From code to clinic: theory and practice for artificial intelligence prediction algorithms

Hond, A.A.H. de

6

Feasibility of machine learning and competing risk analysis algorithms to predict outcomes from the Dutch Arthroplasty Register

Jacobien H. F. Oosterhoff*, Anne A. H. de Hond*, Rinne M. Peters, Liza N. van Steenbergen, Juliette Sorel, Wierd Zijlstra, Rudolf W. Poolman, Paul C. Jutte, Gino M. M. J. Kerkhoffs, Hein Putter, Ewout W. Steyerberg, and Job N. Doornberg **Both authors contributed equally*

6.1 ABSTRACT

6.1.1 Background

Prediction of revision in arthroplasty surgery requires competing risk analysis to calculate the absolute risk of revision. Machine learning (ML) algorithms may improve upon decision support tools based on traditional regression techniques in orthopedics and overcome surgeons' biases in risk stratification. Therefore, this study addressed the following study question: Does ML survival analysis with competing risk perform better than traditional regression models for estimating the risk of revision for patients undergoing arthroplasty surgery?

6.1.2 Methods

We developed a set of time-to-event models with revision as the event of interest and death as the competing risk using 11 datasets from previously published studies from the Dutch Arthroplasty Register. A set of predictors was identified based on the original variable selection of the included studies. We assessed the predictive performance of two state-of-the-art statistical time-to-event models for 1- 2- and 3-year follow-up: a Fine and Gray model and a cause-specific Cox model. These were compared to a ML approach consisting of a random survival forest model. The 11 datasets were all observational cohort studies that previously reported on predictors of outcome or survival following partial or total knee and hip arthroplasty. The sample size of these datasets ranged from 1,037 to 218,214 procedures. Performance was assessed according to discriminative ability as quantified by the c-index (time-dependent area under the receiver operating curve [AUCt]), calibration (slope and intercept), and overall prediction error (scaled Brier score).

6.1.3 Results

The AUCt of the models ranged between 0.52 to 0.68. On average, the differences between the validated performance of different modeling approaches were 0.00 (range -0.04 to 0.03) across 11 data sets.

6.1.4 Conclusions

ML did not outperform traditional regression models. Current predictor variables are insufficient for estimating the risk of revision following arthroplasty surgery either with ML survival- or traditional regression methods. Future

registry efforts should aim at collecting more relevant predictors to improve prediction for individual patients planned for a procedure.

6.1.5 Level of evidence

Prognostic level III

6.2 INTRODUCTION

Various predictive modeling tools have been developed and are used for decision support in healthcare to inform patient and surgeon decision-making. In the field of orthopedic surgery, studies have been designed predicting arthroplasty revision surgery, using competing risk analysis [1-13]. Revision is a procedure that may involve a partial or complete exchange of the prosthesis implanted during the primary -index- surgery. In a classical survival setting, patients only fail from one cause. However, the cumulative incidence of revision (primary outcome) depends not only on the effect of covariates (i.e., age or gender) but also on the survival rate, since patients who have died cannot subsequently undergo revision. Standard survival analyses (Kaplan Meier curves) treat death simply as censored information, but this approach may overestimate revision rates [14]. Therefore, a competing risk analysis should be performed with revision as the primary outcome event and death as competing risk.

New techniques like machine learning (ML) algorithms and the increasing availability of electronic health record data as well as healthcare registries provide new opportunities to improve decision support tools. However, it is unclear whether ML generates better risk estimates than the traditional approach. Some preliminary evidence exists investigating this question. For example, a recent study from our group compared ML and logistic regression algorithms for the prediction of binary events (e.g., reoperation yes or no) in orthopedic trauma in 9 datasets. ML's benefit was shown to be limited [15]. In fields outside of orthopedic surgery, studies have explored ML survival analysis with competing risks: random survival forests (RSFs) [16, 17], a decision tree-based ML algorithm for time-to-event analysis. However, to our knowledge, no study to date has compared competing risk survival models based on ML and traditional regression methods in multiple datasets.

This study aimed to compare the performance of ML survival analysis and traditional regression modeling in a competing risk setting. We hereto analyzed 11 datasets including patients undergoing arthroplasty surgery registered in the Dutch Arthroplasty Register (LROI).

6.3 MATERIALS AND METHODS

6.3.1 Guidelines

This study was conducted according to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research and the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis (TRIPOD) guidelines [18, 19].

6.3.2 Study design and participants (Data sources)

Eligible datasets were derived from previously published studies, including patients registered in the Dutch Arthroplasty Register (LROI) [20] and undergoing a (partial) knee or hip arthroplasty surgery. The overall data completeness for both primary knee and hip arthroplasties was 96% in 2014 and up to 100% in 2020 [20]. In total, 11 datasets were included in the study. All were observational cohort studies that previously reported on predictors of outcome or survival following partial or total knee and hip arthroplasty (Table 6.1) [1-5, 7-11, 21]. The sample size of these datasets ranged from 1,037 to 218,214 procedures. The raw datasets supplied by the LROI were directly derived from the previous studies and contained several processing steps. This resulted in different patients and variables being available across the different datasets. We therefore chose to perform the ML versus traditional statistics comparison in each dataset separately with the same inclusion criteria and set of predictors as applied by the original studies. This also allowed for a direct comparison with the results from the original studies. The baseline characteristics of the 11 included datasets can be found in the original studies [1-5, 7-11, 21]

Table 6.1 Study characteristics of the included studies

Authors	Study title	Methodology	No of patients	Outcome	Predictors included	Study period
Peters et al. (2020)	Patient Characteristics Influence Revision Rate of THA; ASA and BMI Were the Strongest Predictors for Short-Term Revision After Primary THA	LR	218,214	Survival; 1y and 3y revision	Age, gender, ASA, previous operation, smoking, BMI, Charnley	2007-2018
Peters et al. (2018)	The effect of bearing type on the outcome of Total Hip Arthroplasty.	CPH	209,912	Survival; 5y and 9y revision	Age, gender, ASA, diagnosis, previous operation, fixation, head diameter, surgical approach, and period of surgery	2007-2016
van Steenbergen et al. (2020)	Dutch advice not to use large head metal-on-metal hip arthroplasties justifiable – results from the Dutch Arthroplasty Register	CPH	211,002	Survival; 8y revision	age, gender, ASA score, diagnosis (OA vs non-OA), period	2007-2016
van Oost et al. (2020)	Higher risk of revision for partial knee replacements in low absolute volume hospitals: data from 18,134 partial knee replacements in the Dutch Arthroplasty Register	Kaplan Meier, CPH	18,134	Survival	Age category, sex, ASA, year, diagnosis, unicondylar side, type of hospital	2007-2016

Table 6.1 Study characteristics of the included studies (continued)

Authors	Study title	Methodology	No of patients	Outcome	Predictors included	Study period
Burger et al. (2020)	A comprehensive evaluation of lateral uni-compartmental knee arthroplasty short to mid-term survivorship, and the effect of patient and implant characteristics: an analysis of data from Dutch Arthroplasty Register	Kaplan Meier, CPH	19,832	Survival; 5y revision	Age, gender, diagnosis, prior operation, bearing type, and fixation type	2007-2017
Kuijpers et al. (2019)	The risk of revision after total hip arthroplasty in young patients depends on surgical approach, femoral head size and bearing type; an analysis of 19,682 operations in the Dutch arthroplasty register	Kaplan Meier, CPH	19,682	Survival; 5y revision	Age, gender, diagnosis, ASA, surgical approach, fixation, bearing type, head size and year	2007-2017
Bloemheuvel et al. (2019)	Lower 5-year cup re-revision rate for dual mobility cups compared with unipolar cups: report of 15,922 cup revision cases in the Dutch Arthroplasty Register (2007–2016)	Kaplan Meier, CPH	15,922	Survival; 5y re-revision	Gender, age, ASA, fixation	2007-2016

Table 6.1 Study characteristics of the included studies (continued)

Authors	Study title	Methodology	No of patients	Outcome	Predictors included	Study period
Bloemheuvel et al. (2018)	Dual mobility cups in primary total hip arthroplasties: trend over time in use, patient characteristics, and mid-term revision in 3,038 cases in the Dutch Arthroplasty Register (2007–2016)	Kaplan Meier, CPH	3,038	Survival; 5y cup revision	Gender, age, diagnosis, previous operation, ASA, fixation, surgical approach, and femoral head diameter	2007-2016
Spekenbrink et al. (2018)	Higher mid-term revision rates of posterior stabilized compared with cruciate retaining total knee arthroplasties: 133,841 cemented arthroplasties for osteoarthritis in the Netherlands in 2007–2016	CPH	133,841	Survival; 8y revision	Age, gender, ASA, and previous operations	2007-2016
Moerman et al. (2018)	Hemiarthroplasty and total hip arthroplasty in 30,830 patients with hip fractures: data from the Dutch Arthroplasty Register on revision and risk factors for revision	CPH	30,830	Survival; 1y revision	Gender, age, ASA, smoking BMI, approach, and stem fixation	2007-2017
Janssen et al. (2018)	Do Stem Design and Surgical Approach Influence Early Aseptic Loosening in Cementless THA?	CPH	63,354	Survival	Age, sex, diagnosis, ASA, earlier surgeries, and coating and material of stem	2007-2013

NO = number; LR = Logistic Regression; CPH = Cox Proportional Hazards; y = year; ASA = American Society of Anesthesiologists classification; BMI = body mass index; OA = osteoarthritis

6.3.3 Data analysis – Traditional survival approaches

Of the included studies, one study conducted a multivariable logistic regression analysis [1], five applied Kaplan Meier analysis [4, 5, 7, 8, 21], and 10 used multivariable Cox proportional hazard regression analyses [2-5, 7-11, 21]. None of these methods accounted for competing risks.

Logistic Regression Analysis

The study applying logistic regression analysis investigated the differences in revision rates (1- and 3-year) between case-mix subgroups were investigated [1].

Kaplan-Meier Analysis

Kaplan-Meier survival analyses were performed [4, 5, 7, 8, 21] to determine the probability of not experiencing a revision after a specific period of time (surviving) in which the log-rank test (Mantel-Co, 95% CI) was used to compare two groups (e.g., men and women). Kaplan-Meier analysis may overestimate the probability of the event of interest (i.e., revision surgery) [14, 22].

Cox Proportional Hazard Regression Analysis

Cox regression analyses were common in previously published studies [2-5, 7-11, 21], where death was censored. Absolute risk was calculated per time window (e.g., 3-year), and covariates were presented with hazard ratios with a 95% confidence interval (CI).

6.3.4 Data analysis – Survival approaches accounting for competing risks

On the included studies, we developed a set of time-to-event models with revision as the event of interest and death as the competing risk for all included studies separately.

Aalen-Johansen curves

An Aalen-Johansen estimator is a non-parametric estimation of risks, like the Kaplan-Meier estimator in the survival setting (see above). The Aalen-Johansen curve is plotting the CIF of the event of interest (revision) accounting for a competing risk (death) [23].

Fine and Gray Model
A Fine and Gray model [24] is a semi-parametric method (proportional hazards model), estimating the incidence of the outcome of interest (revision) over time in the presence over a competing risk (death), thereby relating covariates to the CIF of the event of interest (revision) [25].

Cause-specific Cox Model
A cause-specific Cox model also is a semi-parametric method. It is an extension of the earlier described Cox regression analyses. In the cause-specific Cox model, the risk of revision is compared among patients who are event free and in follow-up (i.e., patients who have not experienced a revision or the competing risk (death) at a particular time point) [22, 26].

Random Survival Forest
The Random survival forest (RSF) [27] was introduced as a time-to-event extension to a random forest that can account for competing risks. RSF is a machine learning method that uses ensemble learning on many decision trees. It can work with high dimensional and complex (also nonlinear) data.

6.3.5 Data preparation
A set of predictors was identified based on the original variable selection of the included studies (Table 6.1) [26]. The observations for which age or gender were missing were removed from the analysis. All other missing data was imputed using multivariate imputation by chained equations [28] creating 10 imputed datasets.

6.3.6 Model development
For each of the 11 datasets, we plotted the CIF for the outcome of interests (revision) and the competing risk (death) in Aalen-Johansen curves [23]. Subsequently, we compared the predictive performance of two state-of-the-art statistical time-to-event models: a Fine and Gray model and a cause-specific Cox model. These were compared to a ML approach consisting of a random survival forest with competing risks [27].

The time-to-event was set at 1-, 2- and 3-year follow-up for each cohort. The imputed data was split into a train (2/3 of the data) and a test set. This approach

was chosen over more sophisticated train designs (e.g., nested cross-validation) for computational feasibility. The hyperparameters for the random survival forest were set via 5-fold cross-validation on the train data (supplementary Table S1). The models were trained on the train data (with tuned hyperparameters) and applied to the test data.

6.3.7 Model performance

Model performance was evaluated following recent guidance for prediction models in the presence of competing risks [29] that includes: (1) discrimination with a time-dependent area under the receiver operating curve (AUC_t), (2) calibration with the calibration slope and intercept (in line with the method by Cox [30]) and (3) the overall prediction error with the scaled version of the Brier score [29].

The c-index (AUC) ranges from 0.50 to 1.0, with 1.0 indicating the highest discrimination score and 0.50 indicating the lowest. The higher the discrimination score, the better the model's ability to distinguish patients who had the outcome (i.e., patients who received revision from those who had not) [31]. The time-dependent c-index (AUC_t) can be calculated for a single time point of interest (e.g., two-year revision) [29, 32].

A calibration plot plots the primary outcome's estimated and observed probabilities. A perfect calibration plot has an intercept of 0 (<0 reflects overestimation, >0 reflects underestimating the probability of the outcome) and a slope of 1 (the model is performing similarly in training and test sets) [33, 34]. In a small dataset, the slope is often <1, reflecting model overfitting; probabilities are too extreme (low probability too low, high probability too high) [26].

The Brier score calculates a composite of discrimination and calibration, with 0 indicating perfect prediction and a Brier score of 1 the poorest prediction [31]. A scaled version of the Brier score (1-(model Brier score / null model Brier score)) can be interpreted as the amount of prediction error in a null model that the prediction model explains. A 100% scaled Brier score corresponds to a perfect model, 0% to an ineffective model, and <0% to a harmful model [29].

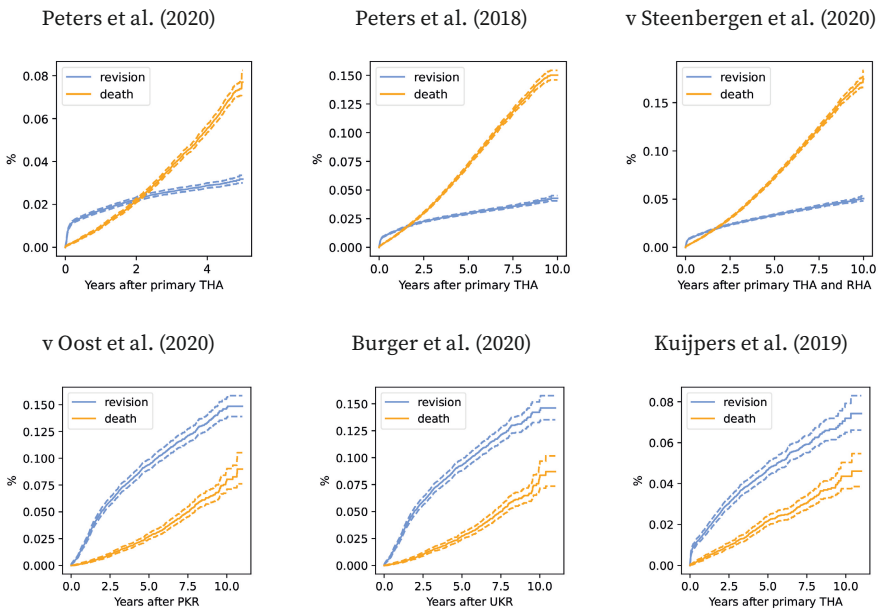
Model performance estimates were pooled across the 10 imputed datasets via Rubin's Rules [35]. We visualized model performance comparison in a beeswarm plot – a scatterplot of the differences in AUCt of each ML and traditional regression pair.

6.3.8 Software

Data pre-processing and analysis was performed using R Version 5.3 ("R: A Language and Environment for Statistical Computing" The R Foundation, Vienna, Austria 2013), R- studio Version 1.2.1335 (R-Studio, Boston, MA, USA) and Python 3.10. The following packages were used: caret, cmprsk, geepack, Hmisc, modelr, prodlim, randomForestSRC, riskRegression, survival, tidyr, tidyverse, and beeswarm. We used the following packages for Python (version 3.7.7): pandas, numpy, matplotlib, lifelines, sksurv, and sklearn.

6.4 RESULTS

The CIFs were plotted for all 11 datasets (Figure 6.1). For most datasets, the absolute risk of death surpasses the risk of revision at some point in time, which concurs with the population generally studied.



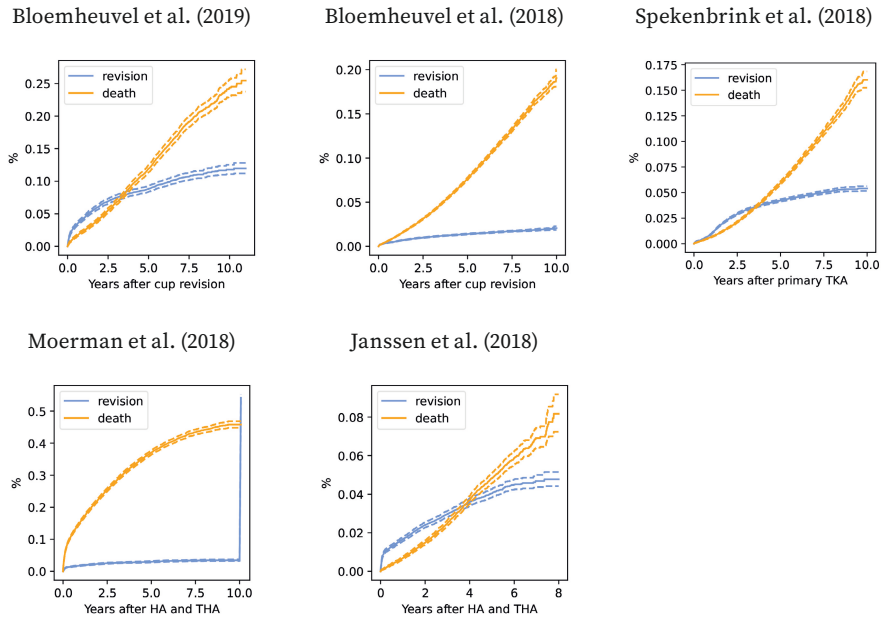


Figure 6.1 Cumulative Incidence Function for all datasets

Abbreviations: *THA* Total Hip Arthroplasty, *RHA* Resurfacing Hip Arthroplasty, *PKR* Partial Knee Replacement, *UKR* Unicompartimental Knee Arthroplasty, *TKA* Total Knee Arthroplasty, *HA* hemiarthroplasty

Next, we compared the predictive performance of a ML approach to two state-of-the-art statistical time-to-event models in 11 datasets. The difference in AUCt for each analysis was on average 0.00 for traditional regression compared to ML, with a range from -0.04 to 0.03, indicating that ML and traditional regression models produce similar probability estimates (Figure 6.2).

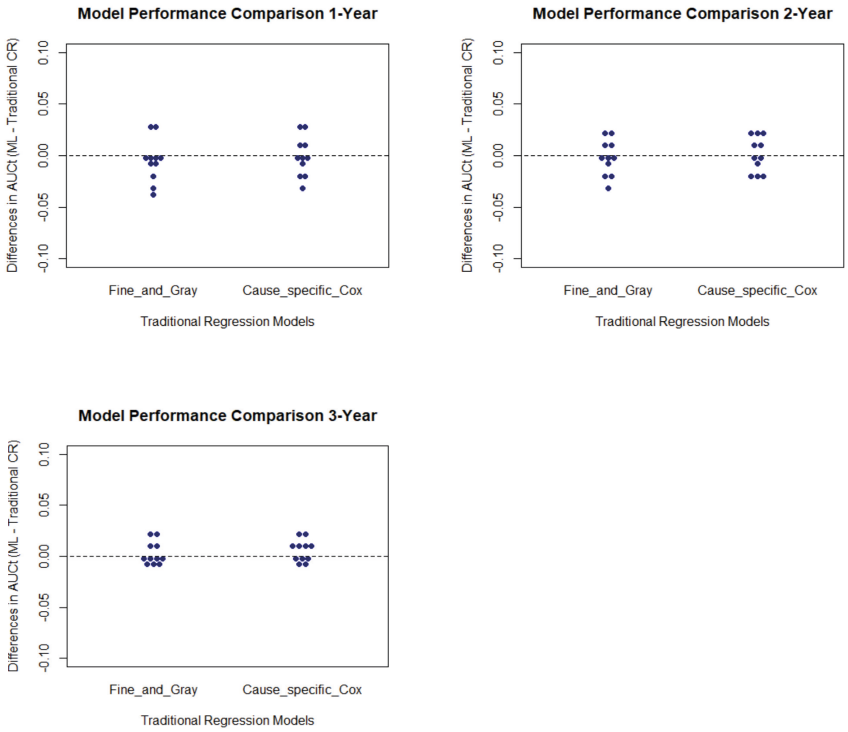


Figure 6.2 Beeswarm plots of model performance time-dependent c-index differences (ΔML – Traditional Regression).

Abbreviations: *ML* Machine Learning, *CR* Competing Risk

Table 6.2 Time-dependent AUC (95% confidence interval)

Dataset	Time	Fine and Gray	Cause-specific Cox	Competing Risk Survival Forest
Peters et al. (2020)	1	0.57 (0.55,0.6)	0.57 (0.55,0.6)	0.56 (0.54,0.59)
	2	0.58 (0.56,0.6)	0.58 (0.56,0.6)	0.56 (0.54,0.59)
	3	0.58 (0.56,0.6)	0.58 (0.56,0.6)	0.56 (0.54,0.59)
Peters et al. (2018)	1	0.61 (0.59,0.63)	0.61 (0.59,0.63)	0.61 (0.59,0.63)
	2	0.62 (0.61,0.64)	0.63 (0.61,0.64)	0.63 (0.61,0.64)
	3	0.66 (0.64,0.67)	0.66 (0.64,0.67)	0.65 (0.64,0.67)
van Steenbergen et al. (2020)	1	0.54 (0.53,0.56)	0.55 (0.53,0.57)	0.57 (0.55,0.59)
	2	0.55 (0.54,0.57)	0.56 (0.54,0.58)	0.57 (0.55,0.59)
	3	0.57 (0.55,0.58)	0.57 (0.56,0.59)	0.58 (0.56,0.59)
van Oost et al. (2020)	1	0.54 (0.49,0.58)	0.54 (0.49,0.58)	0.53 (0.48,0.58)

Table 6.2 Time-dependent AUC (95% confidence interval) (continued)

Dataset	Time	Fine and Gray	Cause-specific Cox	Competing Risk Survival Forest
	2	0.55 (0.52,0.58)	0.55 (0.52,0.59)	0.55 (0.52,0.58)
	3	0.55 (0.52,0.57)	0.55 (0.52,0.58)	0.54 (0.51,0.57)
Burger et al. (2020)	1	0.51 (0.45,0.56)	0.51 (0.46,0.57)	0.52 (0.47,0.58)
	2	0.53 (0.49,0.57)	0.53 (0.49,0.57)	0.53 (0.49,0.58)
	3	0.54 (0.5,0.57)	0.54 (0.5,0.58)	0.55 (0.51,0.58)
Kuijpers et al. (2019)	1	0.59 (0.53,0.65)	0.59 (0.53,0.65)	0.61 (0.54,0.67)
	2	0.58 (0.53,0.63)	0.58 (0.53,0.63)	0.58 (0.53,0.63)
	3	0.58 (0.53,0.62)	0.58 (0.53,0.62)	0.58 (0.53,0.62)
Bloemheuvel et al. (2019)	1	0.5 (0.46,0.53)	0.5 (0.46,0.54)	0.52 (0.49,0.56)
	2	0.53 (0.5,0.56)	0.53 (0.5,0.56)	0.54 (0.51,0.57)
	3	0.54 (0.51,0.57)	0.54 (0.51,0.57)	0.55 (0.52,0.58)
Bloemheuvel et al. (2018)	1	0.56 (0.52,0.59)	0.56 (0.53,0.59)	0.58 (0.54,0.61)
	2	0.57 (0.54,0.6)	0.57 (0.54,0.6)	0.57 (0.54,0.59)
	3	0.57 (0.54,0.59)	0.57 (0.54,0.59)	0.56 (0.54,0.59)
Spekenbrink et al. (2018)	1	0.57 (0.54,0.6)	0.57 (0.54,0.6)	0.58 (0.55,0.62)
	2	0.6 (0.58,0.62)	0.6 (0.58,0.62)	0.6 (0.58,0.62)
	3	0.62 (0.6,0.63)	0.62 (0.6,0.63)	0.61 (0.59,0.63)
Moerman et al. (2018)	1	0.61 (0.56,0.65)	0.61 (0.56,0.65)	0.62 (0.57,0.66)
	2	0.61 (0.57,0.65)	0.61 (0.57,0.65)	0.62 (0.58,0.67)
	3	0.61 (0.58,0.65)	0.61 (0.58,0.65)	0.62 (0.59,0.66)
Janssen et al. (2018)	1	0.52 (0.49,0.56)	0.52 (0.49,0.56)	0.53 (0.49,0.57)
	2	0.52 (0.49,0.55)	0.52 (0.49,0.55)	0.53 (0.5,0.56)
	3	0.5 (0.47,0.53)	0.5 (0.47,0.53)	0.51 (0.48,0.55)

Abbreviations: *AUC* Area Under the Curve

The AUC_t of the Fine and Gray models ranged between 0.52 to 0.66, the cause-specific Cox ranged from 0.53 to 0.66, and the random survival forest ranged from 0.51 to 0.65 (Table 6.2).

The calibration metrics (Table 6.3) and scaled Brier scores (Table 6.4) also produced comparable estimates, showing no advantage of ML over traditional regression models.

Table 6.3 Calibration intercept and slope (95% confidence interval)

Dataset	Time	Fine and Gray		Cause-specific Cox	
		Intercept	Slope	Intercept	Slope
Peters et al. (2020)	1	-0.04 (-0.12,0.05)	0.79 (0.51,1.07)	-0.04 (-0.12,0.05)	0.79 (0.51,1.07)
	2	-0.06 (-0.14,0.02)	0.86 (0.6,1.12)	-0.06 (-0.14,0.02)	0.86 (0.6,1.12)
	3	-0.08 (-0.15,0)	0.92 (0.65,1.2)	-0.08 (-0.15,0)	0.92 (0.65,1.19)
Peters et al. (2018)	1	-0.01 (-0.08,0.05)	1.02 (0.8,1.23)	-0.01 (-0.08,0.05)	1.03 (0.82,1.25)
	2	-0.04 (-0.09,0.02)	0.93 (0.74,1.12)	-0.04 (-0.1,0.02)	0.94 (0.75,1.12)
	3	-0.07 (-0.12,-0.02)	0.89 (0.72,1.07)	-0.07 (-0.13,-0.02)	0.89 (0.72,1.06)
van Steenbergen et al. (2020)	1	-0.04 (-0.1,0.03)	0.61 (0.34,0.87)	-0.03 (-0.09,0.03)	0.72 (0.44,0.99)
	2	-0.04 (-0.1,0.01)	0.72 (0.48,0.95)	-0.04 (-0.1,0.01)	0.77 (0.54,1)
	3	-0.02 (-0.08,0.03)	0.87 (0.66,1.09)	-0.03 (-0.08,0.03)	0.89 (0.67,1.1)
van Oost et al. (2020)	1	0.05 (-0.12,0.21)	0.24 (-0.19,0.67)	0.05 (-0.11,0.22)	0.25 (-0.2,0.69)
	2	0.11 (-0.01,0.22)	0.57 (0.26,0.89)	0.11 (0,0.23)	0.59 (0.27,0.91)
	3	0.14 (0.04,0.24)	0.6 (0.32,0.88)	0.14 (0.04,0.24)	0.62 (0.33,0.9)
Burger et al. (2020)	1	-0.2 (-0.38,-0.02)	0.19 (-0.41,0.78)	-0.19 (-0.37,-0.01)	0.19 (-0.42,0.8)
	2	-0.1 (-0.23,0.02)	0.46 (0.05,0.87)	-0.1 (-0.23,0.03)	0.47 (0.05,0.88)
	3	-0.06 (-0.18,0.06)	0.66 (0.28,1.03)	-0.06 (-0.17,0.06)	0.67 (0.29,1.04)
Kuijpers et al. (2019)	1	-0.26 (-0.47,-0.04)	0.61 (0.1,1.11)	-0.25 (-0.47,-0.03)	0.63 (0.13,1.14)
	2	-0.18 (-0.36,-0.01)	0.55 (0.14,0.96)	-0.18 (-0.35,-0.01)	0.56 (0.15,0.97)
	3	-0.13 (-0.29,0.02)	0.66 (0.25,1.07)	-0.13 (-0.29,0.02)	0.66 (0.25,1.07)
Bloemheuvel et al. (2019)	1	-0.11 (-0.24,0.03)	-0.09 (-0.55,0.38)	-0.09 (-0.23,0.04)	-0.06 (-0.56,0.43)
	2	-0.09 (-0.21,0.02)	0.34 (-0.07,0.75)	-0.08 (-0.2,0.03)	0.37 (-0.05,0.8)
	3	-0.04 (-0.14,0.07)	0.56 (0.18,0.94)	-0.03 (-0.14,0.07)	0.59 (0.2,0.99)

Table 6.3 Calibration intercept and slope (95% confidence interval) (continued)

Dataset	Time	Fine and Gray		Cause-specific Cox	
		Intercept	Slope	Intercept	Slope
Bloemheuvel et al. (2018)	1	-0.17 (-0.28,-0.05)	0.92 (0.54,1.3)	-0.16 (-0.27,-0.05)	1 (0.61,1.38)
	2	-0.14 (-0.23,-0.04)	1.1 (0.76,1.44)	-0.14 (-0.23,-0.04)	1.14 (0.8,1.47)
	3	-0.09 (-0.18,-0.01)	1.08 (0.79,1.38)	-0.09 (-0.18,-0.01)	1.1 (0.8,1.39)
Spekenbrink et al. (2018)	1	-0.12 (-0.23,-0.01)	0.6 (0.34,0.87)	-0.11 (-0.22,0)	0.62 (0.35,0.89)
	2	-0.09 (-0.17,-0.02)	0.88 (0.7,1.06)	-0.09 (-0.16,-0.02)	0.88 (0.7,1.07)
	3	-0.13 (-0.19,-0.06)	0.9 (0.74,1.05)	-0.13 (-0.19,-0.06)	0.9 (0.74,1.06)
Moerman et al. (2018)	1	-0.02 (-0.18,0.14)	0.65 (0.38,0.92)	0 (-0.17,0.16)	0.7 (0.41,0.98)
	2	0 (-0.14,0.15)	0.64 (0.4,0.87)	0.01 (-0.13,0.16)	0.66 (0.41,0.9)
	3	-0.02 (-0.16,0.13)	0.6 (0.38,0.83)	-0.01 (-0.15,0.13)	0.62 (0.39,0.85)
Janssen et al. (2018)	1	-0.01 (-0.14,0.11)	0.35 (-0.24,0.94)	-0.01 (-0.14,0.11)	0.4 (-0.18,0.99)
	2	-0.03 (-0.13,0.08)	0.38 (-0.11,0.86)	-0.03 (-0.13,0.08)	0.41 (-0.08,0.89)
	3	-0.06 (-0.16,0.04)	0.36 (-0.12,0.85)	-0.05 (-0.15,0.05)	0.38 (-0.1,0.86)

Table 6.3 continued Calibration intercept and slope (95% confidence interval)

Dataset	Time	Competing Risk Survival Forest	
		Intercept	Slope
Peters et al. (2020)	1	-0.03 (-0.12,0.07)	0.77 (0.43,1.11)
	2	-0.07 (-0.15,0.01)	0.77 (0.4,1.13)
	3	-0.09 (-0.17,-0.01)	0.77 (0.45,1.09)
Peters et al. (2018)	1	0.04 (-0.03,0.11)	1.4 (1.1,1.71)
	2	0.02 (-0.04,0.08)	1.34 (1.06,1.62)
	3	-0.01 (-0.06,0.04)	1.25 (0.99,1.52)
van Steenbergen et al. (2020)	1	0.05 (-0.02,0.12)	0.84 (0.52,1.16)
	2	0 (-0.06,0.06)	0.86 (0.55,1.16)
	3	0.01 (-0.04,0.07)	0.94 (0.63,1.26)
van Oost et al. (2020)	1	0.19 (0.02,0.37)	0.44 (-0.26,1.15)
	2	0.19 (0.07,0.31)	0.85 (0.41,1.28)
	3	0.18 (0.07,0.28)	0.89 (0.5,1.28)
Burger et al. (2020)	1	-0.07 (-0.26,0.11)	0.37 (-0.23,0.97)
	2	-0.08 (-0.21,0.06)	0.5 (0.1,0.91)
	3	-0.02 (-0.14,0.1)	0.68 (0.27,1.08)
Kuijpers et al. (2019)	1	-0.17 (-0.39,0.05)	1.33 (0.26,2.41)
	2	-0.11 (-0.28,0.06)	0.94 (0.18,1.7)
	3	-0.09 (-0.24,0.06)	0.95 (0.31,1.59)
Bloemheuvel et al. (2019)	1	-0.03 (-0.17,0.11)	0.16 (-0.34,0.67)
	2	-0.07 (-0.19,0.05)	0.38 (-0.05,0.8)
	3	-0.04 (-0.15,0.07)	0.51 (0.14,0.88)
Bloemheuvel et al. (2018)	1	-0.14 (-0.26,-0.01)	1.11 (0.51,1.7)
	2	-0.13 (-0.24,-0.03)	1.02 (0.65,1.39)
	3	-0.09 (-0.19,0)	1 (0.68,1.32)
Spekenbrink et al. (2018)	1	0.14 (0.02,0.26)	1.09 (0.6,1.59)
	2	0.03 (-0.05,0.1)	1.28 (0.97,1.58)
	3	-0.06 (-0.13,0.01)	1.15 (0.91,1.39)
Moerman et al. (2018)	1	0 (-0.18,0.19)	0.7 (0.4,1)
	2	0 (-0.17,0.16)	0.64 (0.39,0.89)
	3	-0.02 (-0.18,0.13)	0.61 (0.38,0.83)
Janssen et al. (2018)	1	0 (-0.13,0.13)	0.45 (-0.04,0.94)
	2	-0.02 (-0.13,0.09)	0.44 (0.02,0.86)
	3	-0.05 (-0.15,0.05)	0.46 (0.02,0.91)

Table 6.4 Scaled Brier score (95% confidence interval)

Dataset	Time	Fine and Gray	Cause-specific Cox	Competing Risk Survival Forest
Peters et al. (2020)	1	0.001 (0,0.002)	0.001 (0,0.002)	0.001 (0,0.002)
	2	0.001 (0,0.003)	0.001 (0,0.003)	0.001 (-0.001,0.002)
	3	0.002 (0.001,0.004)	0.002 (0.001,0.004)	0.001 (-0.001,0.002)
Peters et al. (2018)	1	0.002 (0.001,0.003)	0.002 (0.001,0.003)	0.002 (0.001,0.003)
	2	0.004 (0.003,0.005)	0.004 (0.003,0.005)	0.003 (0.002,0.004)
	3	0.006 (0.005,0.007)	0.007 (0.006,0.007)	0.004 (0.004,0.005)
van Steenbergen et al. (2020)	1	0 (0,0.001)	0 (0,0.001)	0.001 (0,0.001)
	2	0.001 (0,0.001)	0.001 (0,0.002)	0.001 (0.001,0.002)
	3	0.001 (0.001,0.002)	0.002 (0.001,0.002)	0.002 (0.001,0.002)
van Oost et al. (2020)	1	-0.002 (-0.005,0.001)	-0.002 (-0.004,0.001)	-0.001 (-0.004,0.001)
	2	-0.001 (-0.005,0.003)	-0.001 (-0.005,0.003)	0 (-0.004,0.003)
	3	-0.003 (-0.008,0.003)	-0.002 (-0.008,0.003)	-0.001 (-0.005,0.004)
Burger et al. (2020)	1	-0.003 (-0.007,0.001)	-0.003 (-0.007,0.001)	-0.001 (-0.004,0.001)
	2	-0.002 (-0.007,0.002)	-0.002 (-0.007,0.002)	-0.002 (-0.007,0.003)
	3	-0.001 (-0.006,0.005)	-0.001 (-0.006,0.005)	0 (-0.005,0.005)
Kuijpers et al. (2019)	1	-0.001 (-0.005,0.003)	-0.001 (-0.005,0.003)	0.001 (-0.002,0.005)
	2	-0.002 (-0.007,0.003)	-0.002 (-0.007,0.003)	0 (-0.003,0.004)
	3	-0.002 (-0.008,0.005)	-0.002 (-0.008,0.005)	0.001 (-0.003,0.005)
Bloemheuvel et al. (2019)	1	-0.005 (-0.009,-0.002)	-0.004 (-0.008,-0.001)	-0.003 (-0.006,0.001)
	2	-0.003 (-0.008,0.001)	-0.002 (-0.007,0.002)	-0.002 (-0.007,0.003)
	3	-0.001 (-0.006,0.004)	0 (-0.005,0.004)	-0.001 (-0.006,0.004)
Bloemheuvel et al. (2018)	1	0 (-0.001,0.001)	0 (0,0.001)	0 (0,0.001)
	2	0.001 (0,0.001)	0.001 (0,0.001)	0 (0,0.001)
	3	0.001 (0,0.002)	0.001 (0,0.002)	0.001 (0,0.001)
Spekenbrink et al. (2018)	1	0 (-0.001,0.001)	0 (-0.001,0.001)	0.001 (0,0.001)
	2	0.003 (0.001,0.005)	0.003 (0.001,0.005)	0.003 (0.002,0.004)
	3	0.005 (0.002,0.007)	0.005 (0.002,0.007)	0.004 (0.003,0.006)
Moerman et al. (2018)	1	0.002 (-0.002,0.005)	0.002 (-0.001,0.006)	0.003 (-0.001,0.007)
	2	0.003 (-0.001,0.006)	0.003 (-0.001,0.007)	0.004 (-0.001,0.008)
	3	0.003 (-0.002,0.007)	0.003 (-0.001,0.007)	0.004 (-0.001,0.009)
Janssen et al. (2018)	1	0 (-0.001,0)	0 (-0.001,0.001)	0 (-0.001,0.001)
	2	-0.001 (-0.002,0)	-0.001 (-0.002,0)	-0.001 (-0.002,0.001)
	3	-0.001 (-0.003,0)	-0.001 (-0.003,0)	-0.001 (-0.003,0.001)

Abbreviations: *AUC* Area Under the Curve

6.5 DISCUSSION

In this comparative study, we found that ML performed similarly to traditional regression methods. Moreover, current predictor variables are insufficient for estimating the risk of revision for patients undergoing arthroplasty surgery either with ML or conventional traditional regression methods.

6.5.1 Strengths and limitations

An important strength of our study is that we included multiple datasets from previously peer-reviewed published studies, enhancing the reliability of our findings. To the best of our knowledge, this is the first study comparing ML and conventional traditional regression methods with a competing risk in multiple datasets.

The results of this study should also be viewed considering several limitations. First, the data was derived from the Dutch Arthroplasty Registry (LROI) [20] and may not be generalized to an international population. The findings of this study may be more reliable when evaluated on independent registry cohorts. Second, we chose a common set of time-to-event points for a true comparison of model performances across the included datasets. A visual trend was seen where the ML performances were increasing more over time compared to traditional regression model performances (Figure 6.2). Future studies can evaluate longer time-to-event points for individual studies investigating the benefit of ML survival analysis with a competing risk. Third, hyperparameter tuning was carried out on the train data set. We did not carry out nested cross-validation due to the current computation time for training a RSF model. However, we did not expect to have an incremental benefit in model performance in our cohorts with the use of more sophisticated nested cross-validation. Lastly, this was a retrospective study beholden to limitations inherent to such a research design. Future prospective research efforts to predict revision following arthroplasty surgery, should aim at collecting a higher number of relevant predictors per individual patient.

6.5.2 Previous literature

Our findings were comparable to Aram and colleagues [16] evaluating various model approaches for accurate risk estimation in patients undergoing revi-

sion surgery after knee arthroplasty. Their results showed that a fully parametric model (i.e., RSF) is essential for revision prediction; however, their study concluded that such methods did not provide high discriminatory power at the individual level either. Martin and colleagues [36] aimed to predict revision surgery following hip arthroscopy, including different model approaches (e.g., RSF), concluding limited clinical usefulness.

The finding that ML and traditional regression methods were comparable is consistent with a previous study from our group, comparing ML and logistic regression algorithms for predicting binary outcome in Orthopaedic trauma in 9 datasets. In other fields, a study expected ML analysis to outperform Cox proportional hazard regression analysis in breast cancer survival [37]. However again, RSF showed a similar performance to traditional regression analysis, and the ML algorithms outperforming traditional regression analysis did not account for a competing risk.

6.5.3 Implications

These findings have implications for future research aiming to improve decision support tools in the presence of competing risks. First, the observation that ML models are comparable with traditional models in the presence of competing risks suggests that their benefit may be limited in this context. Our findings highlight the importance of not overly relying on ML methods as the 'holy grail' in prediction modelling and questioning the benefit of ML models for low dimensional datasets.

Second, the modelling approaches presented here are insufficient to predict the risk of revision following knee- or hip arthroplasty. The low revision rate ranging between 0.5% to 4.6% may have limited the models' ability to distinguish between procedures with and without a revision in the current study context [38]. Predicting revision in arthroplasty procedures will likely remain challenging for this reason. Imbalance correction techniques could be applied prior to training the models in the future, but this comes at the cost of strong miscalibration [39]. Future research may investigate the comparison between ML and traditional regression methods for other outcomes, such as patient-reported outcome measures (PROMs), and evaluating patients' satisfaction after arthroplasty surgery [40, 41].

Moreover, the registry data at present may not be discriminative enough. Globally, arthroplasty registries are broadening their data collection regarding PROMs and social determinants of health [42]. Future patient-centered strategies can focus on evaluating such measures and their influence on improving decision support tools.

6.5.4 Conclusions

Current predictor variables are insufficient to accurately predict the risk of revision following arthroplasty surgery either with ML or traditional regression approach. Developing prediction models for estimating the risk of revision surgery in patients undergoing arthroplasty surgery offers challenges due to the censored nature of data and the current data availability. Future registry efforts should aim at collecting more relevant predictors for the benefit of individual patients.

REFERENCES

1. Peters, R.M., L.N. van Steenbergen, R.E. Stewart, et al., *Patient Characteristics Influence Revision Rate of Total Hip Arthroplasty: American Society of Anesthesiologists Score and Body Mass Index Were the Strongest Predictors for Short-Term Revision After Primary Total Hip Arthroplasty.* J Arthroplasty, 2020. **35**(1): p. 188-192.e2.
2. Peters, R.M., L.N. Van Steenbergen, M. Stevens, P.C. Rijk, S.K. Bulstra, and W.P. Zijlstra, *The effect of bearing type on the outcome of total hip arthroplasty.* Acta Orthop, 2018. **89**(2): p. 163-169.
3. van Steenbergen, L.N., G.A. Denissen, B.W. Schreurs, W.P. Zijlstra, H.W. Koot, and R.G. Nelissen, *Dutch advice not to use large head metal-on-metal hip arthroplasties justifiable–results from the Dutch Arthroplasty Register.* Nederlands Tijdschrift voor Orthopaedie, 2020. **27**(1).
4. van Oost, I., K.L.M. Koenraadt, L.N. van Steenbergen, S.B.T. Bolder, and R.C.I. van Geenen, *Higher risk of revision for partial knee replacements in low absolute volume hospitals: data from 18,134 partial knee replacements in the Dutch Arthroplasty Register.* Acta Orthop, 2020. **91**(4): p. 426-432.
5. Burger, J.A., L.J. Kleeblad, I.N. Sierevelt, et al., *A Comprehensive Evaluation of Lateral Unicompartmental Knee Arthroplasty Short to Mid-Term Survivorship, and the Effect of Patient and Implant Characteristics: An Analysis of Data From the Dutch Arthroplasty Register.* J Arthroplasty, 2020. **35**(7): p. 1813-1818.
6. Kuijpers, M.F.L., G. Hannink, L.N. van Steenbergen, and B.W. Schreurs, *Outcome of revision hip arthroplasty in patients younger than 55 years: an analysis of 1,037 revisions in the Dutch Arthroplasty Register.* Acta Orthop, 2020. **91**(2): p. 165-170.
7. Kuijpers, M.F.L., G. Hannink, S.B.W. Vehmeijer, L.N. van Steenbergen, and B.W. Schreurs, *The risk of revision after total hip arthroplasty in young patients depends on surgical approach, femoral head size and bearing type; an analysis of 19,682 operations in the Dutch arthroplasty register.* BMC Musculoskelet Disord, 2019. **20**(1): p. 385.
8. Bloemheuvel, E.M., L.N.V. Steenbergen, and B.A. Swierstra, *Lower 5-year cup re-revision rate for dual mobility cups compared with unipolar cups: report of 15,922 cup revision cases in the Dutch Arthroplasty Register (2007-2016).* Acta Orthop, 2019. **90**(4): p. 338-341.
9. Spekenbrink-Spooren, A., L.N. Van Steenbergen, G.A.W. Denissen, B.A. Swierstra, R.W. Poolman, and R. Nelissen, *Higher mid-term revision rates of posterior stabilized compared with cruciate retaining total knee arthroplasties: 133,841 cemented arthroplasties for osteoarthritis in the Netherlands in 2007-2016.* Acta Orthop, 2018. **89**(6): p. 640-645.
10. Moerman, S., N.M.C. Mathijssen, W.E. Tuinebreijer, A.J.H. Vochteloo, and R. Nelissen, *Hemiarthroplasty and total hip arthroplasty in 30,830 patients with hip fractures: data from the Dutch Arthroplasty Register on revision and risk factors for revision.* Acta Orthop, 2018. **89**(5): p. 509-514.
11. Janssen, L., K.A.P. Wijnands, D. Janssen, M. Janssen, and J.W. Morrenhof, *Do Stem Design and Surgical Approach Influence Early Aseptic Loosening in Cementless THA?* Clin Orthop Relat Res, 2018. **476**(6): p. 1212-1220.

12. Zijlstra, W.P., B. De Hartog, L.N. Van Steenbergen, B.W. Scheurs, and R. Nelissen, *Effect of femoral head size and surgical approach on risk of revision for dislocation after total hip arthroplasty.* Acta Orthop, 2017. **88**(4): p. 395-401.
13. Peters, R.M., L.N. van Steenbergen, S.K. Bulstra, et al., *Nationwide review of mixed and non-mixed components from different manufacturers in total hip arthroplasty: a Dutch Arthroplasty Register study.* Acta Orthopaedica, 2016. **87**(4): p. 356-362.
14. Keurentjes, J.C., M. Fiocco, B.W. Schreurs, B.G. Pijls, K.A. Nouta, and R.G.H.H. Nelissen, *Revision surgery is overestimated in hip replacement.* Bone & Joint Research, 2012. **1**(10): p. 258-262.
15. Oosterhoff, J.H.F., B.Y. Gravesteijn, A.V. Karhade, et al., *Feasibility of Machine Learning and Logistic Regression Algorithms to Predict Outcome in Orthopaedic Trauma Surgery.* J Bone Joint Surg Am, 2022. **104**(6): p. 544-551.
16. Aram, P., L. Trela-Larsen, A. Sayers, et al., *Estimating an Individual's Probability of Revision Surgery After Knee Replacement: A Comparison of Modeling Approaches Using a National Data Set.* Am J Epidemiol, 2018. **187**(10): p. 2252-2262.
17. Pickett, K.L., K. Suresh, K.R. Campbell, S. Davis, and E. Juarez-Colunga, *Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker.* BMC Medical Research Methodology, 2021. **21**(1): p. 216.
18. Luo, W., D. Phung, T. Tran, et al., *Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View.* J Med Internet Res, 2016. **18**(12): p. e323.
19. Collins, G.S., J.B. Reitsma, D.G. Altman, and K.G.M. Moons, *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement.* Eur Urol, 2015. **67**(6): p. 1142-1151.
20. Dutch Arthroplasty Register. *LROI Report 2022*. Available from: https://www.lroi-report.nl/app/uploads/2022/04/PDF-LROI-annual-report-2021.pdf.
21. Bloemheuvel, E.M., L.N. van Steenbergen, and B.A. Swierstra, *Dual mobility cups in primary total hip arthroplasties: trend over time in use, patient characteristics, and mid-term revision in 3,038 cases in the Dutch Arthroplasty Register (2007-2016).* Acta Orthop, 2019. **90**(1): p. 11-14.
22. Putter, H., M. Fiocco, and R.B. Geskus, *Tutorial in biostatistics: competing risks and multi-state models.* Statistics in Medicine, 2007. **26**(11): p. 2389-2430.
23. Aalen, O.O. and S. Johansen, *An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations.* Scandinavian Journal of Statistics, 1978. **5**(3): p. 141-150.
24. Fine, J.P. and R.J. Gray, *A Proportional Hazards Model for the Subdistribution of a Competing Risk.* Journal of the American Statistical Association, 1999. **94**(446): p. 496-509.
25. Austin, P.C., E.W. Steyerberg, and H. Putter, *Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1.* Stat Med, 2021. **40**(19): p. 4200-4212.
26. Van Der Pas, S., R. Nelissen, and M. Fiocco, *Different competing risks models for different questions may give similar results in arthroplasty registers in the presence of few events.* Acta Orthop, 2018. **89**(2): p. 145-151.
27. Ishwaran, H., T.A. Gerds, U.B. Kogalur, R.D. Moore, S.J. Gange, and B.M. Lau, *Random survival forests for competing risks.* Biostatistics, 2014. **15**(4): p. 757-773.

28. van Buuren, S. and K. Groothuis-Oudshoorn, *mice: Multivariate Imputation by Chained Equations in R.* Journal of Statistical Software, 2011. **45**(3): p. 1 - 67.
29. van Geloven, N., D. Giardiello, E.F. Bonneville, et al., *Validation of prediction models in the presence of competing risks: a guide through modern methods.* Bmj, 2022. **377**: p. e069249.
30. Cox, D.R., *Two further applications of a model for binary regression.* Biometrika, 1958. **45**(3/4): p. 562-565.
31. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation.* Eur Heart J, 2014. **35**(29): p. 1925-31.
32. Gerds, T.A., M.W. Kattan, M. Schumacher, and C. Yu, *Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring.* Stat Med, 2013. **32**(13): p. 2173-84.
33. Steyerberg, E.W., A.J. Vickers, N.R. Cook, et al., *Assessing the performance of prediction models: a framework for some traditional and novel measures.* Epidemiology (Cambridge, Mass.), 2010. **21**(1): p. 128.
34. Van Calster, B. and A.J. Vickers, *Calibration of risk prediction models: impact on decision-analytic performance.* Med Decis Making, 2015. **35**(2): p. 162-9.
35. Rubin, D.B., *Multiple imputation for nonresponse in surveys*. Vol. 81. 2004: John Wiley & Sons.
36. Martin, R.K., S. Wastvedt, J. Lange, A. Pareek, J. Wolfson, and B. Lund, *Limited clinical utility of a machine learning revision prediction model based on a national hip arthroscopy registry.* Knee Surgery, Sports Traumatology, Arthroscopy, 2022: p. 1-11.
37. Moncada-Torres, A., M.C. van Maaren, M.P. Hendriks, S. Siesling, and G. Geleijnse, *Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival.* Scientific reports, 2021. **11**(1): p. 6968.
38. Labek, G., M. Thaler, W. Janda, M. Agreiter, and B. Stöckl, *Revision rates after total joint replacement: cumulative results from worldwide joint register datasets.* The Journal of bone and joint surgery. British volume, 2011. **93**(3): p. 293-297.
39. van den Goorbergh, R., M. van Smeden, D. Timmerman, and B. Van Calster, *The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression.* Journal of the American Medical Informatics Association, 2022. **29**(9): p. 1525-1534.
40. Sorel, J., E. Veltman, A. Honig, and R. Poolman, *The influence of preoperative psychological distress on pain and function after total knee arthroplasty: a systematic review and meta-analysis.* Bone Joint J, 2019. **101**(1): p. 7-14.
41. Peters, R.M., L.N. van Steenbergen, R.E. Stewart, et al., *Which patients improve most after total hip arthroplasty? Influence of patient characteristics on patient-reported outcome measures of 22,357 total hip arthroplasties in the Dutch Arthroplasty Register.* Hip international, 2021. **31**(5): p. 593-602.
42. Rolfson, O., E. Bohm, P. Franklin, et al., *Patient-reported outcome measures in arthroplasty registries: report of the Patient-Reported Outcome Measures Working Group of the International Society of Arthroplasty Registries Part II. Recommendations for selection, administration, and analysis.* Acta orthopaedica, 2016. **87**(sup1): p. 9-23.

SUPPLEMENTARY MATERIAL

Table S1 Hyper parameters Random Survival Forest

Hyperparameters	Number of trees	Maximum number of end nodes	Maximum depth
	[50,100,200]	[5,10,15]	[5,10,15]
Peters et al. (2020)	50	10	5
Peters et al. (2018)	200	15	5
van Steenbergen et al. (2020)	50	5	5
van Oost et al. (2020)	100	10	5
Burger et al. (2020)	100	5	5
Kuijpers et al. (2019)	200	5	5
Bloemheuvel et al. (2019)	200	5	5
Bloemheuvel et al. (2018)	100	10	5
Spekenbrink et al. (2018)	50	15	5
Moerman et al. (2018)	50	5	5
Janssen et al. (2018)	200	15	15