (57%)	6,613 (58%)
Other	2,484 (12%)	1,491 (13%)
Missing	380	212
Fixation device combination		
Interference screw x2	6,028 (30%)	3,163 (28%)
Interference/suspension	51 (0.3%)	17 (0.2%)
Suspension/cortical device x2	1,646 (8.2%)	1,011 (9.0%)
Suspension/interference	9,635 (48%)	5,410 (48%)
Other	2,634 (13%)	1,577 (14%)
Missing	824	452
Injured side		
Right	10,613 (51%)	5,871 (50%)
Left	10,205 (49%)	5,759 (50%)
Previous surgery on opposite knee	1,526 (7.3%)	786 (6.8%)
Previous surgery on same knee	3,784 (18%)	2,220 (19%)
Time injury to surgery (years)	1.71 (3.36)	1.81 (3.63)
Missing	1,076	546
Systemic Antibiotic Prophylaxis	20,669 (100%)	11,534 (99%)
Missing	51	34

<sup>a</sup> Statistics presented: Mean (SD); n (%).

applied to select variables for each outcome, and those with non-zero coefficients were retained (Fig. 1). Random forests (function *randomForest* from package *randomForest*) were trained for each outcome with minimum node size 5, 10 variables tried per split, 500 trees, and the full set of predictors (hyperparameters selected via cross-validation). GAMs (function *gam* from package *mgcv*) were trained with those variables selected in the lasso for the respective outcomes, using smooth terms for all continuous variables selected. Finally, GBMs (function *gbm* from package *gbm*) were trained using a shrinkage parameter of 0.01,

minimum node size of 10, maximum tree depth of 3, 1000 trees, and the full set of predictors (hyperparameters selected via cross-validation). All four models were restricted to patients with complete data for the predictors used (Table 2a and Table 2b).

#### Model evaluation

Model performance was evaluated by calculating predicted probabilities of subjective failure at 2 years of follow-up for the hold-out test data using the trained models. Model calibration was assessed using the Hosmer–Lemeshow statistic (function *hoslem.test* in package *ResourceSelection*) [23]. Calibration refers to the accuracy of the predicted probabilities, comparing expected to actual observed outcomes. This statistic sums average misclassification in each predicted risk quintile and converts the sum into a chi-squared statistic. Larger calibration statistics correspond to smaller p values, and statistical significance means that the null hypothesis of perfect calibration is rejected. The area under the curve (AUC) was also calculated for each model along with confidence intervals for the AUC using bootstrap resampling (functions *auc* and *ci.auc* from package *pROC*).

#### Missing data

An inverse probability-weighted analysis was conducted to assess whether patients with complete follow-up KOOS QoL score data were fundamentally different from those with missing outcome data based on observed characteristics. Inverse probability weighting assigns each observation a weight based on the inverse of the probability of a patient with similar observed characteristics being present in the dataset. In this case, patients with combinations of predictor variables that are rare in the complete outcome dataset receive high weights. Conversely, patients with common predictor variables are down-weighted to adjust for their over-representation. The result of the weighting is a population that mimics what would have occurred if all patients were to have complete outcome data. The same models are then built on this weighted population and compared to the unweighted analysis. If the weighted models show substantively different results, this indicates that there may be fundamental differences between patients with complete and incomplete outcome data. If there is no substantive difference, this indicates that removing patients with incomplete outcome data does not jeopardise the results.

To assess the effect of excluding patients with missing predictor values from the models, the same four models were trained using multiple imputations to fill in missing values in the training data (function *mice* from package *mice*). As with the weighted models, if there is no substantive difference when using imputation, this indicates that removing patients with incomplete predictor data does not adversely affect the results.

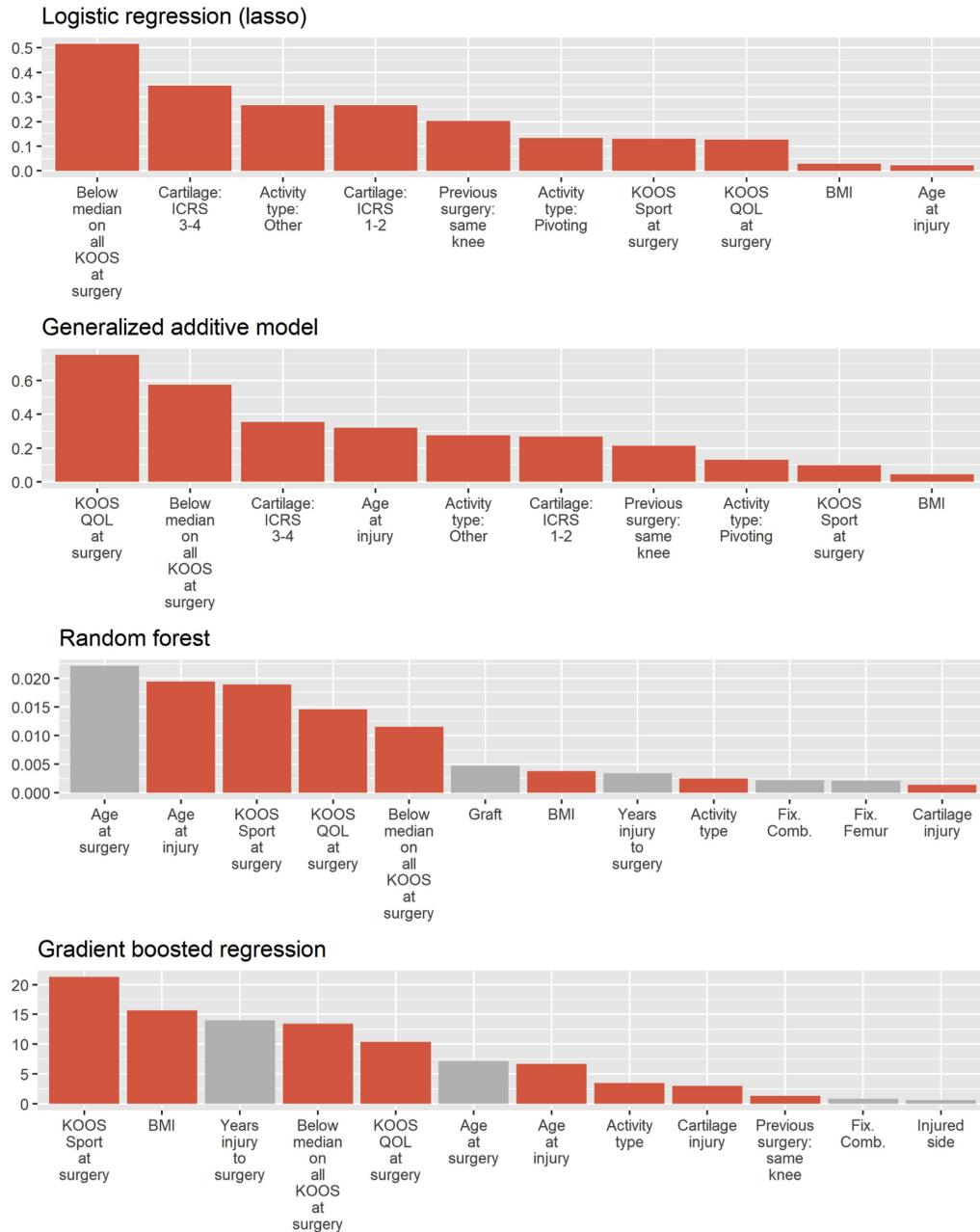
#### Sources of funding

This study was funded by a Norwegian Centennial Chair seed grant. Funding supported the machine learning analysis and interpretation. The funding agencies had no direct role in the investigation.

#### Results

##### Data characteristics

Table 1 describes the characteristics of the registered population at the time of primary surgery and the variables included for analysis. After data cleaning, 20,818 patients met the inclusion criteria (Fig. 2). Of these patients, 11,630 (56%) had complete 2-year follow-up KOOS QoL data. Subjective failure (KOOS QoL score <44) occurred in 2,556 (22%) of the patients with complete outcome data. The population was approximately evenly split between male and female, with an average age (and standard deviation) of 29 ± 11 years at the time of primary surgery.



**Fig. 1.** Variable Importance. The four plots show relative feature importance in each of the machine learning models. The vertical axis is a variable importance score, which differs depending on the model. For the lasso logistic regression and GAM, the vertical axis is the absolute value of the variable coefficient (effect size). For the random forest and GBM, the scale is the decrease in model error rate if the variable were to be removed from the model. The highlighted bars indicate variables that were selected using the lasso and included in the final model used for the in-clinic calculator. GAM, generalized additive model; GBM, gradient boosted regression model.

**Table 2a**

Lasso logistic regression/generalised additive model complete/incomplete case comparison.

Variable*	Incomplete N = 14,810	Complete N = 6,008	Total N = 20,818	P-value**
Years: surgery to data current date (2020-01-12)	9.1 (4.2)	7.6 (2.5)	8.6 (3.9)	<0.001
KOOS QoL <44 at 2 years	1,270 (23%)	1,286 (21%)	2,556 (22%)	0.13
Missing	9,188	0	9,188	
Age at injury	26 (10)	27 (11)	26 (10)	0.006
Missing	1,072	0	1,072	
Pre-surgery BMI	25.1 (3.7)	24.8 (3.7)	25.0 (3.7)	<0.001
Missing	7,244	0	7,244	
Pre-surgery KOOS QoL score (out of 10)	3.48 (1.83)	3.55 (1.85)	3.50 (1.83)	0.016
Missing	4,022	0	4,022	
Pre-surgery KOOS Sports score (out of 10)	4.29 (2.71)	4.42 (2.70)	4.33 (2.71)	0.002
Missing	4,162	0	4,162	
Below median on all pre-surgery KOOS scores	2,199 (20%)	1,086 (18%)	3,285 (19%)	0.001
Missing	3,893	0	3,893	
Activity that led to injury				<0.001
Non-pivoting	2,784 (19%)	1,325 (22%)	4,109 (20%)	
Pivoting	8,433 (59%)	3,574 (59%)	12,007 (59%)	
Other	3,122 (22%)	1,109 (18%)	4,231 (21%)	
Missing	471	0	471	
Cartilage injury				0.015
ICRS 1-2	2,648 (18%)	977 (16%)	3,625 (17%)	
ICRS 3-4	692 (4.7%)	301 (5.0%)	993 (4.8%)	
None	11,470 (77%)	4,730 (79%)	16,200 (78%)	
Previous surgery on same knee	2,824 (19%)	960 (16%)	3,784 (18%)	<0.001

\*Statistics presented: Mean (SD); n (%).

\*\*Statistical tests performed: t-test, chi-square test.

To assess the impact of restricting the analysis to patients with complete KOOS QoL score data, covariate distributions between patients with complete outcomes and the full dataset were compared (Table 1). Covariate distributions between the complete cases for each model and the full dataset were also compared (Table 2a and Table 2b). Due to the large sample sizes, some comparisons produce p values below the significance threshold: those with complete data were newer to the register, had their surgeries at higher-volume hospitals, and were more likely to be female. However, these differences were in general small and of limited clinical significance. An inverse-probability-weighted analysis and an analysis imputing missing covariate data was also performed. Neither alternative analysis showed meaningfully different results from the complete case models (Table 3 and Table 4).

#### Model performance

The lasso logistic regression, GBM, and GAM all demonstrated AUC in the moderate range (0.67–0.68), with the GAM performing best at 0.68 (95% CI 0.64–0.71). Lasso logistic regression, gradient boosted regression, and the GAM were well-calibrated, and the random forest showed evidence of miscalibration (Table 5).

#### Factors predicting outcome

The most important predictors of subjective failure at 2 years following primary surgery in the lasso logistic regression model in order were below the median on all KOOS subscale scores at the time of surgery, cartilage injury at the time of surgery, activity leading to injury, previous surgery on the same knee, KOOS Sports and QoL scores at surgery, body mass index (BMI) at surgery, and age at injury. In the random forest, predictors in the top third by variable importance score also included age at surgery, graft choice, years between injury and surgery, fixation device combination, and femur fixation. The GAM and GBM produced similar rankings of feature importance (Fig. 1). The lasso logistic regression and GAM measure feature importance by effect size associated with the variable. The other models use the difference in model error rate where the feature is to be removed.

#### Risk-prediction calculator

The GAM was selected to create an easy-to-use in-clinic calculator to predict the risk of a patient experiencing a subjective failure at 2 years of follow-up after primary ACL reconstruction ([https://swastvedt.shinyapps.io/calculator\\_koosqol/](https://swastvedt.shinyapps.io/calculator_koosqol/) and Fig. 3). The GAM was chosen out of the four models because it combines performance with simplicity, using fewer predictor variables than the similarly performing GBM. Whereas the overall risk of failure in the register was 22%, this calculator can quantify the risk at a patient-specific level (Video 1).

#### Discussion

The most important finding of this study was that machine learning analysis of a knee ligament register allows the creation of a validated algorithm to predict a patient's risk of experiencing subjective failure of ACL reconstruction with fair accuracy. Additionally, despite having 20 possible prognostic variables contained within the NKLR, the algorithm required only eight factors for the prediction of 2-year risk. Variables required for risk prediction include age at injury, pre-operative KOOS subscale scores, activity leading to an ACL injury, concomitant cartilage injury, history of previous surgery on the same knee, and pre-operative BMI. Using this algorithm, we developed an in-clinic calculator was developed that can estimate the risk of subjective failure.

This represents the first machine learning model for predicting the subjective outcome of ACL reconstruction at a patient-specific level. Estimation of revision risk has been developed previously [9], and together, these two prediction tools can be used to guide the discussion surrounding the surgical options and realistic outcome goals at a patient-specific level. For the clinician, this represents a valuable adjunct to the assessment of patients with ACL deficiency desiring surgical management.

Similar to the previous study of revision risk [9], four models were used to analyse the NKLR and create algorithms predicting the risk of subjective failure after ACL reconstruction. Discrimination (AUC) was similar for the prediction of subjective outcome evaluated with this study

Table 2b

Random forest/gradient boosted regression complete/incomplete case comparison.

Variable*	Incomplete N = 15,040	Complete N = 5,778	Total N = 20,818	P-value**
Years: surgery to data current date (2020-01-12)	9.0 (4.2)	7.5 (2.5)	8.6 (3.9)	<0.001
KOOS QoL <44 at 2 years	1,329 (23%)	1,227 (21%)	2,556 (22%)	0.058
Missing	9,188	0	9,188	
Age at surgery	28 (10)	28 (11)	28 (10)	0.19
Age at injury	26 (10)	27 (11)	26 (10)	0.006
Missing	1,072	0	1,072	
Sex				<0.001
Male	8,890 (59%)	2,779 (48%)	11,669 (56%)	
Female	6,150 (41%)	2,999 (52%)	9,149 (44%)	
Pre-surgery BMI	25.1 (3.7)	24.8 (3.7)	25.0 (3.7)	<0.001
Missing	7,244	0	7,244	
Pre-surgery KOOS QoL score (out of 10)	3.48 (1.82)	3.56 (1.85)	3.50 (1.83)	0.006
Missing	4,022	0	4,022	
Pre-surgery KOOS Sports score (out of 10)	4.28 (2.71)	4.43 (2.71)	4.33 (2.71)	0.001
Missing	4,162	0	4,162	
Below median on all pre-surgery KOOS scores	2,244 (20%)	1,041 (18%)	3,285 (19%)	0.001
Missing	3,893	0	3,893	
Activity that led to injury				<0.001
Non-pivoting	2,846 (20%)	1,263 (22%)	4,109 (20%)	
Pivoting	8,564 (59%)	3,443 (60%)	12,007 (59%)	
Other	3,159 (22%)	1,072 (19%)	4,231 (21%)	
Missing	471	0	471	
Meniscus injury	7,908 (53%)	3,034 (53%)	10,942 (53%)	0.940
Cartilage injury				0.031
ICRS 1-2	2,683 (18%)	942 (16%)	3,625 (17%)	
ICRS 3-4	710 (4.7%)	283 (4.9%)	993 (4.8%)	
None	11,647 (77%)	4,553 (79%)	16,200 (78%)	
Graft choice				<0.001
BPTB autograft	5,454 (36%)	1,880 (33%)	7,334 (35%)	
Hamstring autograft	9,358 (62%)	3,839 (66%)	13,197 (63%)	
Other	228 (1.5%)	59 (1.0%)	287 (1.4%)	
Tibia fixation device				<0.001
Interference screw	12,494 (87%)	5,399 (93%)	17,893 (89%)	
Suspension/cortical device	1,700 (12%)	373 (6.5%)	2,073 (10%)	
Other	146 (1.0%)	6 (0.1%)	152 (0.8%)	
Missing	700	0	700	
Femur fixation device				<0.001
Interference screw	4,671 (32%)	1,654 (29%)	6,325 (31%)	
Suspension/cortical device	7,817 (53%)	3,812 (66%)	11,629 (57%)	
Other	2,172 (15%)	312 (5.4%)	2,484 (12%)	
Missing	380	0	380	
Fixation device combination				<0.001
Interference screw x2	4,391 (31%)	1,637 (28%)	6,028 (30%)	
Interference/suspension	40 (0.3%)	11 (0.2%)	51 (0.3%)	
Suspension/interference	6,177 (43%)	3,458 (60%)	9,635 (48%)	
Suspension/cortical device x2	1,292 (9.1%)	354 (6.1%)	1,646 (8.2%)	
Other	2,316 (16%)	318 (5.5%)	2,634 (13%)	
Missing	824	0	824	
Injured side				0.18
Right	7,711 (51%)	2,902 (50%)	10,613 (51%)	
Left	7,329 (49%)	2,876 (50%)	10,205 (49%)	
Previous surgery on opposite knee	1,157 (7.7%)	369 (6.4%)	1,526 (7.3%)	0.001
Previous surgery on same knee	2,856 (19%)	928 (16%)	3,784 (18%)	<0.001
Time injury to surgery (years)	1.72 (3.31)	1.68 (3.50)	1.71 (3.36)	0.42
Missing	1,076	0	1,076	
Systemic antibiotic prophylaxis	14,897 (99%)	5,772 (100%)	20,669 (100%)	<0.001
Missing	51	0	51	

\*Statistics presented: Mean (SD); n (%).

\*\*Statistical tests performed: t-test, chi-square test.

(0.65–0.68) compared with the revision risk prediction (0.67–0.69), and all models except the random forest demonstrated appropriate calibration. It is interesting to note that while the factors used for predicting revision risk included modifiable surgical details (graft choice, femoral fixation device, and length of time between injury and surgery) [9], the prediction of subjective failure appears to be static. That is, most of the variables used to predict subjective outcome are based on patient-driven factors that are present prior to surgery (age, concomitant chondral injury, history of previous surgery, and activity leading to injury) and may not be amenable to optimisation.

Of the variables identified by the algorithm as important for predicting the risk of subjective failure, the only truly modifiable factor was patient BMI at the time of surgery. The extent to which efforts to decrease BMI prior to surgery may influence the risk of poor functional outcomes is unclear and raises an interesting area for future study. Similarly, given the impact of the pre-surgical KOOS scores on the eventual post-operative subjective outcome, efforts to optimise functional outcomes prior to surgery through physiotherapy or cognitive behavioural coaching may also be beneficial. Regarding variable relative importance (Fig. 1), BMI was the least important variable in the GAM, while KOOS QoL had the

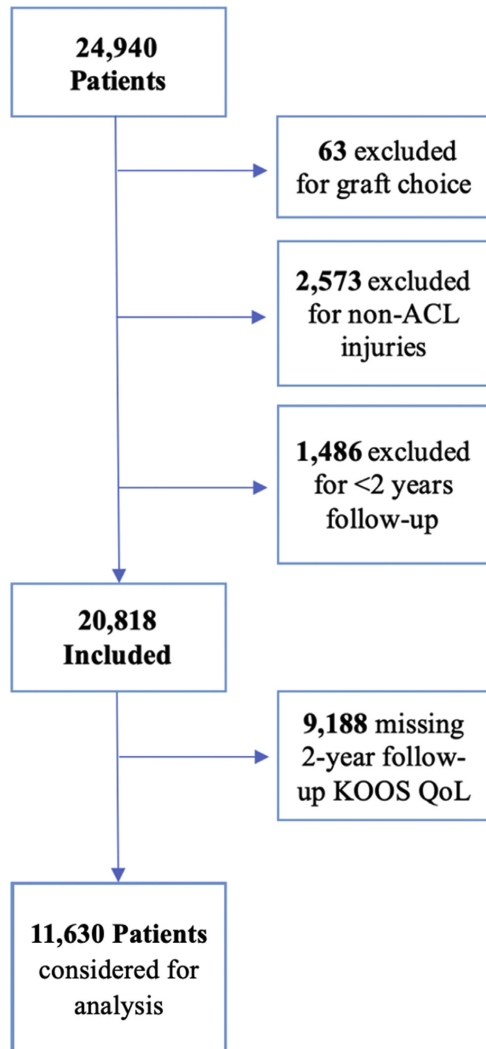


Fig. 2. Patient inclusion flowchart.

highest relative importance. It should be noted, however, that the present study was designed to predict subjective failure risk and does not represent a comparative study to determine the effect of risk factor modification.

**Table 3**  
Inverse probability weighted model performance.

Model	AUC	Weighted calibration statistic	Unweighted calibration	Calibration p-value (unweighted)
Logistic regression (lasso)	0.67	0.020	4.33	0.228
Random forest	0.65	0.054	24.65	<0.001
Gradient boosted regression	0.67	0.017	6.65	0.084
Generalised additive model	0.67	0.019	7.45	0.059

**Table 4**  
Multiple imputation model performance.

Model	AUC	Calibration statistic	Calibration p-value
Logistic regression (lasso)	0.68	2.54	0.468
Random forest	0.67	21.30	0.006
Gradient boosted regression	0.69	1.62	0.656
Generalised additive model	0.68	2.46	0.482

**Table 5**  
Model performance.

Model	AUC	AUC confidence Interval	Calibration statistic	Calibration p-value
Logistic regression (lasso)	0.67	(0.64, 0.71)	4.57	0.206
Random forest	0.65	(0.62, 0.69)	26.83	<0.001
Gradient boosted regression	0.68	(0.64, 0.71)	4.03	0.258
Generalised additive model	0.68	(0.64, 0.71)	4.74	0.192

The primary outcome of the subjective failure of ACL reconstruction was defined as a KOOS QoL score of <44. Other possible measures of subjective outcome include, but are not limited to, the minimal clinically important difference (MCID) or Patient Acceptable Symptom State and may use other assessment tools such as a visual analogue scale or the International Knee Documentation Committee questionnaire. While there are advantages and disadvantages to each measure of functional outcome, KOOS QoL was selected for this study since it has previously been validated as a measure of inadequate knee function associated with prospective ACL reconstructed graft failure and represents a poor outcome after surgery [11]. Further, the prevalence of a KOOS QoL score of <44 was 22%, which suggests that the outcome is clinically relevant across the population.



**Fig. 3.** QR Code for 2-year subjective failure point-of-care risk stratification at the time of primary ACL reconstruction. ACL, anterior cruciate ligament.



### Limitations

The most significant limitation of this study is the missing follow-up KOOS data. Whereas overall compliance with the NKLR is 86% for tracking revision surgery following ACL reconstruction [10], follow-up KOOS scores were only available for 56% of patients at 2 years. While we cannot determine that data were missing completely at random, the inverse probability weighted analysis does provide evidence that the group of patients with complete KOOS follow-up data was not meaningfully different from the group with missing data based on recorded characteristics. Complete PROM follow-up represents a challenge for all national knee ligament registers since patients are typically young and reside throughout the country. Patient compliance is typically higher when research teams and surgeons are actively engaged in the data collection [2], which is not feasible for a large national register like the NKLR. Second, although several machine learning models were evaluated, a model that not considered may have performed better. A third limitation is the fact that the analysis was limited to the variables contained within the register. Although these variables included several known risk factors for ACL reconstruction failure, there are also many other factors that may be associated with the poor outcome that are not recorded in the NKLR. Examples include radiographic variables such as tibial slope and coronal alignment [24–28], physical examination and rehabilitation details [29,30], and surgical technique factors such as tunnel position [31] and graft size [32,33]. Further, while meniscus and chondral injuries were recorded, the surgical treatments employed at the time of surgery were not included as variables and may represent a source of exclusion bias.

There are also limitations regarding the clinical utility of this analysis. The machine learning models use several variables for outcome prediction. To account for this, the GAM was selected for the in-clinic calculator due to its simplicity, requiring fewer input variables without a significant decrease in performance versus the more complex models. Further, this study included patients from a single national register, and the results may not be applicable to other populations. External validity could be established through the evaluation of model performance when applied to patients from other registers or databases. While an advantage of registers like the NKLR is the generalisability and real-world applicability [34], the inclusion of all Norwegian surgeons in the data collection may result in wide variability. Finally, while the machine learning algorithm was well calibrated, the AUC was fair. The accuracy of the model may be improved if radiographic, rehabilitation, and/or other variables not included in the model were assessed.

### Conclusion

Machine learning analysis of a national knee ligament register can predict subjective failure risk following ACL reconstruction with few factors required for outcome prediction and moderate accuracy overall. This algorithm supports the creation of an easy-to-use in-clinic calculator for point-of-care risk stratification. Clinicians can use this calculator to estimate subjective failure risk at a patient-specific level when discussing outcome expectations pre-operatively.

### Institutional review board

Approval not required as consent was obtained by all patients at time of enrolment in the national knee ligament register.

### Sources of funding

This study was funded by a Norwegian Centennial Chair Seed Grant.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jisako.2021.12.005>.

### References

- [1] An VV, Scholes C, Mhaskar VA, et al. Limitations in predicting outcome following primary ACL reconstruction with single-bundle hamstring autograft — a systematic review. *Knee* 2017;24:170–8. <https://doi.org/10.1016/j.knee.2016.10.006>.
- [2] Marx Robert G, Wolfe Isabel A, Turner Brooke E, et al., MOON Knee Group. MOON's strategy for obtaining over eighty percent follow-up at 10 years following ACL reconstruction. *J Bone Joint Surg Am* 2021 Aug 23. <https://doi.org/10.2106/JBJS.21.00166>.
- [3] Nguyen JT, Wasserstein D, Reinke EK, et al. Does the chronicity of anterior cruciate ligament ruptures influence patient-reported outcomes before surgery? *Am J Sports Med* 2017;45:541–9. <https://doi.org/10.1177/0363546516669344>.
- [4] Brophy RH, Huston LJ, Briskin I, et al. Articular cartilage and meniscus predictors of patient-reported outcomes 10 years after anterior cruciate ligament reconstruction: a multicenter cohort study. *Am J Sports Med* 2021;49:2878–88. <https://doi.org/10.1177/03635465211028247>.
- [5] Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. *Front Bioeng Biotechnol* 2018;6:75. <https://doi.org/10.3389/fbioe.2018.00075>.
- [6] Van Eetvelde H, Mendonça LD, Ley C, et al. Machine learning methods in sport injury prediction and prevention: a systematic review. *J Exp Orthop* 2021;8:27. <https://doi.org/10.1186/s40634-021-00346-x>.
- [7] 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:e200198. <https://doi.org/10.1148/ryai.2020200198>.
- [8] Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. *Radiol Artif Intell* 2020;2:e190023. <https://doi.org/10.1148/ryai.2020190023>.
- [9] 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* 2021;104(2):145–53. <https://doi.org/10.2106/JBJS.21.00113>.
- [10] Annual report. Bergen, Norway: Norwegian national advisory unit on arthroplasty and hip fractures. 2021.
- [11] Granan L-P, Baste V, Engebretsen I, et al. Associations between inadequate knee function detected by KOOS and prospective graft failure in an anterior cruciate ligament-reconstructed knee. *Knee Surg Sports Traumatol Arthrosc* 2015;23:1135–40. <https://doi.org/10.1007/s00167-014-2925-5>.
- [12] LaPrade CM, Dornan GJ, Granan L-P, et al. Outcomes after anterior cruciate ligament reconstruction using the Norwegian knee ligament registry of 4691 patients: how does meniscal repair or resection affect short-term outcomes? *Am J Sports Med* 2015;43:1591–7. <https://doi.org/10.1177/0363546515577364>.
- [13] Persson A, Fjeldsgaard K, Gjertsen J-E, et al. Increased risk of revision with hamstring tendon grafts compared with patellar tendon grafts after anterior cruciate ligament reconstruction: a study of 12,643 patients from the Norwegian Cruciate Ligament Registry, 2004–2012. *Am J Sports Med* 2014;42:285–91. <https://doi.org/10.1177/0363546513511419>.
- [14] Persson A, Kjellsen AB, Fjeldsgaard K, et al. Registry data highlight increased revision rates for endobutton/biosure HA in ACL reconstruction with hamstring tendon autograft: a nationwide cohort study from the Norwegian Knee Ligament Registry, 2004–2013. *Am J Sports Med* 2015;43:2182–8. <https://doi.org/10.1177/0363546515584757>.
- [15] Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55–63. <https://doi.org/10.7326/M14-0697>.
- [16] Granan L-P, Bahr R, Steindal K, et al. Development of a national cruciate ligament surgery registry: the Norwegian National Knee Ligament Registry. *Am J Sports Med* 2008;36:308–15. <https://doi.org/10.1177/0363546507308939>.
- [17] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. <https://www.R-project.org/>. [Accessed 19 May 2020].
- [18] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Software* 2010;33(1):1–22. <https://doi.org/10.18637/jss.v033.i01>.
- [19] Breiman L. Random forests. *Mach Learn* 2001;45:5–32. <https://doi.org/10.1023/A:1010933404324>.
- [20] Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;29. <https://doi.org/10.1214/aos/1013203451>.



- [21] Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal* 2002;38: 367–78. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2).
- [22] Wood SN. Generalized additive models: an introduction with R. 2nd ed. Chapman and Hall/CRC; 2017. <https://doi.org/10.1201/9781315370279>.
- [23] Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat Theor Methods* 1980;9:1043–69. <https://doi.org/10.1080/03610928008827941>.
- [24] Bernholt DL, Dornan GJ, DePhillipo NN, et al. High-grade posterolateral tibial plateau impaction fractures in the setting of a primary anterior cruciate ligament tear are correlated with an increased preoperative pivot shift and inferior postoperative outcomes after anterior cruciate ligament reconstruction. *Am J Sports Med* 2020;48:2185–94. <https://doi.org/10.1177/0363546520932912>.
- [25] Bayer S, Meredith SJ, Wilson KW, et al. Knee morphological risk factors for anterior cruciate ligament injury: a systematic review. *J Bone Jt Surg* 2020;102:703–18. <https://doi.org/10.2106/JBJS.19.00535>.
- [26] Li Y, Hong L, Feng H, et al. Posterior tibial slope influences static anterior tibial translation in anterior cruciate ligament reconstruction: a minimum 2-year follow-up study. *Am J Sports Med* 2014;42:927–33. <https://doi.org/10.1177/0363546514521770>.
- [27] Bernhardt AS, Aman ZS, Dornan GJ, et al. Tibial slope and its effect on force in anterior cruciate ligament grafts: anterior cruciate ligament force increases linearly as posterior tibial slope increases. *Am J Sports Med* 2019;47:296–302. <https://doi.org/10.1177/0363546518820302>.
- [28] 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 Off J ESSKA* 2020;28:1119–29. <https://doi.org/10.1007/s00167-019-05770-2>.
- [29] Roe C, Jacobs C, Kline P, et al. Correlations of single-leg performance tests to patient-reported outcomes after primary anterior cruciate ligament reconstruction. *Clin J Sport Med Off J Can Acad Sport Med* 2021;31:e265–70. <https://doi.org/10.1097/JSM.0000000000000780>.
- [30] Grindem H, Snyder-Mackler L, Moksnes H, et al. Simple decision rules can reduce reinjury risk by 84% after ACL reconstruction: the Delaware-Oslo ACL cohort study. *Br J Sports Med* 2016;50:804–8. <https://doi.org/10.1136/bjsports-2016-096031>.
- [31] Liu A, Sun M, Ma C, et al. Clinical outcomes of transtibial versus anteromedial drilling techniques to prepare the femoral tunnel during anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc Off J ESSKA* 2017;25:2751–9. <https://doi.org/10.1007/s00167-015-3672-y>.
- [32] Conte EJ, Hyatt AE, Gatt CJ, et al. Hamstring autograft size can be predicted and is a potential risk factor for anterior cruciate ligament reconstruction failure. *Arthrosc J Arthrosc Relat Surg* 2014;30:882–90. <https://doi.org/10.1016/j.arthro.2014.03.028>. Off Publ Arthrosc Assoc N Am Int Arthrosc Assoc.
- [33] Magnussen RA, Lawrence JTR, West RL, et al. Graft size and patient age are predictors of early revision after anterior cruciate ligament reconstruction with hamstring autograft. *Arthrosc J Arthrosc Relat Surg* 2012;28:526–31. <https://doi.org/10.1016/j.arthro.2011.11.024>. Off Publ Arthrosc Assoc N Am Int Arthrosc Assoc.
- [34] Naylor CD, Guyatt GH. Users' guides to the medical literature. X. How to use an article reporting variations in the outcomes of health services. The Evidence-Based Medicine Working Group. *JAMA* 1996;275(7):554–8. <https://doi.org/10.1001/jama.275.7.554>.

## Appendix A: Machine Learning Models

### Lasso logistic regression

The Lasso logistic regression model applies Lasso (L1) regularization to a logistic regression model for binary 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 Lasso using the *glmnet* package in R, with the tuning parameter selected via cross-validation to balance model simplicity and fit.<sup>1</sup>

### Random Forest

The random forest, as implemented in the *randomForest* R package, uses an ensemble tree method designed for classification of binary outcome data. Each tree uses a bootstrap sample of the data, with variables randomly selected for splitting at each node in the tree. 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 the overall OOB error rate for all trees in the forest.<sup>2</sup>

### Generalized additive model

A generalized additive model (GAM) is a regression model that allows for non-linear relationships between predictors and the outcome. In the R package *mgcv*, which we used for our model, smooth terms are fit using penalized regression splines. The generalized additive model fits a logistic regression model, suitable for binary data.<sup>3</sup>

### Gradient boosted regression

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.<sup>4,5</sup>

## REFERENCES

1. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010;33(1):1-22. doi:10.18637/jss.v033.i01
2. Breiman, L. Random Forests. *Mach Learn.* 2001;45(1):5-32. doi: 10.1023/A:1010933404324
3. Wood SN. *Generalized Additive Models: An Introduction with R.* 2nd ed. Chapman and Hall/CRC; 2017. doi:10.1201/9781315370279
4. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat.* 2001;29(5). doi:10.1214/aos/1013203451

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





## Paper III

Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Lind M, Engebretsen L. Ceiling Effect of the Combined Norwegian and Danish Knee Ligament Registers Limits Anterior Cruciate Ligament Reconstruction Outcome Prediction. *Am J Sports Med.* 2023;51(9):2324-2332. doi:10.1177/03635465231177905





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

R. Kyle Martin,<sup>\*†‡</sup> MD , Solvejg Wastvedt,<sup>§</sup> BA, Ayoosh Pareek,<sup>||</sup> MD, Andreas Persson,<sup>¶\*\*\*</sup> MD, PhD, Håvard Visnes,<sup>\*\*</sup> MD, PhD, Anne Marie Fenstad,<sup>\*\*</sup> MSc, Gilbert Moatshe,<sup>¶#</sup> MD, PhD, Julian Wolfson,<sup>§</sup> PhD, Martin Lind,<sup>††</sup> MD, PhD, and Lars Engebretsen,<sup>¶#</sup> MD, PhD   
*Investigation performed at the University of Minnesota, Minneapolis, Minnesota, USA*

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

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

**Study Design:** Cohort study; Level of evidence, 3.

**Methods:** Machine learning analysis was performed on combined data from the NKLR and DKRR. 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%). There were 4 machine learning models examined: Cox lasso, random survival forest, gradient boosting, and super learner. Concordance and calibration were calculated for all 4 models.

**Results:** The data set included 62,955 patients in which 5% underwent a revision surgical procedure with a mean follow-up of 7.6 ± 4.5 years. The 3 nonparametric models (random survival forest, gradient boosting, and super learner) performed best, demonstrating moderate concordance (0.67 [95% CI, 0.64-0.70]), and were well calibrated at 1 and 2 years. Model performance was similar to that of the previously published model (NKLR-only model: concordance, 0.67-0.69; well calibrated).

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

**Keywords:** ACL revision; outcome prediction; machine learning; artificial intelligence

There has been an increased focus on outcome prediction using machine learning in orthopaedic surgery recently.<sup>22</sup> The primary goal of these early clinical predictive models was to enable patient-specific risk estimation to guide management discussions and expectations. Clinical tools based on machine learning analysis now exist for outcome prediction after anterior cruciate ligament reconstruction (ACLR) including revision surgery<sup>30</sup> and inferior patient-reported outcomes.<sup>31</sup> These models were developed from

analyses of the Norwegian Knee Ligament Register (NKLR), and the revision prediction model has also been externally validated using the Danish Knee Ligament Reconstruction Registry (DKRR).<sup>32</sup>

The accurate prediction of outcomes after 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 his or her 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 predictive algorithms, the general principle is that more data may lead to improved model accuracy. The rationale for this is that more data present more opportunity for the models to “learn” the association between predictors and outcomes. Therefore, the purpose of this study was to apply machine learning to a combined NKLR and DKRR data set, with the aim of predicting revision surgery with improved accuracy relative to a previously published model.<sup>30</sup> The original NKLR model was developed using machine learning analysis of approximately 25,000 patients, whereas the combined NKLR and DKRR data set includes nearly 63,000 patients. The hypothesis was that the additional patient data would result in a more accurate prediction of the revision ACLR risk.

## METHODS

This article was written in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis statement.<sup>6</sup> 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 provided informed consent for the NKLR, and the Norwegian Data Protection Authority granted permission for the register to collect, analyze, and publish health data. Data registration was performed confidentially according to European Union data protection rules, with all data de-identified before retrieval. The regional ethics committee stated that it was not necessary to obtain further ethical approval.<sup>11</sup> Similarly, the DKRR obtained informed consent at the time of enrollment, and patient data were de-identified before retrieval with no further ethical approval required.

## Data Compilation

Patients who underwent primary ACLR between June 2004 and December 2020 were included. Patients missing

data for graft choice, those with a graft choice recorded as “direct suture,” and those missing data for the indicator of revision surgery were excluded. Variables considered for analysis are shown in Table 1.

A predictor indicating if a patient scored below the median score in the respective registry for all preoperative Knee injury and Osteoarthritis Outcome Score (KOOS) subscales was created. Patients who underwent revision ACLR before the follow-up time were considered to have experienced the event.

## Machine Learning Modeling

NKLR and DKRR data were combined and then randomly split into training (75%) and test (25%) sets used to fit and evaluate the models, respectively. The primary outcome was the probability of revision ACLR within 1, 2, and 5 years. R (Version 4.1.11; R Core Team) was used to fit machine learning models that were 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 endpoint. For example, a patient who has been revision-free for 4 years has not yet reached the 5-year selected outcome time point, but his or her revision-free time can still be considered in the analysis for the 5-year revision risk. Censoring also accounts for the fact that patients who have not yet undergone a revision procedure may ultimately undergo revision surgery in the future.

Four models intended for this type of data were used: Cox lasso, random survival forest, gradient boosting, and super learner. These models represent a range of approaches regarding the flexibility of model fitting and the number of variables incorporated. Cox lasso is a semi-parametric, penalized regression model that selects a subset of the most important predictor variables for inclusion.<sup>41</sup> Random survival forest is a nonparametric model, meaning that it does not require prespecification 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.<sup>17</sup> Gradient boosting is also a tree-based, nonparametric model adapted for censored data; this model iteratively updates to improve the fit using all available variables.<sup>9</sup> Super learner is an

\*Address correspondence to R. Kyle Martin, MD, Department of Orthopedic Surgery, University of Minnesota, 2512 South 7th Street, Suite R200, Minneapolis, MN 55455, USA (email: rkylemartin@gmail.com) (Twitter: @RKMartin6).

<sup>†</sup>Department of Orthopedic Surgery, University of Minnesota, Minneapolis, Minnesota, USA.

<sup>‡</sup>Department of Orthopedics, CentraCare, St Cloud, Minnesota, USA.

<sup>§</sup>Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA.

<sup>||</sup>Department of Orthopedic Surgery, Hospital for Special Surgery, New York, New York, USA.

<sup>¶</sup>Department of Orthopaedic Surgery, Oslo University Hospital Ullevål, Oslo, Norway.

<sup>\*\*</sup>Oslo Sports Trauma Research Center, Norwegian School of Sport Sciences, Oslo, Norway.

<sup>††</sup>Norwegian Knee Ligament Register, Haukeland University Hospital, Bergen, Norway.

<sup>†††</sup>Aarhus University Hospital, Aarhus, Denmark.

Submitted October 3, 2022; accepted April 11, 2023.

One or more of the authors has declared the following potential conflict of interest or source of funding: This study was funded by a Norwegian Centennial Chair seed grant. R.K.M. has received consulting fees from Smith & Nephew and support for education from Gemini/Arthrex. G.M. has received consulting fees from Arthrex and IBSA. M.L. has received consulting fees from Smith & Nephew. L.E. has received research support from Biomet and Health South-Eastern Norway and royalties from Arthrex and Smith & Nephew. AOSSM checks author disclosures against the Open Payments Database (OPD). AOSSM has not conducted an independent investigation on the OPD and disclaims any liability or responsibility relating thereto.



TABLE 1  
Patient and Surgical Characteristics<sup>a</sup>

	Value (N = 62,955)
Revision	3205 (5)
Follow-up time or time to revision, mean $\pm$ SD, y	7.6 $\pm$ 4.5
Age at surgery, median (IQR), y	26 (20-36)
Age at injury, median (IQR), y	24 (18-34)
Missing, n	1870
Sex	
Male	36,509 (58)
Female	26,446 (42)
Preoperative KOOS–Quality of Life score (of 10), mean $\pm$ SD	3.63 $\pm$ 1.80
Missing, n	29,512
Preoperative KOOS–Sport score (of 10), mean $\pm$ SD	4.12 $\pm$ 2.69
Missing, n	29,708
All preoperative KOOS scores below median	6372 (19)
Missing, n	29,323
Activity that led to injury	
Nonpivoting	20,391 (32)
Pivoting	35,851 (57)
Other	6162 (10)
Missing, n	551
Meniscal injury	
Injury without repair	20,328 (32)
Injury with repair	10,554 (17)
None	32,061 (51)
Missing, n	12
Cartilage injury	
Grade 1-2	8766 (14)
Grade 3-4	3223 (5)
None	50,878 (81)
Missing, n	88
Graft choice	
Bone–patellar tendon–bone	15,639 (25)
Hamstring tendon	43,518 (69)
Quadriceps tendon	2520 (4)
Other	1278 (2)
Tibial fixation device	
Interference screw	55,792 (89)
Suspension/cortical device	3643 (6)
Other	2356 (4)
Missing, n	1164
Femoral fixation device	
Interference screw	16,434 (26)
Suspension/cortical device	39,742 (63)
Other	4822 (8)
Missing, n	1957
Fixation device combination	
2 interference screws	15,865 (25)
Interference screw (femur) and suspension device (tibia)	236 (0.4)
2 suspension/cortical devices	2994 (5)
Suspension device (femur) and interference screw (tibia)	34,895 (55)
Other	6529 (10)
Missing, n	2436
Injured side	
Right	32,147 (51)

(continued)

TABLE 1  
(continued)

	Value (N = 62,955)
Left	30,807 (49)
Missing, n	1
Previous surgery on opposite knee	4839 (8)
Missing, n	2946
Previous surgery on same knee	10,312 (16)
Missing, n	673
Time from injury to surgery, median (IQR), y	0.61 (0.33-1.32)
Missing, n	2083
Registry	
DKRR	34,554 (55)
NKLRL	28,401 (45)

<sup>a</sup>Data are reported as n (%) unless otherwise indicated. DKRR, Danish Knee Ligament Reconstruction Registry; IQR, interquartile range; KOOS, Knee injury and Osteoarthritis Outcome Score; NKLRL, Norwegian Knee Ligament Register.

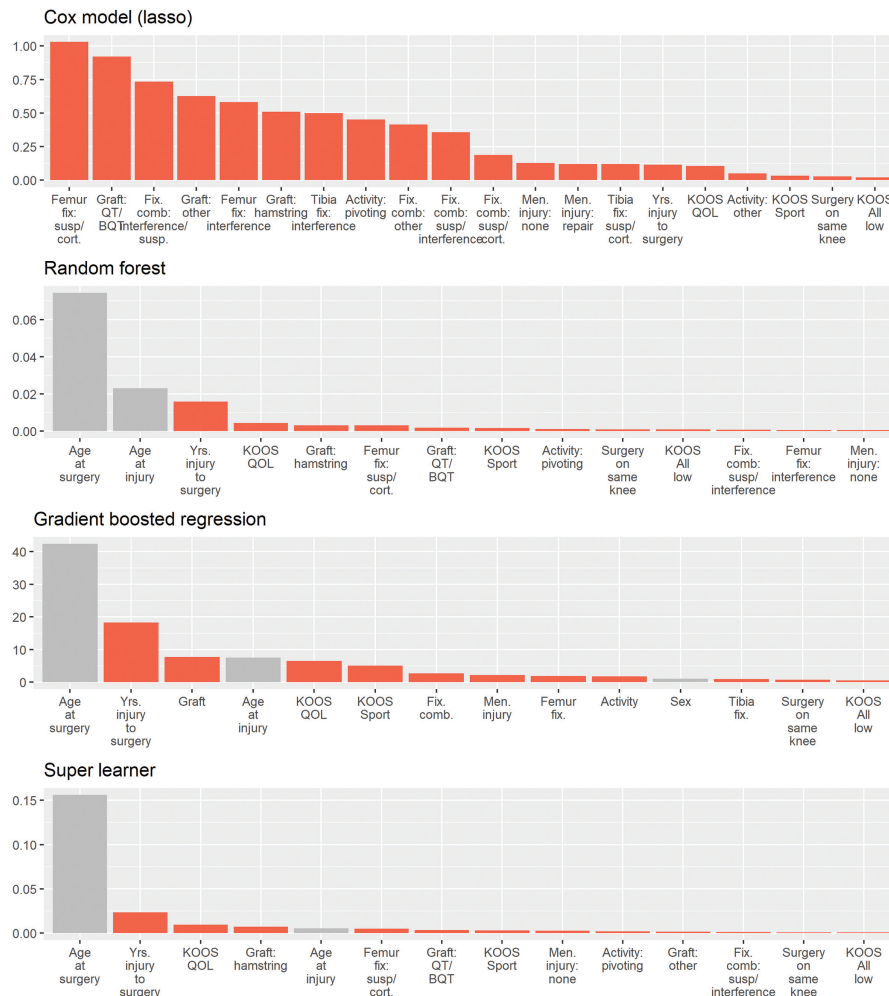
“ensemble” model that creates a weighted average of other machine learning techniques, combining them into 1 overall fit and thereby providing an even more flexible approach<sup>46</sup>; the super learner model combines the random survival forest and gradient boosting models. Further descriptions of each model are included in the Appendix (available in the online version of this article).

Variables with nonzero 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.

For the random survival forest, gradient boosting, 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. The random survival forest model (package *randomForestSRC*) was trained using the following hyperparameters: node size of 300, 10 variables per split, and 500 trees. The gradient boosting model (package *gbm*) was trained using a shrinkage parameter of 0.01, interaction depth of 3, minimum node size of 100, and 1,000 trees. The super learner model was trained using the same hyperparameter values for the random survival forest and gradient boosting models and utilizing the *SuperModel* function (package *MachineShop*) to determine, via cross-validation, the optimal weighting of the component models. All 4 models were restricted to patients with complete data for the predictors used (see Table 1 and Missing Data section).

## 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



**Figure 1.** The 4 plots show the relative feature importance in each of the machine learning models. The highlighted bars indicate features selected for the Cox lasso model. The random survival forest, gradient boosting, and super learner plots show features in the top half according to the 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 model measures feature importance by absolute effect size. The random survival forest and super learner models use permutation-based importance, which measures the relative change in model performance after randomly permuting values of the given feature. The gradient boosting model uses the difference in the error rate if the feature was to be removed, normalized to a total sum of 100. BQT, quadriceps tendon autograft with bone; comb, combined; cort, cortical; fix, fixation; KOOS, Knee injury and Osteoarthritis Outcome Score; Men, meniscus; QOL, Quality of Life; QT, quadriceps tendon autograft; Sport, Sport and Recreation Subscale; susp, suspension; Yrs, years.

calculated using methods adapted for censored data. Concordance was determined using the Harrell C-index at 1-, 2-, and 5-year follow-up. The C-index is a generalization of the common area under the receiver operating

characteristic curve metric. As with the area under the curve, it ranges from 0 to 1, with 1 indicating perfect concordance. The C-index measures the proportion of pairs of observations in which predicted rankings of survival

TABLE 2  
Model Performance With Complete Case Training Data

	Concordance (95% CI)	Calibration Statistic	Calibration <i>P</i> Value
1 y			
Cox lasso	0.59 (0.56-0.61)	7.19	.066
Random survival forest	0.67 (0.64-0.69)	5.54	.136
Gradient boosting	0.67 (0.65-0.70)	7.48	.058
Super learner	0.67 (0.65-0.69)	8.67	.034
2 y			
Cox lasso	0.58 (0.56-0.61)	8.17	.043
Random survival forest	0.67 (0.64-0.69)	6.42	.093
Gradient boosting	0.67 (0.64-0.69)	4.53	.210
Super learner	0.67 (0.64-0.69)	4.10	.250
5 y			
Cox lasso	0.58 (0.56-0.61)	11.37	.010
Random survival forest	0.67 (0.65-0.69)	9.27	.026
Gradient boosting	0.67 (0.64-0.69)	11.07	.011
Super learner	0.67 (0.64-0.69)	11.82	.008

probabilities correspond to actual rankings.<sup>14</sup> Furthermore, calculation of the C-index 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 undergone revision surgery and the other is censored (no revision yet, with the 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 outcomes with actual outcomes. We calculated calibration using a version of the Hosmer-Lemeshow test that accounts for censoring.<sup>47</sup> This statistic sums the average misclassification in each predicted risk quintile and converts the sum into a chi-square statistic. Larger values of calibration indicate worse accuracy and correspond to smaller *P* values, with statistical significance indicating a 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 effect of restricting data to complete cases, we re-trained and re-evaluated the models 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. Multiple imputation by chained equations was conducted with 5 imputations on training and test data (package *mice*). The variables with nonzero coefficients for the Cox lasso model with complete cases were used to refit the model with each imputed training data set, averaging predictions over the 5 imputations. The

random survival forest, gradient boosting, and super learner models were similarly refit. A bootstrap procedure was used to compare the calibration between the complete case and multiply imputed models.

## RESULTS

### Patient Data

Table 1 details the characteristics of the population at the time of surgery and shows all variables included for the analysis. After data cleaning, the combined registries' population consisted of 62,955 patients, with 55% from the DKRR and 45% from the NKLR. The primary outcome, revision surgery, occurred in 5% of patients with a mean follow-up of  $7.6 \pm 4.5$  years. The population was 58% male, with a median age at the time of the primary injury of 24 years (interquartile range, 18-34 years) and a median age at the time of surgery of 26 years (interquartile range, 20-36 years).

### Model Performance

The 3 nonparametric models—random survival forest, gradient boosting, and super learner—had moderate concordance (0.67) at all follow-up times, with 95% CIs ranging from 0.64-0.69 to 0.65-0.70 (Table 2).

The Cox lasso model performed more poorly, with a concordance of 0.58-0.59. The Cox lasso model showed moderate evidence of miscalibration (*P* = .01-.043) at 2 and 5 years. The other 3 models were better calibrated, with the exception of the super learner model at 1 year (*P* = .034) and 5 years (*P* = .008). The random survival forest and gradient boosting models also demonstrated moderate evidence of miscalibration at 5 years. Model performance for the original NKLR algorithm demonstrated similar concordance (0.67-0.69) and calibration.<sup>30</sup>

Model performance with imputation is presented in Table 3.

TABLE 3  
Model Performance With Multiply Imputed Training Data

	Concordance (95% CI)	Calibration Statistic	Calibration <i>P</i> Value
1 y			
Cox lasso	0.59 (0.56-0.61)	8.35	.039
Random survival forest	0.66 (0.64-0.69)	4.17	.244
Gradient boosting	0.68 (0.65-0.70)	7.57	.056
Super learner	0.67 (0.65-0.70)	7.99	.046
2 y			
Cox lasso	0.59 (0.56-0.61)	8.81	.032
Random survival forest	0.67 (0.65-0.70)	8.96	.030
Gradient boosting	0.67 (0.65-0.70)	8.98	.030
Super learner	0.67 (0.65-0.70)	8.34	.039
5 y			
Cox lasso	0.58 (0.56-0.61)	8.30	.040
Random survival forest	0.67 (0.65-0.70)	8.95	.030
Gradient boosting	0.67 (0.65-0.69)	11.53	.009
Super learner	0.67 (0.65-0.69)	14.05	.003

Multiply imputed data did not show notable differences from the complete case analysis. The concordance 95% CIs were nearly identical in all cases. Observed calibration ratios from all 4 models were compared with the bootstrap distribution, and all the observed ratios were within the 95% CI. This suggests that there was no significant difference in calibration between the complete case and multiply imputed models.

#### Factors Predicting Outcome

The most important factors predicting revision surgery, according to the 3 best-performing models, included age at the time of surgery and injury, years between injury and surgery, graft choice, and preoperative KOOS–Quality of Life and KOOS–Sport and Recreation scores. Variables in approximately the top half by feature importance in the random survival forest, gradient boosting, and super learner models are shown in the bottom 3 panels of Figure 1. Variables with nonzero 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 the absolute value of the associated effect size. The gradient boosting model uses the difference in the error rate if the feature was to be removed. The random survival forest and super learner models use permutation-based variable importance, measuring the relative change in model performance after randomly permuting values of the given variable.

#### DISCUSSION

Machine learning analysis of the combined NKLR and DKRR enabled the prediction of revision surgery 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.<sup>30,32</sup>

This suggests that the 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 further the evolution of national ACLR registries regarding variable inclusion and data collection.

Machine learning applications within orthopaedic surgery have been increasing at an exponential rate in recent years.<sup>22</sup> These advanced statistical techniques can evaluate large data sets and recognize complex interactions between variables.<sup>28</sup> “Learning” from these interactions, machine learning models can create algorithms capable of predicting outcomes for patients, often at a level of accuracy superior to expert humans.<sup>3,8,37,39,40,45,50</sup>

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 that contributes to the accuracy of a model. Just as important is the quality of the data. If the data set 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 affects the ability of the model to learn and form accurate associations between predictors and outcomes. Techniques such as imputation can address some data quality inadequacies, but there are limits to what may be overcome.<sup>2</sup>

After nearly 20 years of data collection by the NKLR and DKRR, data quantity is superb, with satisfactory completeness and data accuracy.<sup>7,34–36</sup> However, the present study suggests that for an improvement in our ability to predict outcomes based on 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.<sup>11,29</sup> 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 surgeon, which may affect compliance.

Factors that may improve prediction accuracy and could be considered for supplementation in national registers include data regarding radiographic findings,<sup>4,12,13,18,23,33,48</sup> adjunctive surgical procedures, clinical examination results, rehabilitation details,<sup>38</sup> and alternative patient-reported outcome measures such as psychological factors.<sup>5</sup> Preoperative and postoperative radiographic indices could be manually captured, for example, tibial slope and coronal alignment, or included as raw image files that could then be evaluated using computer vision machine learning techniques.<sup>21</sup> 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 outcomes.<sup>1,10,15,16,24,26,42,52</sup> Clinical examination and rehabilitation information such as preoperative knee laxity grade<sup>25,43</sup> could be obtained via third-party sources such as physical therapists or via natural language processing of patient chart notes.<sup>49</sup> Finally, the KOOS may not be the most appropriate patient-reported outcome tool for the patient population, and an alternative measurement of patient function, such as the baseline Marx activity level, could be considered for inclusion in registries moving forward.<sup>19,27</sup>

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 2 main reasons for this. First, reinjury events leading to revision surgery may occur randomly, such as after a slip on ice or a 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. Although discrimination has often been interpreted as performance  $>0.9$  being excellent,  $>0.8$  being good,  $>0.7$  being fair, and  $<0.7$  being poor,<sup>44</sup> most clinically useful algorithms demonstrate performance in the range of 0.65 to 0.80.<sup>51</sup> In fact, discrimination  $>0.8$  for clinical predictive models may represent data mismanagement or model overfitting.<sup>20</sup>

Modeling using combined DKRR and NKLR data revealed some notable differences between the 2 registries. The poor performance of the Cox lasso model is, in part, caused by the fact that when modeled separately, the 2 registry populations led to the 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 because they were able to fit the data with more flexibility. This observation helps explain the fact that although the Cox lasso model was the best model in the previous study of the NKLR,<sup>30</sup> here, the more flexible models performed better.

The present study has some limitations. First, even though several machine learning methods were considered,

it is possible that another model may have performed differently. Second, there was a high proportion of missing preoperative KOOS data (47%, Table 1), and most patients with this missing variable were from the DKRR. Because preoperative KOOS data have been important in predicting outcomes 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 to 2020. Therefore, this analysis may inherit bias related to temporal changes in the revision surgery risk, as surgical indications, techniques, and trends have evolved over time. These changes were not directly accounted for in the present study but likely represent a 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 5 variables. This means that 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 NKLR and DKRR enabled prediction of the revision ACLR risk with moderate accuracy. However, the resulting algorithms were less user-friendly and did not demonstrate superior accuracy in comparison with the previously developed model based on patients from the NKLR alone, despite the analysis of nearly 63,000 patients. This ceiling effect suggests that simply adding more patients to current national knee ligament registers is unlikely to improve predictive capability and may prompt future changes to increase variable inclusion.

## ORCID iDs

R. Kyle Martin  <https://orcid.org/0000-0001-9918-0264>  
Lars Engebretsen  <https://orcid.org/0000-0003-2294-921X>

## REFERENCES

- 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.
- Buuren SV, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1-67.
- 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.
- Christensen JJ, Krych AJ, Engesser 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 Korskårs Register Rekonstruktions Register Årsrapport 2020/2021. Dansk Korskårsregister, 2021.
  8. Esteve 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, Macchiarella 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 Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA.* 1982;247(18):2543-2546.
  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, Athie 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, Krivich LM, Clapp IM, et al. Machine learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: a systematic review. *Arthroscopy.* 2022;38(6):2090-2105.
  23. 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.
  24. 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.
  25. 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.
  26. 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.
  27. 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 Relat Res.* 2022;480(7):1342-1350.
  28. 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.
  29. 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; 2019:359-369.
  30. 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.* 2022;104(2):145-153.
  31. 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.* 2022;7(3):1-9.
  32. 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.* 2022;30(2):368-375.
  33. Mehl J, Otto A, Kia C, et al. Osseous valgus alignment and postero-medial ligament complex deficiency lead to increased ACL graft forces. *Knee Surg Sports Traumatol Arthrosc.* 2020;28(4):1119-1129.
  34. Middtun 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.
  35. 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.
  36. 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.
  37. 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.* 2021;36(7):2510-2517.e6.
  38. 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.
  39. 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.
  40. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature.* 2020; 577(7792):706-710.
  41. 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):1-13.
  42. 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.
  43. Sonnerby-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.
44. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* 1988;240(4857):1285-1293.
  45. 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.
  46. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol.* 2007;6(1):25.
  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. doi: 10.1177/03635465221119787

## APPENDIX

### Cox Lasso<sup>1</sup>

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<sup>2</sup>

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<sup>3,4</sup>

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<sup>5</sup>

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



## Paper IV

Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Lind M, Engebretsen L. Machine learning algorithm to predict anterior cruciate ligament revision demonstrates external validity. *Knee Surg Sports Traumatol Arthrosc.* 2022;30(2):368-375.  
doi:10.1007/s00167-021-06828-w





## Machine learning algorithm to predict anterior cruciate ligament revision demonstrates external validity

R. Kyle Martin<sup>1,2</sup> · Solvejg Wastvedt<sup>3</sup> · Ayoosh Pareek<sup>4</sup> · Andreas Persson<sup>5,6,7</sup> · Håvard Visnes<sup>5</sup> · Anne Marie Fenstad<sup>5</sup> · Gilbert Moatshe<sup>6,7</sup> · Julian Wolfson<sup>3</sup> · Martin Lind<sup>8</sup> · Lars Engebretsen<sup>6,7</sup>

Received: 10 October 2021 / Accepted: 26 November 2021 / Published online: 1 January 2022  
© The Author(s) 2021

### Abstract

**Purpose** External validation of machine learning predictive models is achieved through evaluation of model performance on different groups of patients than were used for algorithm development. This important step is uncommonly performed, inhibiting clinical translation of newly developed models. Machine learning analysis of the Norwegian Knee Ligament Register (NKLK) recently led to the development of a tool capable of estimating the risk of anterior cruciate ligament (ACL) revision ([https://swastvedt.shinyapps.io/calculator\\_rev/](https://swastvedt.shinyapps.io/calculator_rev/)). The purpose of this study was to determine the external validity of the NKLK model by assessing algorithm performance when applied to patients from the Danish Knee Ligament Registry (DKLR).

**Methods** The primary outcome measure of the NKLK model was probability of revision ACL reconstruction within 1, 2, and/or 5 years. For external validation, all DKLR patients with complete data for the five variables required for NKLK prediction were included. The five variables included graft choice, femur fixation device, KOOS QOL score at surgery, years from injury to surgery, and age at surgery. Predicted revision probabilities were calculated for all DKLR patients. The model performance was assessed using the same metrics as the NKLK study: concordance and calibration.

**Results** In total, 10,922 DKLR patients were included for analysis. Average follow-up time or time-to-revision was 8.4 ( $\pm 4.3$ ) years and overall revision rate was 6.9%. Surgical technique trends (i.e., graft choice and fixation devices) and injury characteristics (i.e., concomitant meniscus and cartilage pathology) were dissimilar between registries. The model produced similar concordance when applied to the DKLR population compared to the original NKLK test data (DKLR: 0.68; NKLK: 0.68–0.69). Calibration was poorer for the DKLR population at one and five years post primary surgery but similar to the NKLK at two years.

**Conclusion** The NKLK machine learning algorithm demonstrated similar performance when applied to patients from the DKLR, suggesting that it is valid for application outside of the initial patient population. This represents the first machine learning model for predicting revision ACL reconstruction that has been externally validated. Clinicians can use this in-clinic calculator to estimate revision risk at a patient specific level when discussing outcome expectations pre-operatively. While encouraging, it should be noted that the performance of the model on patients undergoing ACL reconstruction outside of Scandinavia remains unknown.

**Level of evidence** III.

**Keywords** Machine learning · Artificial intelligence · ACL Reconstruction · ACL revision · Outcome prediction

### Introduction

At the time of primary surgery, how does a surgeon estimate the risk of their patient needing a revision anterior cruciate ligament (ACL) reconstruction in the future? Numerous

studies have defined failure rate epidemiology and identified risk factors such as age [13, 18, 24, 27, 32, 33], graft choice [13, 18, 21] and size [1], activity level [13, 33], body composition [27], ligamentous laxity [14, 18], and tibial slope [10, 31]. Despite this mass of knowledge, the ability to synthesize it and accurately quantify revision risk at a patient-specific level remains elusive and is often influenced by surgeon experience. This uncertainty is rooted in the complex relationships between the known (and unknown) risk factors that may be present to varying degrees in the patient seated

✉ R. Kyle Martin  
rkylemartin@gmail.com

Extended author information available on the last page of the article

in the office. The personal experience of the surgeon combined with their subjective interpretation of these variables in real time leads to the equivalent of an educated guess regarding revision rate.

Machine learning has the potential to add clarity and improve our predictive capability. While relatively new to knee ligament surgery, the application of machine learning is rapidly transforming clinical care in several fields, including orthopaedic surgery. In short, machine learning is a combination of advanced statistical techniques that can interpret large data sets that are more complex than would be possible with traditional statistics. Through analysis of large databases, machine learning can decipher the complex interactions between variables and generate algorithms capable of outcome prediction. Often, the result is accuracy that is comparable to or better than the prediction of experts in the field [5, 8, 23, 25, 26, 29, 34].

Recently, machine learning was used to develop a tool that can quantify revision risk for a patient undergoing primary ACL reconstruction ([https://swastvedt.shinyapps.io/calculator\\_rev/](https://swastvedt.shinyapps.io/calculator_rev/); Fig. 1)[19]. The source of data included nearly 25,000 patients with primary ACL reconstruction recorded in the Norwegian Knee Ligament Register (NKLK). The result was a well-calibrated tool capable of predicting revision risk one, two, and five years after primary ACL reconstruction with moderate accuracy. Following model development, external validation is the next step toward clinical application of new models.



Fig. 1 Link to ACL revision risk prediction in-clinic calculator [19]

The purpose of this study was to determine the external validity of the previously published NKLK ACL revision algorithm by assessing its performance when applied to patients from the Danish Knee Ligament Registry (DKLR). The hypothesis was that model performance would be similar, suggesting validity of the algorithm. This represents the first study to assess external validation of a clinical tool developed using machine learning techniques for outcome prediction following ACL reconstruction. The ability to estimate revision risk at a patient specific level may help guide discussion surrounding outcome expectations pre-operatively.

## 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 TRIPOD statement is a comprehensive set of recommendations for studies that develop and/or validate prediction models. The 22-item checklist aims to improve transparency of prediction model studies through full and clear information reporting, independent of study methods.

## Ethics

At the time of enrollment in the NKLK all patients provide informed consent 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 Norwegian and European Union (EU) data protection rules, with all data de-identified prior to retrieval for analysis. The Regional Ethics Committee (REK) states that it is not necessary to obtain further ethical approval for Norwegian register-based studies [9]. Similarly, the DKLR obtains informed consent at the time of enrollment and patient data was de-identified prior to retrieval for analysis with no further ethical approval required.

## Data source

Original prediction model development was based on machine learning analysis of patients contained within the NKLK while model validation was performed using patients from the DKLR. Both national knee ligament registries prospectively enrol patients undergoing cruciate ligament reconstruction pre-operatively and record demographic, injury, surgical, and follow-up outcome details including subsequent revision reconstruction. The Norwegian registry was established in 2004 and reporting has been mandatory since 2017. Overall compliance with the NKLK was 86% in 2017–18. Patients are registered using their unique

Norwegian national identification number which links identification of subsequent revision surgery performed within Norway, regardless of the provider. The DKLR was founded in 2005 and similarly records longitudinal outcome of ACL reconstruction within Denmark.

### Participants and predictors

In the index study of NKLR patients [19], four machine learning prediction models were assessed for the ability to predict subsequent revision ACL reconstruction after primary surgery. The four models tested were Cox Lasso, survival random forest, generalized additive model, and gradient boosted regression. These four models are among the most commonly used for this type of analysis. The patients in the NKLR were randomly split into training (75%) and test (25%) sets; the algorithm was developed using the training set of patients, and the performance of the algorithm was assessed with the hold-out test set, previously unseen by the models. The Cox Lasso model was the best-performing of the four tested models and was used for the development of an in-clinic revision-risk calculator (Fig. 1).

Regarding outcome prediction, the four models assessed all the available data in the NKLR to “learn” which factors are associated with—and can be used to predict—which patients will eventually undergo revision surgery. Starting with the 24 total predictor variables in the NKLR, the models eliminated variables which do not significantly improve prediction ability, without sacrificing accuracy. The result was an algorithm developed using the Cox Lasso model that only required five variables (out of the 24) for outcome prediction. The model was well calibrated and demonstrated moderate discriminative ability in predicting revision surgery after primary ACL reconstruction [19].

This study sought to validate the previously developed Cox Lasso model from the NKLR. The Cox Lasso model was selected for validation since it was the best performing model and because some of the variables required for the random forest and gradient boosted regression models were not available in the DKLR. Thus, while the full set of patient characteristics are shown in Table 1, only the five predictors selected by the NKLR Cox Lasso model were used in this validation analysis. The five variables required for outcome prediction using the Cox Lasso model were: patient age at primary surgery, KOOS QoL score at primary surgery, graft choice, femur fixation method, and years between injury and ACL reconstruction.

For model validation, patients in the DKLR with primary surgery dates from July 2005 through December 2020 were included ( $N = 34,678$ ). To match variables used in the NKLR model, graft choice and femur fixation device were re-coded as shown in Table 1. New variables were defined for time

**Table 1** Characteristics of Danish registry patients

Variable <sup>a</sup>	$N = 34,678$
Years: surgery to data current date (2021-06-14)	8.3 (4.3)
Missing	1
Revision	1791 (5.2%)
Missing	1
Follow-up time or time to revision	7.6 (4.4)
Missing	1
Age at surgery	29 (10)
Missing	1
Age at injury	27 (10)
Missing	499
Sex	
Female	13,958 (40%)
Male	20,719 (60%)
Missing	1
Pre-surgery KOOS QOL score (out of 10)	3.90 (1.61)
Missing	23,522
Pre-surgery KOOS Sports score (out of 10)	3.80 (2.55)
Missing	23,523
Below median on all pre-surgery KOOS	1868 (17%)
Missing	23,520
Meniscus injury	15,501 (45%)
Cartilage injury	5345 (15%)
Graft choice	
BPTB	3,218 (9.3%)
Hamstring	28,291 (82%)
Unknown/Other	3045 (8.8%)
Missing	124
Tibia fixation device	
Interference screw	30,817 (89%)
Suspension/cortical device	983 (2.8%)
Unknown/Other	2878 (8.3%)
Femur fixation device	
Interference screw	6,072 (18%)
Suspension/cortical device	24,949 (72%)
Unknown/Other	3657 (11%)
Fixation device combination	
Interference screw × 2	5951 (17%)
Interference/Suspension	10 (<0.1%)
Suspension/cortical device × 2	968 (2.8%)
Suspension/Interference	22,308 (64%)
Unknown/Other	5441 (16%)
Injured side	
Right	17,781 (51%)
Left	16,895 (49%)
Missing	2
Previous surgery on opposite knee	2745 (7.9%)
Missing	108
Previous surgery on same knee	28,809 (83%)
Time injury to surgery (years)	1.65 (3.21)
Missing	712

**Table 1** (continued)

Variable <sup>a</sup>	N = 34,678
Systemic antibiotic prophylaxis	34,678 (100%)

<sup>a</sup>Statistics presented: Mean (SD); n (%)

between injury and primary surgery. The Knee Injury and Osteoarthritis Outcome Score (KOOS) Quality of Life (QoL) predictor was scaled to a score out of ten. Patients in the DKLR with missing data for any of the five predictors were excluded from model validation.

### Outcome measures and model performance

The primary outcome in the NKLR Cox Lasso model was probability of revision ACL reconstruction within 1, 2, and/or 5 years. Using R (version: 3.6.1, R Core Team 2019, Vienna, Austria) the NKLR Cox Lasso model was applied to calculate predicted time-to-revision probabilities for all DKLR patients. Performance evaluation included censoring of the time-to-event outcome. “Censoring” refers to the fact that, at any given follow-up time, complete information on outcome is not known for all patients. Some patients have not been in the registry for the requisite number of years, while others have not yet experienced revision and it is unknown when or if they ultimately will.

Performance of the model was assessed using the same metrics as the NKLR study: calibration and concordance at each follow-up time. Calibration refers to the accuracy of the risk estimates and was calculated using a version of the Hosmer–Lemeshow statistic appropriate for censored data [30]. This statistic sums average misclassification in each predicted risk quantile and converts the result into a chi-squared statistic. A larger calibration statistic indicates worse calibration, and statistical significance means the null hypothesis of perfect calibration is rejected. Concordance was computed using Harrell’s C-index [12] at 1, 2, and 5-year follow-up times. The C-index is a generalization of area under the curve (AUC) for censored data that measures the proportion of ranked pairs of observations in which the predicted ranking corresponds with true outcomes. As with AUC, the C-index ranges from 0 to 1 with 1 indicating perfect concordance.

## Results

### Participants

Table 1 describes characteristics of the DKLR population at the time of primary surgery. Patients had an average age at primary surgery of 29 years (SD ± 10) and 60% were male.

Hamstring graft was used in 82% of primary surgeries. Of the DKLR patients, 10,922 had complete data for all five variables required by the NKLR Cox Lasso model. Table 2 compares DKLR patients with complete data for these five variables to the NKLR training-data patients with complete data. The large sample sizes produced p-values below the significance threshold on all characteristics, including a few clinically meaningful differences. The DKLR patients were more likely to have hamstring tendon autograft (DKLR: 81%; NKLR: 59%) and suspension/cortical femur fixation (DKLR: 72%; NKLR: 53%). Additionally, the rate of concomitant meniscus (DKLR: 42%; NKLR: 53%) and chondral (DKLR: 14%; NKLR: 23%) injuries were higher in the NKLR cohort, while overall revision rate was higher in the Danish registry patients (DKLR: 6.9%; NKLR: 5.2%). The DKLR patients with complete data on the five required variables were in general similar to those without complete data, particularly on the five required variables (Supplementary Table 1).

### Model performance

The NKLR Cox Lasso model produced similar concordance with the DKLR population compared to the original NKLR test data (DKLR: 0.68; NKLR: 0.68–0.69). Calibration was poorer for the DKLR population than for the NKLR test data at 1 and 5 years post primary surgery but similar at two years (Table 3).

## Discussion

The most important finding of this study was that a machine learning algorithm developed from the NKLR demonstrated similar performance when applied to patients from the DKLR. Despite different injury profiles including concomitant meniscus/chondral injury rates and variation in surgical technique trends between the two nations, the concordance was nearly identical to that achieved with the index study of NKLR patients. This suggests that the algorithm is valid for application outside of the initial patient population and represents the first machine learning model for predicting revision ACL reconstruction that has been externally validated. The original model was developed to help guide the clinical discussion regarding surgical options and outcome expectations at a patient-specific level [19].

Machine learning models explore large datasets divided into inputs (predictors) and outputs (outcomes), to establish connections and relationships between them. These relationships may be more complex than could be identified through standard statistical analysis. When a machine learning algorithm can determine a link between the predictors and outcome of interest, it can then create a tool capable of



**Table 2** Characteristics of patients with complete data on Norwegian Cox lasso variables

Variable*	Danish N = 10,922	Norwegian N = 14,161	P value**
Years: surgery to data current date (Danish: 06–14-2021; Norwegian: 01–12-2020)	9.3 (4.1)	8.4 (4.1)	< 0.001
Revision	755 (6.9%)	743 (5.2%)	< 0.001
Follow-up time or time to revision	8.4 (4.3)	7.0 (4.2)	< 0.001
Age at surgery	29 (11)	28 (10)	< 0.001
Age at injury	27 (10)	26 (10)	< 0.001
Missing	9	0	
Sex			n.s
Female	4916 (45%)	6376 (45%)	
Male	6006 (55%)	7785 (55%)	
Pre-surgery KOOS QOL score (out of 10)	3.90 (1.61)	3.48 (1.87)	< 0.001
Pre-surgery KOOS Sports score (out of 10)	3.80 (2.55)	4.27 (2.73)	< 0.001
Missing	1	137	
Below median on all pre-surgery KOOS	1825 (17%)	2799 (20%)	< 0.001
Meniscus injury	4584 (42%)	7537 (53%)	< 0.001
Cartilage injury	1579 (14%)	3318 (23%)	< 0.001
Graft choice			< 0.001
BPTB	1133 (10%)	5522 (39%)	
Hamstring	8866 (81%)	8369 (59%)	
Unknown/Other	923 (8.5%)	270 (1.9%)	
Tibia fixation device			< 0.001
Interference screw	9925 (91%)	10,841 (77%)	
Suspension/cortical device	155 (1.4%)	1468 (10%)	
Unknown/Other	842 (7.7%)	1852 (13%)	
Femur fixation device			< 0.001
Interference screw	2025 (19%)	4763 (34%)	
Suspension/cortical device	7891 (72%)	7522 (53%)	
Unknown/Other	1006 (9.2%)	1876 (13%)	
Fixation device combination			< 0.001
Interference screw × 2	1978 (18%)	4645 (33%)	
Interference/Suspension	2 (< 0.1%)	90 (0.6%)	
Suspension/cortical device × 2	153 (1.4%)	1095 (7.7%)	
Suspension/Interference	7218 (66%)	5529 (39%)	
Unknown/Other	1571 (14%)	2802 (20%)	
Injured side			n.s
Right	5512 (50%)	7149 (50%)	
Left	5409 (50%)	7012 (50%)	
Missing	1	0	
Previous surgery on opposite knee	549 (5.0%)	1001 (7.1%)	< 0.001
Missing	27	0	
Previous surgery on same knee	9014 (83%)	2412 (17%)	< 0.001
Time injury to surgery (years)	1.75 (3.34)	1.66 (3.35)	0.040
Systemic antibiotic prophylaxis			< 0.001
Yes	10,922 (100%)	14,089 (99%)	
No	0 (0%)	46 (0.3%)	
Missing	0 (0%)	26 (0.2%)	

\*Statistics presented: Mean (SD); n (%)

\*\*Statistical tests: Welch Two Sample *t* test; Pearson's Chi-squared test

**Table 3** Model performance

Probability of Revision	Model	Concordance	Calibration statistic	Calibration p-value
1 year	Original Norwegian Algorithm	0.686	4.89	n.s
	Danish Knee Ligament Registry	0.678	22.24	<0.001
2 years	Original Norwegian Algorithm	0.684	11.35	0.01
	Danish Knee Ligament Registry	0.676	11.82	0.008
5 years	Original Norwegian Algorithm	0.683	6.19	n.s
	Danish Knee Ligament Registry	0.678	13.98	0.003

predicting this outcome for other patients. After a prediction model has been developed, the TRIPOD Statement strongly recommends external validation, achieved through evaluation of model performance on new and different groups of patients than were used in the development of the algorithm [6]. However, this important step is uncommonly performed, inhibiting the clinical translation of newly developed models [28].

The original machine learning model was created based on a database including nearly 25,000 patients with 24 variables considered. Four machine learning models were evaluated, and the Cox Lasso model was selected for the development of an in-clinic prediction tool. This tool required the input of only five variables for the prediction of subsequent revision ACL reconstruction risk. Although the performance of this model was assessed using hold-out data that was not included in the learning phase, it only included patients from one nation, limiting its applicability to patients from other countries [19].

This study found that accuracy of the NKLR Cox Lasso model holds when applied to a large data set from another country with different injury characteristics and surgical technique trends. The prediction model demonstrated similar model performance when tested on patients from Denmark that had not been previously seen by the algorithm. It was initially developed using 75% of the patients in the NKLR and validated using the remaining 25%. This study validates the algorithm using an additional 11,000 patients from the DKLR and represents a necessary step toward clinical utility. While this is encouraging, it should be noted that the performance of the model on patients undergoing ACL reconstruction outside of Scandinavia remains unknown. Additionally, there are currently no other published prediction models with which to compare the performance of this model.

Study population variance between the DKLR and NKLR populations may help explain differences in model calibration at one and five years post primary surgery. The DKLR patients with complete data had higher proportions of hamstring tendon autograft and suspension/cortical femur fixation than patients in the NKLR test data. Both these variables are used in the NKLR Cox Lasso model. Thus, the relationship between graft choice and/or femur fixation and

revision risk codified in the model may not be as accurate for patient populations with a substantially different distribution on these variables, such as those in the DKLR. Regarding the fact that the validation data set was limited to approximately one-third of the overall DKLR registry population due to missing values for the required predictors, the objective of this paper was to test the machine learning model on a new population and the inclusion of nearly 11,000 patients represents a suitable data set for this purpose.

While this novel technique represents a new frontier for health-related research, limitations regarding the clinical utility of machine learning algorithms remain. Most importantly, the quality of the model is largely related to the quality of the data that it is developed from. The concordance of the revision ACL prediction tool is moderate based on both the initial and subsequent validation studies. As noted in the original paper, this may be related to data quality since several risk factors for failure of ACL reconstruction are not captured in the NKLR [19]. Examples of these factors include radiographic variables such as tibial slope and coronal alignment [2–4, 10, 15, 20, 31], physical examination and rehabilitation details [11, 14, 18, 22], and surgical technique factors such as tunnel position [16] and graft size [1, 7, 17]. The addition of these variables into the national knee ligament registers may improve future machine learning prediction endeavours.

There is an additional limitation concerning this external validation study. Since pre-operative KOOS QoL score at the time of surgery was one of the input variables required for outcome prediction, all patients in the DKLR without a pre-operative KOOS score were excluded from the analysis. This resulted in the exclusion of approximately two-thirds of the patients contained in the DKLR since pre-surgical compliance with patient reported outcome measures is relatively low in the registry. Despite this, nearly 11,000 patients were still included in the model evaluation which is sufficient for validation.

Machine learning analysis of large health-care registries have the potential for great impact on patient care. These advanced statistical techniques can assess and interpret large volumes of data and recognize complex associations between predictor variables and patient-specific outcome. The resulting algorithm, as is the case with the present study, can be implemented into clinical care as an adjunct for the

orthopaedic surgeon. Supplementing their personal experience and interpretation of the relevant risk factors, clinicians can use this in-clinic calculator to individualize their discussions and quantify the risk of revision ACL reconstruction for their patients.

## Conclusion

The NKLR machine learning algorithm demonstrated similar performance when applied to patients from the DKLR, suggesting that it is valid for application outside of the initial patient population. This represents the first machine learning model for predicting revision ACL reconstruction that has been externally validated. Clinicians can use this in-clinic calculator to estimate revision risk at a patient specific level when discussing outcome expectations pre-operatively. While encouraging, it should be noted that the performance of the model on patients undergoing ACL reconstruction outside of Scandinavia remains unknown.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00167-021-06828-w>.

**Funding** This study was funded by a Norwegian Centennial Chair seed grant. Funding supported the machine learning analysis and interpretation. The funding agencies had no direct role in the investigation.

## Declarations

**Conflict of interest** The authors declare that they have no competing interest.

**Institutional review board** Approval not required as consent was obtained by all patients at time of enrollment in the national knee ligament register.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alkhalaf FNA, Hanna S, Alkhalidi MSH, Alenezi F, Khaja A (2021) Autograft diameter in ACL reconstruction: size does matter. SICOT-J 7:16
- Bayer S, Meredith SJ, Wilson KW, de Sa D, Paulyo T, Byrne K, McDonough CM, Musahl V (2020) Knee morphological risk factors for anterior cruciate ligament injury: a systematic review. J Bone Jt Surg 102:703–718
- Bernhardson AS, Aman ZS, Dornan GJ, Kemler BR, Storaci HW, Brady AW, Nakama GY, LaPrade RF (2019) Tibial slope and its effect on force in anterior cruciate ligament grafts: anterior cruciate ligament force increases linearly as posterior tibial slope increases. Am J Sports Med 47:296–302
- Bernholt DL, Dornan GJ, DePhillipo NN, Aman ZS, Kennedy MI, LaPrade RF (2020) High-grade posterolateral tibial plateau impaction fractures in the setting of a primary anterior cruciate ligament tear are correlated with an increased preoperative pivot shift and inferior postoperative outcomes after anterior cruciate ligament reconstruction. Am J Sports Med 48:2185–2194
- Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, Cheon J-E, Ha JY (2020) Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. Invest Radiol 55:101–110
- Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 162:55–63
- Conte EJ, Hyatt AE, Gatt CJ, Dhawan A (2014) Hamstring autograft size can be predicted and is a potential risk factor for anterior cruciate ligament reconstruction failure. Arthroscopy 30:882–890
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118
- Granan L-P, Bahr R, Steindal K, Furnes O, Engebretsen L (2008) Development of a national cruciate ligament surgery registry: the Norwegian National Knee Ligament Registry. Am J Sports Med 36:308–315
- Grassi A, Signorelli C, Urrizola F, Macchiarola L, Raggi F, Mosca M, Samuelsson K, Zaffagnini S (2019) Patients with failed anterior cruciate ligament reconstruction have an increased posterior lateral tibial plateau slope: a case–controlled study. Arthroscopy 35:1172–1182
- Grindem H, Snyder-Mackler L, Moksnes H, Engebretsen L, Risberg MA (2016) Simple decision rules can reduce reinjury risk by 84% after ACL reconstruction: the Delaware-Oslo ACL cohort study. Br J Sports Med 50:804–808
- Harrell FE (1982) Evaluating the yield of medical tests. JAMA 247:2543
- Kaeding CC, Pedroza AD, Reinke EK, Huston LJ, Spindler KP (2015) 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 43:1583–1590
- Krebs NM, Barber-Westin S, Noyes FR (2021) Generalized joint laxity is associated with increased failure rates of primary anterior cruciate ligament reconstructions: a systematic review. Arthroscopy 37:2337–2347
- Li Y, Hong L, Feng H, Wang Q, Zhang J, Song G, Chen X, Zhuo H (2014) Posterior tibial slope influences static anterior tibial translation in anterior cruciate ligament reconstruction: a minimum 2-year follow-up study. Am J Sports Med 42:927–933
- Liu A, Sun M, Ma C, Chen Y, Xue X, Guo P, Shi Z, Yan S (2017) Clinical outcomes of transtibial versus anteromedial drilling techniques to prepare the femoral tunnel during anterior cruciate ligament reconstruction. Knee Surg Sports Traumatol Arthrosc 25:2751–2759
- Magnussen RA, Lawrence JTR, West RL, Toth AP, Taylor DC, Garrett WE (2012) Graft size and patient age are predictors of

- early revision after anterior cruciate ligament reconstruction with hamstring autograft. *Arthroscopy* 28:526–531
18. Marmura H, Getgood AMJ, Spindler KP, Kattan MW, Briskin I, Bryant DM (2021) Validation of a risk calculator to personalize graft choice and reduce rupture rates for anterior cruciate ligament reconstruction. *Am J Sports Med* 49:1777–1785
  19. Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Engebretsen L (2021) Predicting anterior cruciate ligament reconstruction revision: a machine learning analysis utilizing the Norwegian Knee Ligament Register. *J Bone Jt Surg*. <https://doi.org/10.2106/JBJS.21.00113>
  20. Mehl J, Otto A, Kia C, Murphy M, Obopilwe E, Imhoff FB, Feucht MJ, Imhoff AB, Arciero RA, Beitzel K (2020) Osseous valgus alignment and posteromedial ligament complex deficiency lead to increased ACL graft forces. *Knee Surg Sports Traumatol Arthrosc* 28:1119–1129
  21. Persson A, Fjeldsgaard K, Gjertsen J-E, Kjellsen AB, Engebretsen L, Hole RM, Fevang JM (2014) Increased risk of revision with hamstring tendon grafts compared with patellar tendon grafts after anterior cruciate ligament reconstruction: a study of 12,643 patients from the Norwegian Cruciate Ligament Registry, 2004–2012. *Am J Sports Med* 42:285–291
  22. Roe C, Jacobs C, Kline P, Lucas K, Johnson D, Ireland ML, Lattermann C, Noehren B (2021) Correlations of single-leg performance tests to patient-reported outcomes after primary anterior cruciate ligament reconstruction. *Clin J Sport Med* 31:e265–e270
  23. Rouzrokh P, Wyles CC, Philbrick KA, Ramazanian T, Weston AD, Cai JC, Taunton MJ, Lewallen DG, Berry DJ, Erickson BJ, Maradit Kremers H (2021) A deep learning tool for automated radiographic measurement of acetabular component inclination and version after total hip arthroplasty. *J Arthroplasty* 36:2510–2517
  24. Sanders TL, Pareek A, Hewett TE, Levy BA, Dahm DL, Stuart MJ, Krych AJ (2017) Long-term rate of graft failure after ACL reconstruction: a geographic population cohort analysis. *Knee Surg Sports Traumatol Arthrosc* 25:222–228
  25. Schock J, Truhn D, Abrar DB, Merhof D, Conrad S, Post M, Mittelstrass F, Kuhl C, Nebelung S (2021) Automated analysis of alignment in long-leg radiographs by using a fully automated support system based on artificial intelligence. *Radiol Artif Intell* 3:e200198
  26. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Židek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D (2020) Improved protein structure prediction using potentials from deep learning. *Nature* 577:706–710
  27. Snaebjörnsson T, Svantesson E, Sundemo D, Westin O, Sansone M, Engebretsen L, Hamrin-Senorski E (2019) Young age and high BMI are predictors of early revision surgery after primary anterior cruciate ligament reconstruction: a cohort study from the Swedish and Norwegian knee ligament registries based on 30,747 patients. *Knee Surg Sports Traumatol Arthrosc* 27:3583–3591
  28. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG (2013) Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 10:e1001381
  29. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N (2019) Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol* 48:239–244
  30. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Johnson PE, Vazquez-Benitez G, O'Connor PJ (2016) 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* 61:119–131
  31. Webb JM, Salmon LJ, Leclerc E, Pinczewski LA, Roe JP (2013) Posterior tibial slope and further anterior cruciate ligament injuries in the anterior cruciate ligament-reconstructed patient. *Am J Sports Med* 41:2800–2804
  32. Webster KE, Feller JA, Leigh WB, Richmond AK (2014) Younger patients are at increased risk for graft rupture and contralateral injury after anterior cruciate ligament reconstruction. *Am J Sports Med* 42:641–647
  33. Wiggins AJ, Grandhi RK, Schneider DK, Stanfield D, Webster KE, Myer GD (2016) Risk of secondary injury in younger athletes after anterior cruciate ligament reconstruction: a systematic review and meta-analysis. *Am J Sports Med* 44:1861–1876
  34. Yamada Y, Maki S, Kishida S, Nagai H, Arima J, Yamakawa N, Iijima Y, Shiko Y, Kawasaki Y, Kotani T, Shiga Y, Inage K, Orita S, Eguchi Y, Takahashi H, Yamashita T, Minami S, Ohtori S (2020) 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* 91:699–704

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

## Authors and Affiliations

R. Kyle Martin<sup>1,2</sup> · Solvejg Wastvedt<sup>3</sup> · Ayoosh Pareek<sup>4</sup> · Andreas Persson<sup>5,6,7</sup> · Håvard Visnes<sup>5</sup> · Anne Marie Fenstad<sup>5</sup> · Gilbert Moatshe<sup>6,7</sup> · Julian Wolfson<sup>3</sup> · Martin Lind<sup>8</sup> · Lars Engebretsen<sup>6,7</sup>

<sup>1</sup> Department of Orthopedic Surgery, University of Minnesota, 2512 South 7th Street, Suite R200, Minneapolis, MN 55455, USA

<sup>2</sup> Department of Orthopaedic Surgery, CentraCare, Saint Cloud, MN, USA

<sup>3</sup> Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA

<sup>4</sup> Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN, USA

<sup>5</sup> Norwegian Knee Ligament Register, Haukeland University Hospital, Bergen, Norway

<sup>6</sup> Oslo Sport Trauma Research Center, Norwegian School of Sports Science, Oslo, Norway

<sup>7</sup> Orthopaedic Clinic, Oslo University Hospital Ullevål, Oslo, Norway

<sup>8</sup> Aarhus University Hospital, Aarhus, Denmark

**Supplementary Table 1: Comparison of Danish registry patients with complete vs. incomplete data on Norwegian Cox lasso variables**  
(Yellow highlights show Norwegian cox lasso variables)

Variable*	Complete N = 10,922	Incomplete N = 23,756	P-value**
Years: surgery to data current date (06-14-2021)	9.3 (4.1)	7.9 (4.3)	<0.001
Missing	0	1	
Revision	755 (6.9%)	1,036 (4.4%)	<0.001
Missing	0	1	
Follow-up time or time to revision	8.4 (4.3)	7.2 (4.3)	<0.001
Missing	0	1	
Age at surgery	29 (11)	28 (10)	0.006
Missing	0	1	
Age at injury	27 (10)	27 (10)	n.s.
Missing	9	490	
Sex			<0.001
Female	4,916 (45%)	9,042 (38%)	
Male	6,006 (55%)	14,713 (62%)	
Missing	0	1	
Pre-surgery KOOS QOL score (out of 10)	3.90 (1.61)	3.74 (1.58)	n.s.
Missing	0	23,522	
Pre-surgery KOOS Sports score (out of 10)	3.80 (2.55)	3.76 (2.57)	n.s.
Missing	1	23,522	
Below median on all pre-surgery KOOS	1,825 (17%)	43 (18%)	n.s.
Missing	0	23,520	
Meniscus injury	4,584 (42%)	10,917 (46%)	<0.001
Cartilage injury	1,579 (14%)	3,766 (16%)	<0.001
Graft choice			<0.001
BPTB	1,133 (10%)	2,085 (8.8%)	
Hamstring	8,866 (81%)	19,425 (82%)	
Unknown/Other	923 (8.5%)	2,122 (9.0%)	
Missing	0	124	
Tibia fixation device			<0.001
Interference screw	9,925 (91%)	20,892 (88%)	
Suspension/cortical device	155 (1.4%)	828 (3.5%)	
Unknown/Other	842 (7.7%)	2,036 (8.6%)	
Femur fixation device			<0.001
Interference screw	2,025 (19%)	4,047 (17%)	
Suspension/cortical device	7,891 (72%)	17,058 (72%)	
Unknown/Other	1,006 (9.2%)	2,651 (11%)	
Fixation device combination			<0.001
Interference screw x2	1,978 (18%)	3,973 (17%)	
Interference/Suspension	2 (<0.1%)	8 (<0.1%)	
Suspension/cortical device x2	153 (1.4%)	815 (3.4%)	

Suspension/Interference	7,218 (66%)	15,090 (64%)	0.043
Unknown/Other	1,571 (14%)	3,870 (16%)	
Injured side			0.043
Right	5,512 (50%)	12,269 (52%)	
Left	5,409 (50%)	11,486 (48%)	
Missing	1	1	
Previous surgery on opposite knee	549 (5.0%)	2,196 (9.3%)	<0.001
Missing	27	81	
Previous surgery on same knee	9,014 (83%)	19,795 (83%)	n.s.
Time injury to surgery (years)	1.75 (3.34)	1.60 (3.14)	<0.001
Missing	0	712	
Systemic Antibiotic Prophylaxis	10,922 (100%)	23,756 (100%)	

\*Statistics presented: Mean (SD); n (%)

\*\*Statistical tests: Welch Two Sample t-test; Pearson's Chi-squared test; Fisher's exact test for fixation device combination variable

## Paper V

Martin RK, Marmura H, Wastvedt S, Pareek A, Persson A, Moatshe G, Bryant D, Wolfson J, Engebretsen L, Getgood A. External validation of the Norwegian anterior cruciate ligament reconstruction revision prediction model using patients from the STABILITY 1 Trial. *Knee Surg Sports Traumatol Arthrosc.* 2024;32(2):206-213. doi:10.1002/ksa.12031





# External validation of the Norwegian anterior cruciate ligament reconstruction revision prediction model using patients from the STABILITY 1 Trial

R. Kyle Martin<sup>1,2,3</sup> | Hana Marmura<sup>4</sup> | Solvejg Wastvedt<sup>5</sup> | Ayoosh Pareek<sup>6</sup> |  
 Andreas Persson<sup>3,7</sup> | Gilbert Moatshe<sup>3,7</sup> | Dianne Bryant<sup>8</sup> | Julian Wolfson<sup>5</sup> |  
 Lars Engebretsen<sup>3,7</sup> | Alan Getgood<sup>4</sup>

<sup>1</sup>Department of Orthopaedic Surgery, University of Minnesota, Minneapolis, Minnesota, USA

<sup>2</sup>Department of Orthopaedic Surgery, CentraCare, Saint Cloud, Minnesota, USA

<sup>3</sup>Oslo Sport Trauma Research Center, Norwegian School of Sports Science, Oslo, Norway

<sup>4</sup>Department of Orthopaedic Surgery, University of Western Ontario, London, Ontario, Canada

<sup>5</sup>Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, USA

<sup>6</sup>Department of Orthopaedic Surgery, Hospital for Special Surgery, New York, New York, USA

<sup>7</sup>Orthopaedic Clinic, Oslo University Hospital Ullevål, Oslo, Norway

<sup>8</sup>School of Physical Therapy, University of Western Ontario, London, Ontario, Canada

## Correspondence

R. Kyle Martin, Department of Orthopaedic Surgery, University of Minnesota, 2512 South 7th Street, Suite R200, Minneapolis, MN 55455, USA.  
 Email: rkylemartin@gmail.com

## Funding information

Norwegian Centennial Chair Seed Grant

## Abstract

**Purpose:** A machine learning-based anterior cruciate ligament (ACL) revision prediction model has been developed using Norwegian Knee Ligament Register (NKLK) data, but lacks external validation outside Scandinavia. This study aimed to assess the external validity of the NKLK model ([https://swastvedt.shinyapps.io/calculator\\_rev/](https://swastvedt.shinyapps.io/calculator_rev/)) using the STABILITY 1 randomized clinical trial (RCT) data set. The hypothesis was that model performance would be similar.

**Methods:** The NKLK Cox Lasso model was selected for external validation owing to its superior performance in the original study. STABILITY 1 patients with all five predictors required by the Cox Lasso model were included. The STABILITY 1 RCT was a prospective study which randomized patients to receive either a hamstring tendon autograft (HT) alone or HT plus a lateral extra-articular tenodesis (LET). Since all patients in the STABILITY 1 trial received HT ± LET, three configurations were tested: 1: all patients coded as HT, 2: HT + LET group coded as bone-patellar tendon-bone (BPTB) autograft, 3: HT + LET group coded as unknown/other graft choice. Model performance was assessed via concordance and calibration.

**Results:** In total, 591/618 (95.6%) STABILITY 1 patients were eligible for inclusion, with 39 undergoing revisions within 2 years (6.6%). Model performance was best when patients receiving HT + LET were coded as BPTB. Concordance was similar to the original NKLK prediction model for 1- and 2-year revision prediction (STABILITY: 0.71; NKLK: 0.68–0.69). Concordance 95% confidence interval (CI) ranged from 0.63 to 0.79. The model was well calibrated for 1-year prediction while the 2-year prediction demonstrated evidence of miscalibration.

**Conclusion:** When patients in STABILITY 1 who received HT + LET were coded as BPTB in the NKLK prediction model, concordance was similar to

**Abbreviations:** ACL, anterior cruciate ligament; AUC, area under the curve; BPTB, bone-patellar tendon-bone; CI, confidence interval; HT, Hamstring tendon; IRB, Institutional Review Board; KOOS-QOL, Knee Injury and Osteoarthritis Outcome Score Quality of Life subscale; LET, lateral extra-articular tenodesis; NKLK, Norwegian Knee Ligament Register; RCT, randomized clinical trial; REK, Regional Ethics Committee; STABILITY 1, randomized clinical trial investigating outcomes of hamstring tendon autograft (HT) ACLR with or without lateral extra-articular tenodesis (LET).

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Knee Surgery, Sports Traumatology, Arthroscopy* published by John Wiley & Sons Ltd on behalf of European Society of Sports Traumatology, Knee Surgery and Arthroscopy.

the index study. However, due to a wide 95% CI, the true performance of the prediction model with this Canadian and European cohort is unclear and a larger data set is required to definitively determine the external validity. Further, better calibration for 1-year predictions aligns with general prediction modelling challenges over longer periods. While not a large enough sample size to elicit the true accuracy and external validity of the prediction model when applied to North American patients, this analysis provides more support for the notion that HT plus LET performs similarly to BPTB reconstruction. In addition, despite the wide confidence interval, this study suggests optimism regarding the accuracy of the model when applied outside of Scandinavia.

**Level of Evidence:** Level 3, cohort study.

#### KEYWORDS

ACL, external validation, machine learning, outcome prediction

## INTRODUCTION

Anterior cruciate ligament (ACL) reconstruction (ACLR) is a commonly performed procedure aimed at reducing instability and restoring normal knee biomechanics after injury. Unfortunately, graft rupture and subsequent revision surgery remains an issue of concern—especially among young, active patients [1, 2]. By now, several risk factors for ACLR failure have been identified and include both modifiable and non-modifiable traits [3–6]. Recognition of these factors enables the clinician to coarsely risk-stratify patients who can influence surgical decision-making and outcome expectations. However, due to the sheer number of potential risk factors and the complex interactions between them, fine-level risk estimation remains challenging.

The emergence of machine learning applications into the orthopaedic literature has been heralded as a potential adjunctive tool capable of improving outcome prediction accuracy [7, 8]. These advanced statistical techniques can identify and interpret complex and non-linear interactions between variables leading to a more accurate understanding of how risk factors may affect surgical outcomes, both together and in isolation. This opens the door to the possibility of patient-specific risk estimation, surgical discussion and outcome optimization.

Preliminary machine learning-derived models for the prediction of ACLR outcome, including revision surgery, have recently been developed based on patients in the Norwegian Knee Ligament Register (NKLR) [9, 10]. The revision surgery prediction model includes an open-access online clinical calculator ([https://swastvedt.shinyapps.io/calculator\\_rev/](https://swastvedt.shinyapps.io/calculator_rev/)) and has undergone further external validation using patients from the Danish Knee Ligament Registry [11]. In general, external validation of clinical machine learning models in orthopaedic surgery is uncommonly performed and represents a crucial step prior to widespread adoption and implementation

in clinical practice. External validation is valuable for several reasons, including assessment of model generalizability, minimizing bias and model overfitting, and increasing trust and acceptance among patients and clinicians regarding the utility of the model.

The purpose of this study was to determine the external validity of the previously published ACL revision prediction model when applied to patients enrolled in the STABILITY 1 randomized clinical trial investigating outcomes of hamstring tendon autograft (HT) ACLR with or without lateral extra-articular tenodesis (LET) [12]. The hypothesis was that model performance would be similar to the index study, suggesting validity of the algorithm. This study represents the first attempt to assess external validation using patients from outside of Scandinavia. If successful, this algorithm may be used to estimate revision risk at a patient-specific level and help guide discussion surrounding outcome expectations preoperatively.

## MATERIALS AND METHODS

Development of the Norwegian ACL revision surgery prediction algorithm was based on data from the NKLR. At the time of enrolment in this national registry, all patients provide informed consent and the Norwegian Data Inspectorate granted permission to the register for collection, analysis and publication on this health-related data. All data were de-identified prior to retrieval for model development and no further ethical approval is required from the Regional Ethics Committee (REK) for NKLR-based studies [13]. The original prediction model development was performed at the University of Minnesota and the respective Institutional Review Board similarly concluded that the study was exempt from full review (#00012552). The STABILITY 1 trial was approved by the Western Ontario Health Sciences Research Ethics Board (#104524).

## Index model development

The original prediction model was developed through machine learning analysis of all patients contained within the NKLR who underwent primary ACLR [9]. This national knee ligament registry prospectively enrolls patients undergoing cruciate ligament reconstruction preoperatively and records demographic, injury, surgical and follow-up outcome details including subsequent revision surgery. The NKLR was established in 2004 and reporting has been mandatory since 2017. Overall compliance with the NKLR approximates 88% [14]. Patients are registered using their unique Norwegian national identification number which links identification of subsequent revision surgery performed within Norway, regardless of the provider.

In the index study of NKLR patients [9], four machine learning prediction models were assessed for the ability to predict subsequent revision ACLR after primary surgery and the primary outcome was probability of revision ACLR within 1, 2 and/or 5 years. The four models tested were Cox Lasso, survival random forest, generalized additive model and gradient boosted regression. These four models are among the most commonly used for this type of analysis. The patients in the NKLR were randomly split into training (75%) and test (25%) sets, whereby the algorithm was developed using the training set of patients, and the performance of the algorithm was assessed with the hold-out test set, previously unseen by the models. The Cox Lasso model was the best-performing of the four tested models and was used for the development of an in-clinic revision-risk calculator.

Regarding outcome prediction, the four models considered all the available data in the NKLR to 'learn' which factors are associated with—and can be used to predict—which patients will eventually undergo revision surgery. Starting with the 24 total predictor variables in the NKLR, the models eliminated variables which do not significantly contribute to prediction ability, without sacrificing accuracy. The result was an algorithm developed using the Cox Lasso model that only required five variables (out of the 24) for outcome prediction. The model was generally well calibrated and demonstrated moderate discriminative ability in predicting revision surgery after primary ACLR [9].

## Data source

The validation data for this study were extracted from the STABILITY 1 study, a randomized clinical trial conducted across nine sites (seven in Canada and two in Europe) [12]. This study investigated the 2-year outcomes of patients under 25 years of age who were undergoing a primary ACLR. Patients were randomized to undergo a HT ACLR either with or without an LET.

The patients in this trial were classified as being at high risk of re-injury and/or surgical failure based on meeting at least two of the following criteria: a pivot shift Grade 2 or higher, a desire to return to high-risk/pivoting sports, and/or generalized ligamentous laxity. Outcome data for these patients were obtained at 3, 6, 12, and 24 months postoperatively.

## Participants and predictors

This current study sought to validate the previously developed Cox Lasso model from the NKLR. The Cox Lasso model was selected for validation since it was the best performing model. The Cox Lasso is a penalized regression model, which selects a subset of available variables for inclusion [15]. Thus, while a more extensive set of patient characteristics were assessed in development of the model, only the five predictors required for the NKLR Cox Lasso model were used in this validation analysis. The five variables required for outcome prediction using the Cox Lasso model were: patient age at primary surgery, Knee Injury and Osteoarthritis Outcome Score Quality of Life subscale (KOOS-QOL) score at primary surgery, graft choice, femur fixation method, and time between injury and ACLR.

The graft choice options in the NKLR model are HT, bone-patellar tendon-bone (BPTB), or other/unknown. STABILITY 1 participants were randomized to have an HT ± LET; therefore, all patients had an HT for the graft choice. The two STABILITY 1 graft type groups were coded in three different ways (1: all patients coded as HT, 2: HT + LET group coded as BPTB, 3: HT + LET group coded as unknown/other) to understand which would be most appropriate and which group the HT + LET group behaved most similar to in the model. This approach was chosen based on a previous study that demonstrated the addition of LET to an HT behaved similarly to a BPTB [16]. Since the STABILITY 1 trial followed patients for 2 years post-operatively, 5-year data and predictions were not included. Two patients without a documented revision date were included, with their graft rupture date substituted for the revision date (both 21 months postoperative). Patient characteristics for the STABILITY 1 validation data set are shown in Table 1 [12].

## Model performance

Performance of the model was assessed using the same metrics as the NKLR study: calibration and concordance (discrimination) at each follow-up time. Performance evaluation included censoring of the time-to-event outcome. 'Censoring' refers to the fact that, at any given follow-up time, complete information

**TABLE 1** Characteristics of patients in validation data set.

Characteristics	All patients, <i>N</i> = 618 (mean ± SD) or <i>n</i> (%)	Patients with complete model data, <i>N</i> = 591 (mean ± SD) or <i>n</i> (%)
Age, years	18.9 ± 3.2	19.0 ± 3.2
Missing	1 (0.2)	0 (0)
Sex		
Male	298 (48.2)	287 (48.6)
Female	319 (51.6)	304 (51.4)
Missing	1 (0.2)	0 (0)
BMI, kg/m <sup>2</sup>	24.1 ± 3.8	24.1 ± 3.8
Missing	8 (1.3)	2 (0.3)
KOOS-QOL, 0–100	33.2 ± 17.8	33.2 ± 17.8
Missing	16 (2.6)	0 (0)
Graft		
HT	311 (50.3)	296 (50.1)
BPTB	0 (0)	0 (0)
Other/Unknown	0 (0)	0 (0)
HT + LET	307 (49.7)	295 (49.9)
Femur fixation method		
Interference screw	0 (0)	0 (0)
Suspension or cortical device	618 (100)	591 (100)
Unknown or other	0 (0)	0 (0)
Years between injury & surgery	0.72 ± 1.49	0.72 ± 1.50
Missing	19 (3.1)	0 (0)
Revision		
Yes	40 (6.5)	39 (6.6)
Within 1 year	9 (1.5)	8 (1.4)
Between 1 and 2 years	22 (3.6)	22 (3.7)
After 2 years	9 (1.5)	9 (1.5)
No	570 (92.2)	552 (93.4)
Missing	9 (1.5)	0 (0)

Abbreviations: BMI, body mass index; BPTB, bone-patellar tendon-bone; HT, hamstring tendon; LET, lateral extra-articular tenodesis; KOOS-QOL, knee osteoarthritis outcome score quality of life scale.

on outcome is not known for all patients. Some patients have not been followed in the study for the requisite number of years, while others have not yet experienced revision and it is unknown when or if they ultimately will.

### Statistical analysis

The program R (RStudio 2022.07.1) was used to calculate predicted survival probabilities for all patients in the validation data set. Calibration refers to the

accuracy of the risk estimates. In the NKLR study, calibration was calculated using a version of the Hosmer–Lemeshow statistic [17]. This statistic sums average misclassification in each predicted risk quantile and converts the result into a chi-squared statistic. However, for the validation analysis, the low number of revisions in the validation data necessitated a slightly different approach. Rather than divide the data into risk quantiles, it was divided into three groups as follows: 0–25th percentile, 26–50th percentile, and 51–100th percentile of predicted survival probability. This change

ensured an adequate number of revisions in each group while retaining the statistical validity of the original method. For both calibration methods, a larger calibration statistic indicates worse calibration, and statistical significance means the null hypothesis of perfect calibration is rejected. Concordance was computed using Harrell's C-index [18] at 1- and 2-year follow-up times. The C-index is a generalization of area under the curve (AUC) that measures the proportion of ranked pairs of observations in which the predicted ranking corresponds with true outcomes. As with AUC, the C-index ranges from 0 to 1 with 1 indicating perfect concordance.

## RESULTS

### Participants

Of the 618 participants randomized in the STABILITY 1 study, 591 (95.6%) had complete data on Norwegian Cox lasso variables (five predictor variables and outcome). Of note, there were only 39 (6.6%) revision events in the analysed data set, 30 of which occurred by 2-year follow-up (5.1% of analysed sample). Eight patients had revision surgery within 1-year of their primary surgery while another 22 patients had surgery between the first- and second-year time points. An additional nine patients underwent revision after the 2-year follow-up timepoint.

In contrast to the original NKLR study cohort, which included all patients undergoing ACLR in Norway, the STABILITY 1 trial patients had a narrow age range (14–25 years old) and all patients received HT with suspensory fixation on the femur. Further, time from injury to surgery was shorter and more consistent for the STABILITY 1 patients relative to the NKLR data set.

### Model performance

Model performance was best for both 1- and 2-year revision prediction when patients randomized to HT + LET in the STABILITY 1 trial were coded as BPTB in the prediction calculator (Table 2). The model concordance (discriminative ability) was similar in the validation data set (0.71 and 0.71) compared to the development data set (0.69 and 0.68) at 1- and 2-year follow-up, respectively, with 95% confidence intervals (CI) ranging from 0.63 to 0.79. The concordance values were slightly better in the STABILITY 1 data set compared to the Norwegian registry; however, the associated 95% CI were much wider (Table 2). The calibration statistic for the model predicting 1-year outcomes was adequately low (2.6) and the associated non-significant *p* value (0.10) indicates the model is well calibrated (no significant difference between observed and predicted probabilities of revision). The 2-year prediction model demonstrated evidence of

misclassification in the validation data set (high calibration statistic [11.7] and significant  $p < 0.01$ ), similar to the results seen in the Norwegian and Danish data [11].

## DISCUSSION

The most important finding of this study was that when patients in the STABILITY 1 trial who received HT + LET were coded as BPTB in the Norwegian prediction calculator, the model concordance was similar to the index study. However, the 95% CI for the validation set was wider than the original model, suggesting that more data are required to definitively determine the external validity. Further, model calibration was better for predicting revision surgery within 1-year and worse when predicting 2-year outcomes. This finding is in keeping with the original study and with prediction modelling in general, as predicting outcomes over longer time-periods is typically more challenging due to the increased outcome variability observed over time.

At first glance, the performance of the NKLR model presented in Table 2 may appear impressive—the discriminative ability of the algorithm to properly order patients regarding their revision risk (concordance) was higher in the external validation set relative to both the initial model validation and the external validation using Danish patients [9, 11]. Closer inspection, however, reveals a wide CI that extends beyond both ends of the NKLR model performance CI. This is an important distinction as it suggests the true accuracy of the model for the STABILITY 1 patient population remains unknown. A larger sample size would be necessary to narrow this CI and more clearly ascertain the performance of the NKLR model on this different patient data set.

Model performance was worse when the patients receiving HT + LET were coded as having received either HT or 'Unknown/Other' graft choices. The finding that the failure rate for HT + LET ACLR is most similar to BPTB ACLR is consistent with the literature [16]. In a previous study, the STABILITY 1 data set was used for external validation of the Multicenter Orthopaedic Outcomes Network (MOON) autograft risk calculator, which included either HT or BPTB as the graft type [16]. The validation analysis was run once with the HT + LET group coded as HT only and once coded as BPTB. Mirroring results of the present study, the risk calculator was most predictive (AUC = 0.73) when the LET group was coded as the BPTB group. Hamstring tendon and BPTB autograft both have a long history of use for ACLR, but several studies have found higher failure rates among patients receiving HT [19–21]. This difference is especially apparent in young active patients and has influenced both clinical practice and innovation within the specialty. In Norway, HT was the graft of choice until approximately 2015 when a NKLR-based study found higher failure rates versus BPTB [20]. In 2012, 79% of patients received HT

TABLE 2 Model performance.

Probability of revision	Model	Concordance	Calibration statistic (quintile method)	Calibration <i>p</i> value
1 year	Original Norwegian Algorithm Performance <sup>a</sup>	0.686 (0.652–0.721)	4.9	n.s.
	STABILITY data	0.713 (0.634–0.791)	2.6	n.s.
	HT = HT			
	HT + LET = BPTB			
	STABILITY data	0.609 (0.528–0.691)	10.6	<0.01*
	HT = HT			
	HT + LET = Unknown			
	STABILITY data	0.674 (0.597–0.751)	8.7	<0.01*
2 years	All patients = HT			
	Original Norwegian Algorithm Performance <sup>a</sup>	0.684 (0.650–0.718)	11.3	0.01*
	STABILITY data	0.713 (0.637–0.789)	11.7	<0.01*
	HT = HT			
	HT + LET = BPTB			
	STABILITY data	0.608 (0.530–0.688)	8.9	<0.01*
	HT = HT			
	HT + LET = Unknown			
	STABILITY data	0.673 (0.598–0.747)	10.2	<0.01*
	All patients = HT			

Abbreviations: BPTB, bone-patellar tendon-bone autograft; HT, hamstring tendon autograft; LET, lateral extra-articular tenodesis; n.s., not statistically significant.

<sup>a</sup>See Martin et al. [9].

\*Statistical significance,  $p \leq 0.05$ .

autograft while in 2016 that number dropped to 32%, with BPTB representing 61% of all ACLR grafts that year. The study of LET as an augment to HT ACLR represents an important innovation in response to inferior outcomes with HT alone and has consistently demonstrated lower failure rates than HT alone [12, 22, 23]. These results reflect why HT + LET acts more like BPTB than HT alone or other types of grafts in predictive models of graft failure/revision surgery.

The orthopaedic literature has seen an exponential increase in studies applying a machine learning approach to data analysis. Most of these studies have sought automatic radiologic diagnostics (computer vision), language interpretation (natural language processing) or outcome prediction. While the number of novel machine learning and deep learning models has proliferated, very few have completed the important step of external validation.

While it is crucial to perform prior to widespread adoption and implementation of these models, external validation can present several challenges for clinician scientists. First, a large volume of data is required for model validation, and it can be difficult to find a suitably large study population with the necessary variables

required for external validation. Ideally, patient populations should be similar with regard to the nature of data collection and tracking of outcomes for appropriate model evaluation. Another limitation is the possibility of data transfer barriers between nations or health regions due to local legislation and privacy concerns. For this reason, it is often easiest to share machine learning algorithms rather than patient data and requires collaborative efforts between study groups. Finally, most machine learning algorithms demonstrate a drop-off in performance during external validation, which raises the question of what constitutes acceptable model performance in this setting.

The debate regarding acceptable model performance is especially pertinent when evaluating models to predict outcomes such as ACL revision surgery which will likely never achieve excellent or perfect performance. These designations have historically been reserved for models with discrimination values greater than 0.90. Given the randomness associated with subsequent ACL graft rupture and the multiple variables which may contribute to the decision to proceed with revision surgery, the traditional interpretation of model performance is not realistic for clinical models like this one. There is also concern that models that do demonstrate 'excellent'



discriminative ability may be the product of model overfitting, limiting their real-world performance. Ultimately, the development of clinically useful outcome prediction models relies on a three-step approach consisting of model development, external validation and comparison with expert human prediction performance. It is this final step which establishes the baseline above which prediction models must perform in order to be clinically relevant. At this time, no such comparison exists for ACL outcome prediction and represents the next step in ACL outcome prediction research.

There were some limitations to the current study. First, the patient populations were different in several ways. Due to the standardization in the STABILITY 1 randomized trial, the data set includes a narrow age range from 14 to 25 years old compared to a much wider age range represented in the NKLR. Since all patients had HT autograft ACLR, the femoral fixation was universally suspensory/cortical leading to no variation in this predictor. Further, due to the nature of the STABILITY 1 trial, the chronicity of injury (time from injury to surgery in years) is much smaller and more standardized than observed in the NKLR. Another limitation is the fact that the sample size used for external validation was small, with few observed revision surgery events within 2 years ( $n = 30$ , 5%). This required a change in methodology for the calculation of model calibration compared with the index study. The calculation of model calibration was further complicated by the fact that patients in the STABILITY 1 trial were enrolled based on inclusion criteria that identified them as being particularly high-risk for ACLR failure.

Although the results of this study did not confirm the external validity of the NKLR revision prediction model, model performance with this separate cohort of patients was encouraging and will prompt further evaluation using larger patient data sets. Additionally, LET was only added as a variable collected by the NKLR in June 2019 and therefore was not considered during the original prediction model development. As the rate of LET during ACLR increases, it is important to consider how this adjunctive procedure may affect outcome predictions. This study, in keeping with the findings of Marmura et al. [16], suggests that outcome prediction for patients receiving LET in addition to HT ACLR may be more accurate if they are coded as BPTB in the revision prediction tool.

## CONCLUSION

When patients in STABILITY 1 who received HT + LET were coded as BPTB in the NKLR prediction model, concordance was similar to the index study. However, due to a wide 95% CI, the true performance of the prediction model with this Canadian and European cohort is unclear and a larger data set is required to definitively determine the external validity. Further, better calibration for 1-year predictions aligns with

general prediction modelling challenges over longer periods. While not a large enough sample size to elicit the true accuracy and external validity of the prediction model when applied to North American patients, this analysis provides more support for the notion that HT plus LET performs similarly to BPTB reconstruction. In addition, despite the wide confidence interval, this study suggests optimism regarding the accuracy of the model when applied outside of Scandinavia.

## AUTHOR CONTRIBUTIONS

R. Kyle Martin was involved in study conception, design, interpretation of data analysis, and manuscript preparation and editing. Hana Marmura was involved in study design, carried out the data acquisition, analysis, and interpretation, and was involved in manuscript preparation and editing. Solvejg Wastvedt was involved in study design, data analysis and interpretation, and was involved in manuscript preparation and editing. Ayoosh Pareek was involved in study conception, design, interpretation of data analysis, and manuscript editing. Andreas Persson was involved in study conception, design, interpretation of data analysis, and manuscript preparation and editing. Gilbert Moatshe was involved in study conception, design, interpretation of data analysis, and manuscript preparation and editing. Dianne Bryant was involved in design and manuscript editing. Julian Wolfson was involved in study design, data interpretation, and was involved in manuscript editing. Lars Engebretsen was involved in study conception, design, interpretation of data analysis, and manuscript preparation and editing. Alan Getgood was involved in study conception, design, interpretation of data analysis, and manuscript preparation and editing. All authors have given final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## ACKNOWLEDGEMENTS

This study was funded by a Norwegian Centennial Chair Seed Grant.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## ETHICS STATEMENT

The original prediction model development was performed at the University of Minnesota and the respective Institutional Review Board similarly concluded that the study was exempt from full review (#00012552). The STABILITY 1 trial was approved by the Western Ontario Health Sciences Research Ethics Board (#104524). Approval not required as informed consent was obtained by all patients at the time of enrolment in the national knee ligament register.

## ORCID

R. Kyle Martin  <http://orcid.org/0000-0001-9918-0264>

## REFERENCES

- Webster, K.E. & Feller, J.A. (2016) Exploring the high reinjury rate in younger patients undergoing anterior cruciate ligament reconstruction. *The American Journal of Sports Medicine*, 44(11), 2827–2832.
- Webster, K.E., Feller, J.A., Leigh, W.B. & Richmond, A.K. (2014) Younger patients are at increased risk for graft rupture and contralateral injury after anterior cruciate ligament reconstruction. *The American Journal of Sports Medicine*, 42(3), 641–647.
- Cristiani, R., Forssblad, M., Edman, G., Eriksson, K. & Stålman, A. (2021) Age, time from injury to surgery and quadriceps strength affect the risk of revision surgery after primary ACL reconstruction. *Knee Surgery, Sports Traumatology, Arthroscopy*, 29(12), 4154–4162.
- Gifstad, T., Foss, O.A., Engebretsen, L., Lind, M., Forssblad, M., Albrektsen, G. et al. (2014) Lower risk of revision with patellar tendon autografts compared with hamstring autografts: a registry study based on 45,998 primary ACL reconstructions in Scandinavia. *The American Journal of Sports Medicine*, 42(10), 2319–2328.
- Jaeger, V., Drouven, S., Naendrup, J.-H., Kanakamedala, A.C., Pfeiffer, T. & Shafizadeh, S. (2018) Increased medial and lateral tibial posterior slopes are independent risk factors for graft failure following ACL reconstruction. *Archives of Orthopaedic and Trauma Surgery*, 138(10), 1423–1431.
- Kaeding, C.C., Pedroza, A.D., Reinke, E.K., Huston, L.J., Spindler, K.P., Amendola, A. et al. (2015) 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. *The American Journal of Sports Medicine*, 43(7), 1583–1590.
- Martin, R.K., Ley, C., Pareek, A., Groll, A., Tischer, T. & Seil, R. (2022) Artificial intelligence and machine learning: an introduction for orthopaedic surgeons. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30(2), 361–364.
- Pruneski, J.A., Pareek, A., Kunze, K.N., Martin, R.K., Karlsson, J., Oeding, J.F. et al. (2023) Supervised machine learning and associated algorithms: applications in orthopedic surgery. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31(4), 1196–1202.
- Martin, R.K., Wastvedt, S., Pareek, A., Persson, A., Visnes, H., Fenstad, A.M. et al. (2021) Predicting anterior cruciate ligament reconstruction revision: a machine learning analysis utilizing the Norwegian Knee Ligament Register. *Journal of Bone and Joint Surgery*, 104(2), 145–153.
- Martin, R.K., Wastvedt, S., Pareek, A., Persson, A., Visnes, H., Fenstad, A.M. et al. (2022) Predicting Subjective Failure of ACL Reconstruction: A Machine Learning Analysis of the Norwegian Knee Ligament Register and Patient Reported Outcomes. *Journal of ISAKOS*, 7(3), 1–9.
- Martin, R.K., Wastvedt, S., Pareek, A., Persson, A., Visnes, H., Fenstad, A.M. et al. (2022) Machine learning algorithm to predict anterior cruciate ligament revision demonstrates external validity. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30(2), 368–375.
- Getgood, A.M.J., Bryant, D.M., Litchfield, R., Heard, M., McCormack, R.G. & Rezansoff, A. et al. (2020) Lateral extra-articular tenodesis reduces failure of hamstring tendon autograft anterior cruciate ligament reconstruction: 2-year outcomes from the STABILITY Study randomized clinical trial. *The American Journal of Sports Medicine*, 48(2), 285–297.
- Granan, L.-P., Bahr, R., Steindal, K., Furnes, O. & Engebretsen, L. (2008) Development of a national cruciate ligament surgery registry: the Norwegian National Knee Ligament Registry. *The American Journal of Sports Medicine*, 36(2), 308–315.
- Norwegian Knee Ligament Register: Annual Report 2022. p. 283. Accessed June 2023. <https://www.helse-bergen.no/nasjonalt-kvalitets-og-kompetansenettverk-for-leddproteser-og-hoftebrudd/arsrapporter>
- Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 1–13.
- Marmura, H., Getgood, A.M.J., Spindler, K.P., Kattan, M.W., Briskin, I. & Bryant, D.M. (2021) Validation of a risk calculator to personalize graft choice and reduce rupture rates for anterior cruciate ligament reconstruction. *The American Journal of Sports Medicine*, 49(7), 1777–1785.
- Vock, D.M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P.E., Vazquez-Benitez, G. et al. (2016) Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, 61, 119–131.
- Harrell, F.E. (1982) Evaluating the yield of medical tests. *JAMA*, 247(18), 2543–2546.
- Spindler, K.P., Huston, L.J., Zajichek, A., Reinke, E.K., Amendola, A., Andrich, J.T. et al. (2020) Anterior cruciate ligament reconstruction in high school and college-aged athletes: does autograft choice influence anterior cruciate ligament revision rates? *The American Journal of Sports Medicine*, 48(2), 298–309.
- Persson, A., Fjeldsgaard, K., Gjertsen, J.-E., Kjellsen, A.B., Engebretsen, L., Hole, R.M. et al. (2014) Increased risk of revision with hamstring tendon grafts compared with patellar tendon grafts after anterior cruciate ligament reconstruction: a study of 12,643 patients from the Norwegian Cruciate Ligament Registry, 2004–2012. *The American Journal of Sports Medicine*, 42(2), 285–291.
- Schuetz, H.B., Kraeutler, M.J., Houck, D.A. & McCarty, E.C. (2017) Bone–Patellar tendon–bone versus Hamstring tendon autografts for primary anterior cruciate ligament reconstruction: a systematic review of overlapping meta-analyses. *Orthopaedic Journal of Sports Medicine*, 5(11), 232596711773648.
- Borque, K.A., Jones, M., Laughlin, M.S., Balendra, G., Willinger, L., Pinheiro, V.H. et al. (2022) Effect of lateral extra-articular tenodesis on the rate of revision anterior cruciate ligament reconstruction in elite athletes. *The American Journal of Sports Medicine*, 50(13), 3487–3492.
- Park, Y.-B., Lee, H.-J., Cho, H.-C., Pujol, N. & Kim, S.H. (2023) Combined lateral extra-articular tenodesis or combined anterolateral ligament reconstruction and anterior cruciate ligament reconstruction improves outcomes compared to isolated reconstruction for anterior cruciate ligament tear: a network meta-analysis of randomized controlled trials. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 39(3), 758–776.

**How to cite this article:** Martin, R.K., Marmura, H., Wastvedt, S., Pareek, A., Persson, A., Moatshe, G. et al. (2024) External validation of the Norwegian anterior cruciate ligament reconstruction revision prediction model using patients from the STABILITY 1 Trial. *Knee Surgery, Sports Traumatology, Arthroscopy*, 32, 206–213. <https://doi.org/10.1002/ksa.12031>



## Paper VI

Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, Moatshe G, Wolfson J, Lind M, Engebretsen L. Unsupervised Machine Learning of the Combined Danish and Norwegian Knee Ligament Registers: Identification of 5 Distinct Patient Groups With Differing ACL Revision Rates. *Am J Sports Med.* 2024;52(4):881-891. doi:10.1177/03635465231225215





# Unsupervised Machine Learning of the Combined Danish and Norwegian Knee Ligament Registers

## Identification of 5 Distinct Patient Groups With Differing ACL Revision Rates

R. Kyle Martin,<sup>\*†‡§</sup> MD , Solvejg Wastvedt,<sup>||</sup> BA, Ayoosh Pareek,<sup>¶</sup> MD, Andreas Persson,<sup>§#\*\*</sup> MD, PhD, Håvard Visnes,<sup>§\*\*††</sup> MD, PhD, Anne Marie Fenstad,<sup>\*\*</sup> MSc, Gilbert Moatshe,<sup>§#</sup> MD, PhD, Julian Wolfson,<sup>||</sup> PhD, Martin Lind,<sup>‡‡</sup> MD, PhD, and Lars Engebretsen,<sup>§#</sup> MD, PhD

*Investigation performed at the University of Minnesota, Minneapolis, Minnesota, USA*

**Background:** Most clinical machine learning applications use a supervised learning approach using labeled variables. In contrast, unsupervised learning enables pattern detection without a prespecified outcome.

**Purpose/Hypothesis:** The purpose of this study was to apply unsupervised learning to the combined Danish and Norwegian knee ligament register (KLR) with the goal of detecting distinct subgroups. It was hypothesized that resulting groups would have differing rates of subsequent anterior cruciate ligament reconstruction (ACLR) revision.

**Study Design:** Cohort study; Level of evidence, 3.

**Methods:** K-prototypes clustering was performed on the complete case KLR data. After performing the unsupervised learning analysis, the authors defined clinically relevant characteristics of each cluster using variable summaries, surgeons' domain knowledge, and Shapley Additive exPlanations analysis.

**Results:** Five clusters were identified. Cluster 1 (revision rate, 9.9%) patients were young (mean age, 22 years; SD, 6 years), received hamstring tendon (HT) autograft (91%), and had lower baseline Knee injury and Osteoarthritis Outcome Score (KOOS) Sport and Recreation (Sports) scores (mean, 25.0; SD, 15.6). Cluster 2 (revision rate, 6.9%) patients received HT autograft (89%) and had higher baseline KOOS Sports scores (mean, 67.2; SD, 16.5). Cluster 3 (revision rate, 4.7%) patients received bone-patellar tendon-bone (BPTB) or quadriceps tendon (QT) autograft (94%) and had higher baseline KOOS Sports scores (mean, 65.8; SD, 16.4). Cluster 4 (revision rate, 4.1%) patients received BPTB or QT autograft (88%) and had low baseline KOOS Sports scores (mean, 20.5; SD, 14.0). Cluster 5 (revision rate, 3.1%) patients were older (mean age, 42 years; SD, 7 years), received HT autograft (89%), and had low baseline KOOS Sports scores (mean, 23.4; SD, 17.6).

**Conclusion:** Unsupervised learning identified 5 distinct KLR patient subgroups and each grouping was associated with a unique ACLR revision rate. Patients can be approximately classified into 1 of the 5 clusters based on only 3 variables: age, graft choice (HT, BPTB, or QT autograft), and preoperative KOOS Sports subscale score. If externally validated, the resulting groupings may enable quick risk stratification for future patients undergoing ACLR in the clinical setting. Patients in cluster 1 are considered high risk (9.9%), cluster 2 patients medium risk (6.9%), and patients in clusters 3 to 5 low risk (3.1%-4.7%) for revision ACLR.

**Keywords:** ACL revision; outcome prediction; machine learning; artificial intelligence; unsupervised learning

Machine learning represents an increasingly used approach within the orthopaedic literature due to the ability to process large volumes of complex data and develop clinically useful diagnostic, prognostic, or data collection

models.<sup>30,32</sup> The 3 main categories of machine learning approaches are supervised learning, unsupervised learning, and reinforcement learning. Most of the orthopaedic studies to date have applied a supervised learning approach, referring to the analysis of labeled data. In the supervised learning approach, the computer algorithm is provided with variables that are labeled as either a "predictor" or an "outcome," and the model is tasked with predicting a specified outcome. In contrast, unsupervised learning involves the analysis of unlabeled data whereby the model

is tasked with independently finding patterns in the data set. This process enables the interpretation and simplification of highly complex data through the identification of hidden structures and patterns.<sup>7</sup>

Within orthopaedic research, unsupervised learning approaches have recently been used to stratify groups of patients according to their risk of hip osteoarthritis progression<sup>17</sup> and to identify subphenotypes of osteoarthritis based on blood-based biochemical markers.<sup>2</sup> These examples highlight how a novel approach to a common problem can provide insight into the factors associated with complex clinical conditions. Outcome after anterior cruciate ligament (ACL) injury and subsequent ACL reconstruction (ACLR) is one such example of a clinical condition that evades complete understanding, despite troves of literature on the subject. Studies from the national knee ligament registers, Multicenter Orthopaedic Outcomes Network, and others have helped identify age, activity level, graft choice, fixation device, and posterior tibial slope as some factors that influence failure risk.<sup>5,10,16,28,29,34,41</sup> Despite recognition of these and other risk factors for a poor outcome,<sup>11,19,25,35</sup> along with recent advancements in surgical decision-making and techniques,<sup>6,8,27,33,39</sup> highly accurate clinical prediction models remain elusive. One constraint to accurate patient-specific outcome prediction is the sheer volume of risk factors that may contribute to a patient's outcome and, specifically, the limited ability to synthesize the complex and often unrecognized interactions between these factors.

The Norwegian Knee Ligament Register (NKLK) and Danish Knee Ligament Reconstruction Registry (DKRR) have been prospectively collecting data related to ACLR in their respective countries for nearly 20 years.<sup>9,31</sup> Since their inception, these national registers have produced several studies on ACL treatment and outcomes and have recently developed preliminary outcome prediction models using supervised machine learning methodology.<sup>15,20-23</sup> The present study sought to further investigate the factors associated with subsequent ACLR revision through the application of unsupervised learning techniques to the combined Norwegian and Danish knee ligament register (KLR). The primary goal of this analysis was to identify distinct

subgroups of patients within the registers and determine if the rate of subsequent revision ACLR differs between the patient clusters. The hypothesis was that unsupervised learning would facilitate the grouping of patients based on common characteristics and that this would enable the identification of high- and low-risk groups of patients.

## METHODS

### Ethics

Informed consent was obtained prospectively from all patients enrolled in the NKLK and the Norwegian Data Inspectorate grants permission for the NKLK to collect, analyze, and publish on these health data. Data registration was performed according to European Union data protection rules, with all data deidentified before retrieval. The regional ethics committee stated that further ethics approval was not necessary.<sup>9</sup> Similarly, the DKRR prospectively obtained informed consent at the time of enrollment and patient data were deidentified before retrieval with no further ethics approval required.

### Data Preparation

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

The activity that reportedly led to ACL injury was classified as a pivoting sport, nonpivoting sport, or other activity. Meniscal injuries were classified as present with repair, present without repair (no treatment or partial meniscectomy), or no meniscal injury. Cartilage injuries were grouped according to the International Cartilage Regeneration & Joint Preservation Society grading system and recorded as grade 1 or 2, grade 3 or 4, or no cartilage injury. Additionally, a predictor indicating if a patient

\*Address correspondence to R. Kyle Martin, MD, Department of Orthopedic Surgery, University of Minnesota, 2512 South 7th St, Suite R200, Minneapolis, MN 55455, USA (email: rkylemartin@gmail.com) (X: @UMNOrthoSurg; @RKMartin6; @nih\_physicperf).

<sup>1</sup>Department of Orthopedic Surgery, University of Minnesota, Minneapolis, Minnesota, USA.

<sup>2</sup>Department of Orthopedic Surgery, CentraCare, Saint Cloud, Minnesota, USA.

<sup>3</sup>Oslo Sport Trauma Research Center, Norwegian School of Sports Science, Oslo, Norway.

<sup>4</sup>Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA.

<sup>5</sup>Department of Orthopedic Surgery, Hospital for Special Surgery, New York, New York, USA.

<sup>6</sup>Orthopaedic Clinic, Oslo University Hospital Ullevål, Oslo, Norway.

<sup>7</sup>Norwegian Knee Ligament Register, Haukeland University Hospital, Bergen, Norway.

<sup>8</sup>Department of Orthopedics, Sorlandet Hospital, Kristiansand, Norway.

<sup>9</sup>Aarhus University Hospital, Aarhus, Denmark.

Submitted September 17, 2023; accepted November 29, 2023.

One or more of the authors has declared the following potential conflict of interest or source of funding: This study was funded by a Norwegian Centennial Chair seed grant. R.K.M. has received support for education from Gemini/Arthrex and Smith & Nephew and hospitality payments from Foundation Medical and Medical Device Business Services. G.M. has received consulting fees from Arthrex and IBSA. M.L. has received consulting fees from Smith & Nephew. L.E. has received research support from Biomet and Health Southeast Norway and royalties from Arthrex and Smith & Nephew. A.Pareek. has received support for education from Smith & Nephew and hospitality payments from Medical Device Business Services. AOSSM checks author disclosures against the Open Payments Database (OPD). AOSSM has not conducted an independent investigation on the OPD and disclaims any liability or responsibility relating thereto.

TABLE 1  
Patient Characteristics<sup>a</sup>

Variable	Combined Data, n = 62,955	Complete Case Data, N = 28,631
Revision	3205 (5.1)	1770 (6.2)
Mean follow-up time or time to revision, y	7.6 ± 4.5	8.2 ± 4.5
Mean age at surgery, y	28 ± 11	28 ± 10
Mean age at injury, y	27 ± 10	27 ± 10
Missing	1870	
Sex		
Male	36,509 (58)	15,671 (55)
Female	26,446 (42)	12,960 (45)
Mean presurgery KOOS QOL score	36.3 ± 18.0	36.5 ± 17.9
Missing	29,512	
Mean presurgery KOOS Sports score	41.2 ± 26.9	41.2 ± 26.8
Missing	29,708	
Below median on all presurgery KOOS Subscales	6372 (19)	5259 (18)
Missing	29,323	
Activity that led to injury		
Nonpivoting	20,391 (33)	8175 (29)
Pivoting	35,851 (57)	16,747 (58)
Other	6162 (9.9)	3709 (13)
Missing	551	
Meniscal injury		
Injury without repair	20,328 (32)	9568 (33)
Injury with repair	10,554 (17)	4640 (16)
None	32,061 (51)	14,423 (50)
Missing	12	
Cartilage injury (ICRS grade)		
1 or 2	8766 (14)	4195 (15)
3 or 4	3223 (5.1)	1627 (5.7)
None	50,878 (81)	22,809 (80)
Missing	88	
Graft choice		
BPTB	15,639 (25)	9000 (31)
Hamstring	43,518 (69)	18,356 (64)
QT/BQT	2520 (4.0)	888 (3.1)
Other	1278 (2.0)	387 (1.4)
Tibial fixation device		
Interference screw	55,792 (90)	25,759 (90)
Suspension/cortical device	3643 (5.9)	2031 (7.1)
Other	2356 (3.8)	841 (2.9)
Missing	1164	
Femoral fixation device		
Interference screw	16,434 (27)	8793 (31)
Suspension/cortical device	39,742 (65)	17,502 (61)
Other	4822 (7.9)	2336 (8.2)
Missing	1957	
Fixation device combination		
Interference screw × 2	15,865 (26)	8467 (30)
Interference/suspension	236 (0.4)	150 (0.5)
Suspension/cortical device × 2	2994 (4.9)	1540 (5.4)
Suspension/interference	34,895 (58)	15,493 (54)
Other	6529 (11)	2981 (10)
Missing	2436	
History of previous surgery on same knee <sup>b</sup>	10,312 (17)	4540 (16)
Missing	673	
History of previous cruciate ligament injury to opposite knee <sup>b</sup>	4839 (8.1)	1977 (6.9)
Missing	2946	

(continued)

TABLE 1  
(continued)

Variable	Combined Data, n = 62,955	Complete Case Data, N = 28,631
Median time injury to surgery, y	1.63 [0.33-1.32]	0.61 [0.33-1.29]
Missing	2083	
Register		
DKRR	34,554 (55)	10,487 (37)
NKLR	28,401 (45)	18,144 (63)

<sup>a</sup>Data are presented as n, n (%), mean  $\pm$  SD, or median [IQR]. BPTB, bone-patellar tendon-bone autograft; DKRR, Danish Knee Ligament Register; ICRS, International Cartilage Regeneration & Joint Preservation Society; KOOS, Knee injury and Osteoarthritis Outcome Score; NKLR, Norwegian Knee Ligament Register; QOL, Quality of Life subscale; QT/BQT, quadriceps tendon autograft, with or without bone; Sports, Sport and Recreation subscale.

<sup>b</sup>Surgery performed before primary anterior cruciate ligament reconstruction and enrollment in the register.

was below the median score in the respective register on all presurgery Knee injury and Osteoarthritis Outcome Score (KOOS) variables was also created.

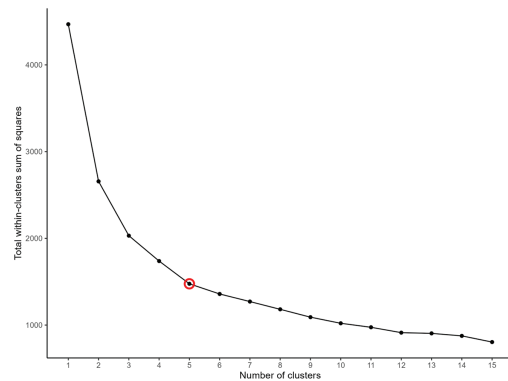
#### Missing Data

A previous study applying supervised machine learning models to the combined KLR data guided the approach to missing data for this study.<sup>20</sup> Briefly, supervised learning models were trained and evaluated using complete data on all variables. This was then repeated using multiple imputation to assess the effect of restricting data to complete cases. 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 and found that multiply imputed data were not notably different from the complete case analysis. This means that there was no meaningful advantage of data imputation for the predictive modeling. Therefore, for this study only patients with complete data on all predictors were included in the analysis.

#### Unsupervised Learning

The machine learning methods used in this analysis were all unsupervised, meaning the models were not trained to produce predictions for a specific outcome variable. Instead, unsupervised methods model how the data were organized with respect to a given set of predictor variables.<sup>13</sup> The applied unsupervised methods produced groups, or clusters, of observations with similar relationships among the predictor variables. Because unsupervised learning does not train and then test predictions, the sample was not split. The entire sample of patients was used to build the unsupervised models and characterize the resulting clusters. All analyses were conducted in R (Version 4.1.1; R Core Team).

Three unsupervised clustering methods were applied: k-means (function *kmeans*; package *stats*), agglomerative hierarchical clustering (function *hclust*; package *fastcluster*<sup>26</sup>), and k-prototypes (function *kproto*; package



**Figure 1.** The elbow method to determine the number of clusters for unsupervised learning analysis. The location where the line bends sharply upward (circle) signifies the elbow, representing the optimal number of clusters.

*clustMixType*<sup>36</sup>). K-means clustering required the user to prespecify the number of clusters. The algorithm then grouped the observations to minimize the sum of squares from points to the cluster centers.<sup>12</sup> To determine the number of clusters (*k*), a common technique called the elbow method was used. In this approach, clusters were computed for various possible values of *k*, and the within-cluster sums of squares were calculated and plotted against the value of *k*. The point at which this line bent sharply upward (the elbow) dictated the optimal number of clusters (Figure 1).

This represented the fewest clusters that could be created without a sharp increase in within-cluster heterogeneity.<sup>38</sup> Agglomerative hierarchical clustering began with each observation in its own cluster and yielded many possible partitions of decreasing complexity, requiring the user to select a level of complexity (by specifying a desired number of clusters).

K-means and agglomerative hierarchical clustering only accommodates continuous predictor variables. To overcome this limitation, a third method, k-prototypes,

which accommodated mixed type predictors, was used. K-prototypes is similar to k-means in that it minimizes within-cluster distance from the cluster mean when assigning observations to a prespecified number of clusters. The distance metric was a weighted combination of Euclidean distance for continuous variables and the count of mismatched category labels for categorical variables. A data-driven technique was used to select the weighting parameter. The cluster “mean” was the mean for continuous variables and the mode for categorical variables. The elbow and silhouette methods were used to define the optimal number of k-prototypes clusters. The silhouette method identified the number of clusters that maximized between-cluster and minimized within-cluster dissimilarity.<sup>36</sup>

### Measures of Cluster Quality

Unlike with supervised learning where models are trained on a training set and evaluated against observed labels on a test set, with unsupervised learning there are no labels for comparison. Assessing the quality of model results is therefore more challenging and typically relies on heuristic arguments and domain knowledge.<sup>13</sup> Therefore, a combination of 2 data-driven methods (elbow and silhouette) and domain knowledge was used to choose the number of clusters.

### Model Interpretability and Clinical Relevance

To identify the defining characteristics of each cluster, 7 orthopaedic surgeons (R.K.M., A.Pareek., A.Persson., H.V., G.M., M.L., L.E.) with subspecialty training in sports medicine reviewed the patient groups and highlighted the clinically relevant features based on their domain knowledge and variable summaries. The goal was to define each cluster in terms that would enable the assignment of future patients to 1 of the 5 clusters. To aid in cluster interpretation, SHapley Additive exPlanations (SHAP) analysis was also performed.<sup>18</sup> This required a 2-step process: (1) build a classification model predicting clusters from input variables and (2) compute SHAP values for this classification model. First, a gradient boosting model was trained to predict the cluster number using all predictor variables originally used for clustering (R package *xgboost*). Gradient boosting is a tree-based machine learning method that can be used for classification with multiple classes, such as in this situation.<sup>4</sup> Next, SHAP values were computed for this model using built-in functions in the *xgboost* package. The SHAP values explained the contributions of input variables in each cluster by summarizing their influence on individual predictions. Cluster-specific Kaplan-Meier curves were created to describe each cluster's mean risk of revision surgery.

## RESULTS

### Participants

After data cleaning, a process whereby incorrect, duplicate, or incomplete data were removed or corrected, the

combined register population consisted of 62,955 patients, 55% from the DKRR and 45% from the NKLR. The primary outcome, revision surgery, occurred in 5.1% of patients during a mean follow-up time of 7.6 years (SD, 4.5 years). The population was 55% male with median ages at primary injury and surgery of 24 years (IQR, 18-34 years) and 26 years (IQR, 20-36 years), respectively. After removing patients with missing predictor variables, the study population consisted of 28,631 patients. Characteristics of the study population at the time of surgery along with the complete case data set are presented in Table 1.

### Clustering Results

The k-prototypes method was chosen because it accommodated both continuous and categorical predictors. The optimal number of clusters was set at 5 via a combination of the data-driven elbow and silhouette methods and domain knowledge (Figure 1). A description of the 5 clusters is presented in Table 2 and Figure 2. Figure 3 presents the SHAP values for all clusters. Cluster-specific Kaplan-Meier curves demonstrating the revision risk profiles for the 5 patient groups are presented in Figure 4.

Surgeon domain knowledge and SHAP values were used to interpret the variable summaries and simplify the distinguishing characteristics of each cluster for clinical relevance. Cluster 1 (revision rate, 9.9%) patients were young (mean age, 22 years; SD, 6 years) and more often female (60%), received hamstring tendon (HT) autograft (91%), and had lower baseline KOOS Sport and Recreation (Sports) scores (mean, 25.0; SD, 15.6). Cluster 2 (revision rate, 6.9%) patients received HT autograft (89%), were more often male (68%), and had higher baseline KOOS Sports scores (mean, 67.2; SD, 16.5). Cluster 3 (revision rate, 4.7%) patients received bone-patellar tendon-bone (BPTB) or quadriceps tendon (QT) autograft (94%) and had higher baseline KOOS Sports scores (mean, 65.8; SD, 16.4). Cluster 4 (revision rate, 4.1%) patients received BPTB or QT autograft (88%) and had low baseline KOOS Sports scores (mean, 20.5; SD, 14.0). Cluster 5 (revision rate, 3.1%) patients were older (mean, 42; SD, 7 years), underwent ACLR with HT autograft (89%), and had low baseline KOOS Sports scores (mean, 23.4; SD, 17.6).

## DISCUSSION

The most important finding of this study was that unsupervised learning analysis of the combined KLR identified 5 distinct patient subgroups among patients undergoing primary ACLR, which are clinically distinguishable based on age, graft type, and baseline KOOS Sports score. Each grouping was associated with its own unique rate of subsequent ACLR revision. If externally validated, the results of this analysis could be applied in the clinical setting to classify patients into 1 of the 5 clusters. This would enable rapid estimation of the risk of subsequent revision ACLR and could be used to guide preoperative discussions and

TABLE 2  
Characteristics of Clusters Using k-Prototypes Method<sup>a</sup>

Variable	Cluster 1, n = 7038	Cluster 2, n = 7693	Cluster 3, n = 4118	Cluster 4, n = 4852	Cluster 5, n = 4930
Revision	695 (9.9)	532 (6.9)	193 (4.7)	198 (4.1)	152 (3.1)
Mean follow-up time or time to revision, y	8.2 ± 4.3	8.5 ± 4.3	7.5 ± 4.8	8.0 ± 5.0	8.8 ± 4.2
Mean age at surgery	22 ± 6	25 ± 9	25 ± 9	30 ± 10	42 ± 7
Mean age at injury	21 ± 6	24 ± 8	23 ± 8	28 ± 9	40 ± 8
Sex					
Male	2808 (40)	5198 (68)	2473 (60)	3036 (63)	2156 (44)
Female	4230 (60)	2495 (32)	1645 (40)	1816 (37)	2774 (56)
Mean presurgery KOOS QOL score	29.7 ± 13.9	49.1 ± 16.1	47.6 ± 15.7	25.5 ± 13.4	28.4 ± 14.4
Mean presurgery KOOS Sports score	25.0 ± 15.6	67.2 ± 16.5	65.8 ± 16.4	20.5 ± 14.0	23.4 ± 17.6
Below median on all presurgery KOOS Subscales	1852 (26)	0 (0)	0 (0)	1738 (36)	1669 (34)
Activity that led to injury					
Nonpivoting	1524 (22)	1746 (23)	931 (23)	1069 (22)	2905 (59)
Pivoting	4863 (69)	5273 (69)	2730 (66)	2796 (58)	1085 (22)
Other	651 (9.2)	674 (8.8)	457 (11)	987 (20)	940 (19)
Meniscal injury					
Injury without repair	2182 (31)	2467 (32)	1277 (31)	1723 (36)	1919 (39)
Injury with repair	1305 (19)	1183 (15)	774 (19)	905 (19)	473 (9.6)
None	3551 (50)	4043 (53)	2067 (50)	2224 (46)	2538 (51)
Cartilage injury (ICRS grade)					
1 or 2	808 (11)	930 (12)	632 (15)	898 (19)	927 (19)
3 or 4	262 (3.7)	280 (3.6)	176 (4.3)	389 (8.0)	520 (11)
None	5968 (85)	6483 (84)	3310 (80)	3565 (73)	3483 (71)
Graft choice					
BPTB	424 (6.0)	579 (7.5)	3565 (87)	4035 (83)	397 (8.1)
Hamstring	6388 (91)	6884 (89)	224 (5.4)	478 (9.9)	4382 (89)
QT/BQT	152 (2.2)	142 (1.8)	270 (6.6)	252 (5.2)	72 (1.5)
Other	74 (1.1)	88 (1.1)	59 (1.4)	87 (1.8)	79 (1.6)
Tibial fixation device					
Interference screw	6101 (87)	6688 (87)	3980 (97)	4594 (95)	4396 (89)
Suspension/cortical device	700 (9.9)	771 (10)	59 (1.4)	123 (2.5)	378 (7.7)
Other	237 (3.4)	234 (3.0)	79 (1.9)	135 (2.8)	156 (3.2)
Femoral fixation device					
Interference screw	118 (1.7)	10 (0.1)	3955 (96)	4369 (90)	341 (6.9)
Suspension/cortical device	6284 (89)	6902 (90)	10 (0.2)	117 (2.4)	4189 (85)
Other	636 (9.0)	781 (10)	153 (3.7)	366 (7.5)	400 (8.1)
Fixation device combination					
Interference screw × 2	96 (1.4)	0 (0)	3845 (93)	4211 (87)	315 (6.4)
Interference screw femur/suspension tibia	15 (0.2)	8 (0.1)	50 (1.2)	63 (1.3)	14 (0.3)
Suspension/cortical device × 2	587 (8.3)	619 (8.0)	9 (0.2)	13 (0.3)	312 (6.3)
Suspension femur/interference screw tibia	5523 (78)	6102 (79)	0 (0)	97 (2.0)	3771 (76)
Other	817 (12)	964 (13)	214 (5.2)	468 (9.6)	518 (11)
History of previous surgery on opposite knee <sup>b</sup>	379 (5.4)	467 (6.1)	266 (6.5)	434 (8.9)	431 (8.7)
History of previous surgery on same knee <sup>b</sup>	1043 (15)	1002 (13)	472 (11)	878 (18)	1145 (23)
Median time injury to surgery, y	0.54 [0.30-1.13]	0.64 [0.37-1.36]	0.63 [0.34-1.22]	0.58 [0.30-1.29]	0.68 [0.36-1.56]

<sup>a</sup>Data are presented as n (%), mean ± SD, or median [IQR]. BPTB, bone–patellar tendon–bone autograft; ICRS, International Cartilage Regeneration & Joint Preservation Society; KOOS, Knee injury and Osteoarthritis Outcome Score; QOL, Quality of Life subscale; QT/BQT, quadriceps tendon autograft, with or without bone.

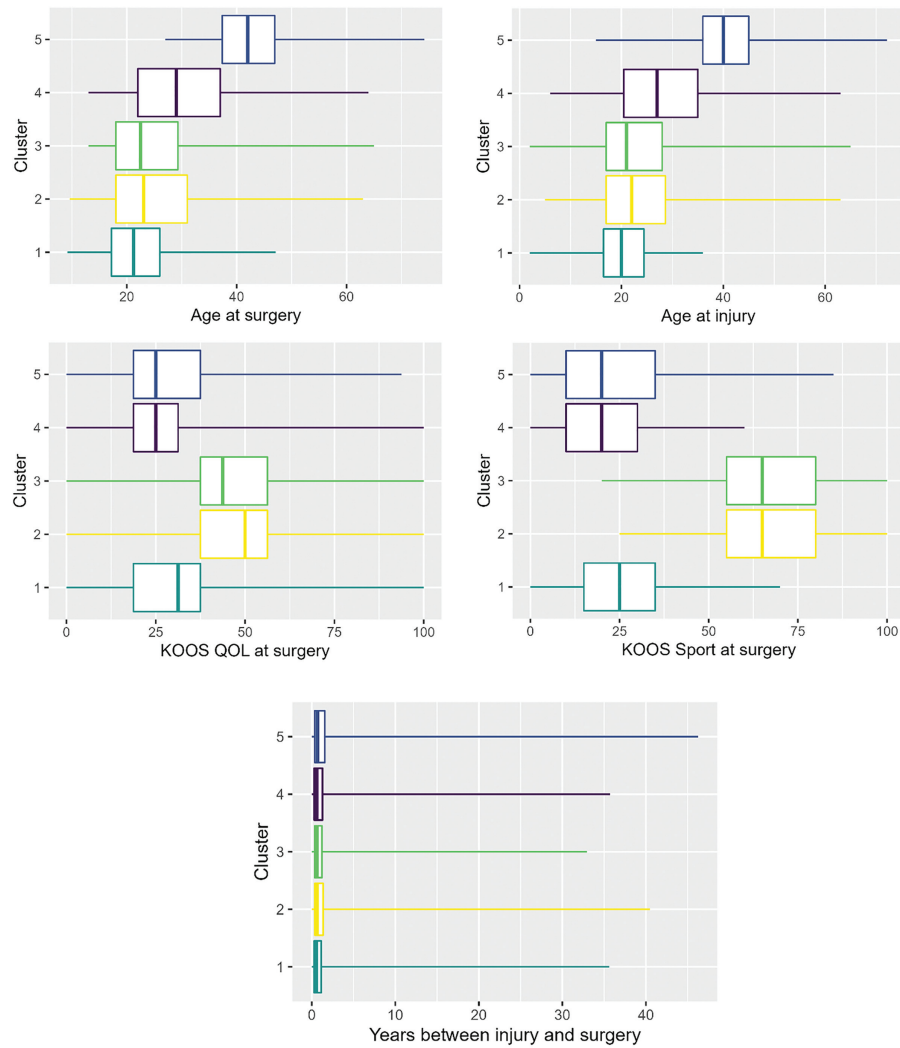
<sup>b</sup>Surgery performed before primary anterior cruciate ligament reconstruction and enrollment in the register.

surgical decision-making with patients undergoing primary ACLR.

To our knowledge, this is the first unsupervised learning analysis of an ACLR database. Unsupervised learning is a useful adjunct to clinical risk prediction efforts, as it may find patterns in data sets like the KLR without manual specification, which can be used to guide decision-making and prognostication.<sup>7</sup> Unsupervised learning

models consider all variables in the data set that are categorized as predictors and are blind to the outcome for each patient (in this case, revision surgery). The algorithm is then tasked with finding common groups of patients within the data set, breaking them into different clusters. These clusters are arrived upon through complex analysis that is not explicitly directed by human instruction. Once the clusters have been identified, the outcome can be assessed



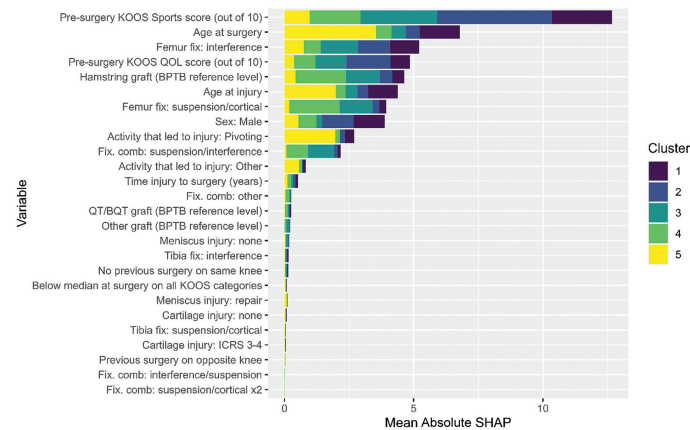


**Figure 2.** Continuous variable summaries by cluster. Box plots summarize the distributions of continuous predictor variables for each of the 5 patient subgroups identified with the unsupervised learning procedure. KOOS, Knee injury and Osteoarthritis Outcome Score; QOL, Quality of Life subscale; Sport, Sport and Recreation subscale.

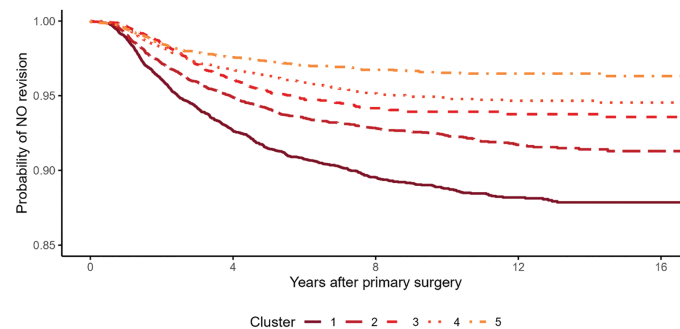
in each group. In this study, revision rate was the primary outcome of interest, and this rate was different among each of the 5 patient groups. Similarly, the survival for each cluster was also distinct, allowing for a time-dependent cluster-based estimation of revision risk.

Accurately assigning a patient to 1 of the 5 clusters requires consideration of all variables included in the

analysis. However, with so many predictor variables to consider, clinical interpretation and application of the patient subgroups can be challenging. To increase the clinical utility, the 5 patient clusters were reviewed by 7 subspecialty-trained orthopaedic sports medicine surgeons for defining characteristics. The recently developed SHAP analysis<sup>18</sup> was also applied to increase the explainability



**Figure 3.** The plot shows mean absolute SHapley Additive exPlanations (SHAP) values by variable for all clusters. Colors in the plot show the contributions from observations assigned to each cluster. BPTB, bone–patellar tendon–bone autograft; comb, combined; fix., fixation; ICRS, International Cartilage Regeneration & Joint Preservation Society; KOOS, Knee injury and Osteoarthritis Outcome Score; QOL, Quality of Life subscale; QT/BQT, quadriceps tendon autograft, with or without bone; Sports, Sport and Recreation subscale.



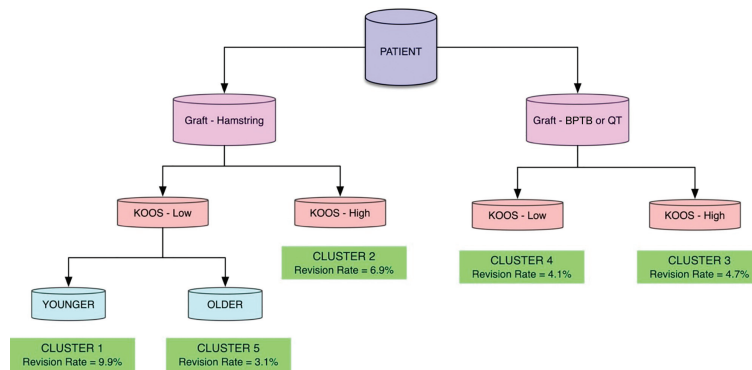
**Figure 4.** Kaplan-Meier survival curve for all 5 clusters.

of the model and decrease the black-box effect. The clusters were subsequently simplified into the following categories (Figure 5):

- Cluster 1: young patient with HT autograft and low baseline KOOS Sports score
- Cluster 2: patient with HT autograft and high baseline KOOS Sports score
- Cluster 3: patient with BPTB or QT autograft and high baseline KOOS Sports score
- Cluster 4: patient with BPTB or QT autograft and low baseline KOOS Sports score
- Cluster 5: older patient with HT autograft and low baseline KOOS Sports score

Based on the revision rates of each cluster, cluster 1 is considered high risk for revision surgery, cluster 2 is considered moderate risk for revision, and clusters 3 to 5 are considered low risk. While the overall revision rate in the KLR was 5.1%, nearly half (49%) of the patients fell into one of the low-risk categories (clusters 3-5) with a revision rate of 3.1% to 4.7%. On the other end of the spectrum, cluster 1 patients demonstrated a revision rate of nearly 10%.

Closer inspection of the highest risk cluster (cluster 1) reveals some interesting trends, including a higher proportion of patients with HT autograft, young age, female sex, and inferior baseline KOOS Sports scores. These factors become especially apparent when compared with clusters



**Figure 5.** Tree diagram for approximate patient classification by cluster. BPTB, bone-patellar tendon-bone autograft; KOOS, Knee Injury and Osteoarthritis Outcome Score (Sports subscale); QT, quadriceps tendon autograft, with or without bone.

2 and 5, which also consisted primarily of HT reconstruction but demonstrated revision rates closer to the mean. ACLR with HT autograft has previously been associated with higher revision surgery rates based on the NKLR.<sup>28</sup> Additionally, young age is a recognized risk factor for failure of ACLR.<sup>14,40,41</sup> Interestingly, the finding that young women receiving HT autograft (cluster 1) may be considered to have the highest risk for subsequent revision surgery is a novel finding. While it is generally accepted that female sex is associated with a higher risk of initial ACL injury,<sup>37</sup> it has not been found to be associated with higher ACLR revision rates.<sup>1,3,14,24</sup> Similarly, the authors are not aware of any literature associating preoperative KOOS Sports scores and subsequent revision risk. This unsupervised learning analysis suggests that because of the complex nature of the interactions between predictor variables, for some patients in certain circumstances, variables such as sex and preoperative patient-reported outcome measures may be important risk factors.

There are limitations to the present study. First, complete case data were available for less than half of the KLR, decreasing the number of patients available for analysis. Despite the missing data, however, >28,000 patients were included, which is sufficient for the purpose of unsupervised machine learning model development, and the inclusion of patients from 2 national databases increases generalizability. Another limitation is that the KLR is primarily composed of patients who received either HT, BPTB, or QT autograft. There were not enough patients receiving other graft choices such as allograft to have a meaningful effect on the clustering. These additional data would be useful in future studies to evaluate whether patients receiving allograft would form their own distinct clusters. The primary outcome measure of revision surgery represents another limitation, as some patients who experience graft failure or inferior clinical surgical outcome do not undergo subsequent revision surgery. Additionally, it is possible that an alternative unsupervised learning method may have yielded different results. There are

several alternative approaches to unsupervised learning, such as principal component analysis, anomaly detection, and divisive hierarchical clustering, among others. However, the 3 unsupervised learning methods evaluated with this study represent the most common and appropriate for the data type and goals of this study. Finally, other factors potentially associated with failure of ACLR, such as pivot-shift grade, tibial slope, rehabilitation details, and surgical adjuncts such as lateral extra-articular tenodesis or anterolateral ligament reconstruction, were not captured in the KLR and were not considered in the analysis. The inclusion of these variables in future data collection may yield different clustering results.



There are also limitations to the clinical interpretability of this unsupervised analysis because of the complex determination of cluster characteristics. The simplified summary of each cluster may not consider certain relevant characteristics, which may lead to inaccurate risk estimation in the office setting. Considering, for example, that nearly 12% of the patients in cluster 4 received grafts other than BPTB or QT, suggests that there is more to the groupings than simply graft choice and KOOS Sports score. Similarly, continuous variables such as age and KOOS values can be challenging to interpret, for example, when defining what constitutes the cutoff point for high or low preoperative KOOS values. Finally, because of the nature of the study investigating revision rates of unsupervised learning-based clusters, the accuracy of the risk estimates was not externally validated. This represents the most important next step before prospective clinical application is recommended.

## CONCLUSION

Unsupervised learning enabled the identification of 5 distinct KLR patient subgroups, and each grouping was associated with a unique ACLR revision rate. Patients can be approximately classified into 1 of the 5 clusters based on only 3 variables: age, graft choice (HT, BPTB, or QT

autograft), and preoperative KOOS Sports subscale score. If externally validated, the resulting groupings may enable quick risk stratification for future patients undergoing ACLR in the clinical setting. Patients in cluster 1 are considered high risk (9.9%) for subsequent revision ACLR, patients in cluster 2 medium risk (6.9%), and patients in clusters 3 to 5 low risk (3.1%-4.7%).

## ORCID iDs

R. Kyle Martin  <https://orcid.org/0000-0001-9918-0264>  
Lars Engebretsen  <https://orcid.org/0000-0003-2294-921X>

## REFERENCES

- Andermord D, Desai N, Björnsson H, Ylander M, Karlsson J, Samuelsson K. Patient predictors of early revision surgery after anterior cruciate ligament reconstruction: a cohort study of 16,930 patients with 2-year follow-up. *Am J Sports Med.* 2015;43(1):121-127.
- Angelini F, Widaa P, Mobasheri A, et al. Osteoarthritis endotype discovery via clustering of biochemical marker data. *Ann Rheum Dis.* 2022;81(5):666-675.
- Capogna BM, Mahure SA, Mollon B, Duenes ML, Rokito AS. Young age, female gender, Caucasian race, and workers' compensation claim are risk factors for reoperation following arthroscopic ACL reconstruction. *Knee Surg Sports Traumatol Arthrosc.* 2020;28(7):2213-2223.
- Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM; 2016:785-794.
- 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.
- Devitt BM, Neri T, Fritsch BA. Combined anterolateral complex and anterior cruciate ligament injury: anatomy, biomechanics and management—state-of-the-art. *J ISAKOS.* 2023;8(1):37-46.
- Eckhardt CM, Madjarova SJ, Williams RJ, et al. Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surg Sports Traumatol Arthrosc.* 2023;31(2):376-381.
- 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.
- 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.
- 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.
- Hamrin Senorski E, Svantesson E, Baldari A, et al. Factors that affect patient reported outcome after anterior cruciate ligament reconstruction—a systematic review of the Scandinavian knee ligament registers. *Br J Sports Med.* 2019;53(7):410-417.
- Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. *Appl Stat.* 1979;28(1):100.
- Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* Springer New York; 2009.
- Hettrich CM, Dunn WR, Reinke EK; MOON Group; Spindler KP. The rate of subsequent surgery and predictors after anterior cruciate ligament reconstruction: two- and 6-year follow-up results from a multicenter cohort. *Am J Sports Med.* 2013;41(7):1534-1540.
- Kaarre J, Zsidai B, Narup E, et al. Scoping review on ACL surgery and registry data. *Curr Rev Musculoskelet Med.* 2022;15(5):385-393.
- 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.
- Ko S, Pareek A, Jo C, et al. Automated risk stratification of hip osteoarthritis development in patients with femoroacetabular impingement using an unsupervised clustering algorithm: a study from the Rochester Epidemiology Project. *Orthop J Sports Med.* 2021;9(11):232596712110506.
- Lundberg S, Lee SI. A unified approach to interpreting model predictions. Preprint published online November 24, 2017. *arxiv.* doi:10.48550/arXiv.1705.07874
- Ma Y, Ao YF, Yu JK, Dai LH, Shao ZX. Failed anterior cruciate ligament reconstruction: analysis of factors leading to instability after primary surgery. *Chin Med J (Engl).* 2013;126(2):280-285.
- Martin RK, Wastvedt S, Pareek A, et al. Ceiling effect of the combined Norwegian and Danish Knee Ligament Registers limits anterior cruciate ligament reconstruction outcome prediction. *Am J Sports Med.* 2023;51(9):2324-2332.
- 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.* 2022;30(2):368-375.
- 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.* 2022;104(2):145-153.
- 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.* 2022;7(3):1-9.
- Mok AC, Fancher AJ, Vopat ML, et al. Sex-specific outcomes after anterior cruciate ligament reconstruction: a systematic review and meta-analysis. *Orthop J Sports Med.* 2022;10(2):232596712210768.
- MOON Knee Group; Spindler KP, Huston LJ, Chagin KM, et al. Ten-year outcomes and risk factors after anterior cruciate ligament reconstruction: a MOON longitudinal prospective cohort study. *Am J Sports Med.* 2018;46(4):815-825.
- Müllner D. fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. *J Stat Softw.* 2013;53(9):1-18.
- Pang L, Li P, Li T, Li Y, Zhu J, Tang X. Arthroscopic anterior cruciate ligament repair versus autograft anterior cruciate ligament reconstruction: a meta-analysis of comparative studies. *Front Surg.* 2022;9:887522.
- Persson A, Fjeldsgaard K, Gjertsen JE, et al. Increased risk of revision with hamstring tendon grafts compared with patellar tendon grafts after anterior cruciate ligament reconstruction: a study of 12,643 patients from the Norwegian Cruciate Ligament Registry, 2004-2012. *Am J Sports Med.* 2014;42(2):285-291.
- Persson A, Kjellsen AB, Fjeldsgaard K, Engebretsen L, Espehaug B, Fevang JM. Registry data highlight increased revision rates for Endobutton/Biosure HA in ACL reconstruction with hamstring tendon autograft: a nationwide cohort study from the Norwegian Knee Ligament Registry, 2004-2013. *Am J Sports Med.* 2015;43(9):2182-2188.
- Pruneski JA, Williams RJ, Nwachukwu BU, et al. The development and deployment of machine learning models. *Knee Surg Sports Traumatol Arthrosc.* 2022;30(12):3917-3923.
- Rahr-Wagner L, Lind M. The Danish Knee Ligament Reconstruction Registry. *Clin Epidemiol.* 2016;8:531-535.
- Ramkumar PN, Pang M, Polissetty T, Helm JM, Karnuta JM. Meaningless applications and misguided methodologies in artificial intelligence-related orthopaedic research propagates hype over hope. *Arthroscopy.* 2022;38(9):2761-2766.
- Riediger MD, Stride D, Coke SE, Kurz AZ, Duong A, Ayeni OR. ACL reconstruction with augmentation: a scoping review. *Curr Rev Musculoskelet Med.* 2019;12(2):166-172.

34. Salmon LJ, Heath E, Akrawi H, Roe JP, Linklater J, Pinczewski LA. 20-year outcomes of anterior cruciate ligament reconstruction with hamstring tendon autograft: the catastrophic effect of age and posterior tibial slope. *Am J Sports Med.* 2018;46(3):531-543.
35. Svantesson E, Hamrin Senorski E, Baldari A, et al. Factors associated with additional anterior cruciate ligament reconstruction and register comparison: a systematic review on the Scandinavian knee ligament registers. *Br J Sports Med.* 2019;53(7):418-425.
36. Szepannek G. clustMixType: user-friendly clustering of mixed-type data in R. *R J.* 2019;10(2):200.
37. The female ACL: why is it more prone to injury? *J Orthop.* 2016; 13(2):A1-A4.
38. University of Cincinnati. K-means cluster analysis. *UC Business Analytics R Programming Guide.* Accessed January 8, 2024. [https://uc-r.github.io/kmeans\\_clustering](https://uc-r.github.io/kmeans_clustering)
39. Vadhera AS, Knapik DM, Gursoy S, et al. Current concepts in anterior tibial closing wedge osteotomies for anterior cruciate ligament deficient knees. *Curr Rev Musculoskelet Med.* 2021;14(6):485-492.
40. Webster KE, Feller JA. Exploring the high reinjury rate in younger patients undergoing anterior cruciate ligament reconstruction. *Am J Sports Med.* 2016;44(11):2827-2832.
41. Webster KE, Feller JA, Leigh WB, Richmond AK. Younger patients are at increased risk for graft rupture and contralateral injury after anterior cruciate ligament reconstruction. *Am J Sports Med.* 2014;42(3):641-647.

