A Comparison of Survival Models for Prediction of Eight-year Revision Risk Following Total Knee and Hip Arthroplasty

Alana R Cuthbert (✉ alana.cuthbert@sahmri.com)

University of South Australia

Lynne C Giles

The University of Adelaide

Gary Glonek

The University of Adelaide

Lisa M Kalisch Ellett

University of South Australia

Nicole L Pratt

University of South Australia

Research Article

Keywords: Prediction model, machine learning, time-to-event data, flexible parametric survival model, parametric survival model, random survival forest, hip replacement, knee replacement

Posted Date: August 13th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-778992/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

There is increasing interest in the development and use of clinical prediction models, but a lack of evidence-supported guidance on the merits of different modelling approaches. This is especially true for time-to-event outcomes, where limited studies have compared the vast number of modelling approaches available. This study compares prediction accuracy and variable importance measures for four modelling approaches in prediction of time-to-revision surgery following total knee arthroplasty (TKA) and total hip arthroplasty (THA).

Methods

The study included 321 945 TKA and 151 113 THA procedures performed between 1 January 2003 and 31 December 2017. Accuracy of the Cox model, Weibull parametric model, flexible parametric model, and random survival forest were compared, with patient age, sex, comorbidities, and prosthesis characteristics considered as predictors. Prediction accuracy was assessed using the Index of Prediction Accuracy (IPA), c-index, and smoothed calibration curves. Variable importance rankings from the Cox model and random survival forest were also compared.

Results

Overall, the Cox and flexible parametric survival models performed best for prediction of both TKA (integrated IPA 0.056 (95% CI [0.054, 0.057]) compared to 0.054 (95% CI [0.053, 0.056]) for the Weibull parametric model), and THA revision. (0.029 95% CI [0.027, 0.030] compared to 0.027 (95% CI [0.025, 0.028]) for the random survival forest). The c-index showed broadly similar discrimination between all modelling approaches. Models were generally well calibrated, but random survival forest underfitted the predicted risk of TKA revision compared to regression approaches. The most important predictors of revision were similar in the Cox model and random survival forest for TKA (age, opioid use, and patella resurfacing) and THA (femoral cement, depression, and opioid use).

Conclusion

The Cox and flexible parametric models had superior overall performance, although all approaches performed similarly. Notably, this study showed no benefit of a tuned random survival forest over regression models in this setting.

Background

There is increasing interest in the development and use of clinical prediction models⁽¹⁾. Accurate prediction models can assist in informed decision-making by estimating a patient's risk of a health outcome based on their individual characteristics, rather than relying on crude population-level estimates. However, developing an accurate prediction model requires the researcher to choose from many available

modelling approaches and limited studies have compared the advantages and disadvantages of each method.

For time-to-event outcomes the Cox model is the most common approach, but parametric survival models, including flexible parametric models, may be preferable depending on the complexity of the data. Alternatively, machine learning methods hold promise of improved prediction accuracy through automatic modelling of non-linearities, interactions, and time-varying effects in predictor variables. These methods make fewer (or no) assumptions about the underlying structure of the data(2), but as a consequence they are generally less efficient and require much larger sample sizes to obtain stable predictions(3). Another drawback of machine learning methods is lack of interpretability; understanding which variables are important for prediction and how they influence the outcome are critical to the utility of such models(4). Many machine learning methods provide measures of how 'important' each variable, but do not indicate effect size or direction. A raft of machine learning methods to assist in the development of prediction models are now available, but few studies have systematically compared their performance to traditional regression approaches(5).

An important clinical area for the development of prediction models is joint replacement surgery. Arthroplasty of the hip or knee hip is an effective treatment for end stage osteoarthritis, an increasingly common disease and one of the leading causes of global disability(6). While joint replacements are expected to last at least 25 years on average, a small proportion will fail within a shorter time frame and require revision surgery(7, 8). Premature revision surgery is a major burden for both patients and the healthcare system, resulting in worse outcomes for patients and billions in hospital costs(9, 10). Improved prediction of the risk of revision, by taking into account patient-, surgeon- and prosthesis-related factors, will better inform patients of their likely risks when undergoing elective surgery, as well as enable hospitals to predict expected health care burden. Prospective joint replacement recipients are concerned with both their risk of revision surgery and the ways in which their personal characteristics influence this risk, highlighting the need for prediction models that are both accurate and interpretable(11).

In the present study, four survival modelling approaches for predicting time-to-revision within eight years of joint arthroplasty surgery are compared: Cox regression, parametric regression with a Weibull distribution, flexible parametric regression, and random survival forests. Variable importance rankings from the Cox model and random survival forests are also compared.

Methods

Data Source

This study used data from elective primary Total Knee Arthroplasty (TKA) and Total conventional Hip Arthroplasty (THA) procedures recorded in the Australian Orthopaedic Association National Joint Replacement Registry (AOANJRR) between 1 July 2003 and 31 December 2017. The Registry collects data on patient age, sex, indication for surgery, and prosthesis type and features. Patient comorbidities

were identified through record linkage with the Pharmaceutical Benefits Scheme administrative claims database. This database is maintained by the Australian Government Department of Human Services and contains information on the dispensing of prescription medicines. A total of 47 morbid conditions were identified using the validated Rx-Risk coding of patient prescriptions(12). A patient was considered to have a morbid condition if they were dispensed at least one medicine indicative of that condition in the 12 months prior to their joint replacement surgery.

The AOANJRR captured approximately 98% of knee and hip arthroplasties in Australia over the study period. Using probabilistic data linkage, 95% of the procedures were linked to Pharmaceutical Benefits Scheme data. Revision procedures were identified through internal record linkage and patient death was identified through record linkage with the National Death Index, enabling near-complete follow-up of all procedures.

Participants and Variables

The inclusion criteria for TKAs were: primary indication of osteoarthritis, patients aged 45 to 89 years at the time of surgery; minimally and posterior-stabilised prostheses only, and no missing prosthesis attributes.

The inclusion criteria for THAs were: primary indication of osteoarthritis, patients aged 40 to 89 years at the time of surgery; modern bearings (metal-on-cross-linked polyethylene, ceramic-on-cross-linked polyethylene, and ceramic-on-ceramic), and no missing prosthesis attributes.

In addition, only patients receiving concessional benefits were included. These patients are eligible healthcare cardholders or pensioners who pay a lower co-payment towards the cost of medicines subsidised by the Australian Government and represented 80% of the total joint replacement population. Flowcharts showing details of inclusion criteria are provided in Additional File 1.

Once selection criteria were applied, 321 945 TKA and 151 113 THA procedures were available for model development and validation. A summary of the study population is given in Table 1 for TKAs and Table 2 for THAs.

Table 1
Demographics, comorbid conditions and prosthesis use in patients undergoing TKA, stratified by outcome

Variable	Revised N = 10470	All TKA N = 321945
Median (IQR) age in years	69 (64–75)	72 (67–78)
N (%) female	6052 (57.8)	195668 (60.8)
N (%) cemented femoral	5772 (55.1)	188734 (58.6)
N (%) cemented tibial	7843 (74.9)	259187 (80.5)
N (%) patella used	4162 (39.8)	164333 (51)
N (%) fixed bearing	7701 (73.6)	257008 (79.8)
N (%) minimally stabilised	7135 (68.1)	234827 (72.9)
N (%) cross-linked polyethylene	2115 (20.2)	107456 (33.4)
N (%) computer navigated	1589 (15.2)	60164 (18.7)
N (%) anticoagulants	1338 (12.8)	42398 (13.2)
N (%) antiplatelet medications	2210 (21.1)	64011 (19.9)
N (%) anxiety	1541 (14.7)	34863 (10.8)
N (%) Arrhythmia	595 (5.7)	18373 (5.7)
N (%) congestive heart failure	864 (8.3)	25281 (7.9)
N (%) depression	3341 (31.9)	83445 (25.9)
N (%) diabetes	1509 (14.4)	47041 (14.6)
N (%) Gastro-oesophageal reflux disease	5368 (51.3)	156820 (48.7)
N (%) glaucoma	562 (5.4)	20382 (6.3)
N (%) gout	1047 (10)	30540 (9.5)
N (%) hyperlipidaemia	5070 (48.4)	160478 (49.8)
N (%) hypertension	5649 (54)	181127 (56.3)
N (%) hypothyroidism	1102 (10.5)	33546 (10.4)
N (%) ischaemic heart disease (angina)	867 (8.3)	21925 (6.8)
N (%) ischaemic heart disease (hypertension)	3586 (34.3)	117801 (36.6)
N (%) osteoporosis/Paget's	889 (8.5)	30706 (9.5)
N (%) pain	5861 (56)	151805 (47.2)

Variable	Revised N = 10470	All TKA N = 321945
N (%) inflammation pain	6430 (61.4)	176883 (54.9)
N (%) chronic airways disease	2498 (23.9)	71122 (22.1)
N (%) steroid responsive	2006 (19.2)	54271 (16.9)

Table 2
Demographics, comorbid conditions and prosthesis use in patients undergoing THA,
stratified by outcome

Variable	Revised N = 4748	All THA N = 151113
Median (IQR) age in years	73 (67–78)	73 (68–79)
N (%) female	2692 (56.7)	90153 (59.7)
N (%) cemented femoral	1660 (35)	66537 (44)
N (%) cemented acetabular	102 (2.1)	4231 (2.8)
N (%) bearing surface		
Ceramic/ceramic	1134 (23.9)	28886 (19.1)
Ceramic/cross-linked polyethylene	728 (15.3)	27249 (18)
Metal/cross-linked polyethylene	2886 (60.8)	94978 (62.9)
N (%) head size		
≤28mm	1870 (39.4)	29724 (19.7)
32mm	1623 (34.2)	64854 (42.9)
36mm	1151 (24.2)	53376 (35.3)
≥40mm	104 (2.2)	3159 (2.1)
N (%) anticoagulants	698 (14.7)	20990 (13.9)
N (%) antiplatelet medications	961 (20.2)	27645 (18.3)
N (%) anxiety	672 (14.2)	16146 (10.7)
N (%) arrhythmia	308 (6.5)	9092 (6)
N (%) congestive heart failure	424 (8.9)	11708 (7.7)
N (%) depression	1471 (31)	37159 (24.6)
N (%) diabetes	512 (10.8)	16689 (11)
N (%) Gastro-oesophageal reflux disease	2314 (48.7)	65911 (43.6)
N (%) glaucoma	300 (6.3)	9824 (6.5)
N (%) gout	405 (8.5)	11325 (7.5)
N (%) hyperlipidaemia	2110 (44.4)	70365 (46.6)
N (%) hypertension	2418 (50.9)	78101 (51.7)
N (%) hypothyroidism	467 (9.8)	14282 (9.5)

Variable	Revised N = 4748	All THA N = 151113
N (%) ischaemic heart disease (angina)	334 (7)	9576 (6.3)
N (%) ischaemic heart disease (hypertension)	1613 (34)	50495 (33.4)
N (%) osteoporosis/Paget's	530 (11.2)	15157 (10)
N (%) pain	2902 (61.1)	82787 (54.8)
N (%) inflammation pain	2962 (62.4)	83078 (55)
N (%) chronic airways disease	1022 (21.5)	29633 (19.6)
N (%) steroid responsive	860 (18.1)	22730 (15)

Modelling Approaches

Cox regression is the most popular method for predicting time-to-event outcomes, largely due to simplicity, as the distributional form of the baseline hazard does not need to be specified. Parametric survival models are an alternative approach in which the baseline hazard is assumed to follow a particular distribution, allowing for survival estimates at continuous time points and extrapolation. The Weibull distribution on a proportional hazards scale, while commonly used in health data, may be inappropriate for complex hazards that are not monotonically increasing or decreasing. To address this, the Royston-Parmar flexible parametric model was proposed(13). The baseline hazard is modelled using a restricted cubic spline, allowing flexibility in the functional form of the baseline log cumulative hazard. This hybrid approach combines the flexibility of the Cox model with the desirable properties of parametric survival models. Despite the benefits of flexible parametric models, they are rarely used for prognostic models in medical settings and have not been systematically compared to other survival analysis approaches(14).

Both the Cox and parametric models on the proportional hazards scale assume that the effect of covariates is constant over time, which is rarely a realistic assumption in medical settings. Parametric and semi-parametric approaches also require interactions between variables to be explicitly specified, which may be intractable when the number of predictors is large. In contrast, the random survival forest algorithm is a fully non-parametric machine learning approach that does not assume proportional hazards and can automatically account for possible interaction effects(15). Introduced in 2008, it is an extension of the random forest algorithm that makes predictions for new patients by aggregating predicted survival curves from a series of survival trees. Random survival forests can also provide fully nonparametric measures of variable importance(16).

Statistical Analysis

The outcome of interest was time-to-first revision within eight years of primary joint arthroplasty. In the Cox and parametric models, age was treated as a continuous variable and modelled with a restricted cubic spline with four knots. All other variables included in the prediction models were categorical. Patient

death was treated as a censoring event as ignoring competing risks has been shown to have a negligible impact on patient-level predictions in this setting(17). If a patient had bilateral TKAs or THAs, each side was treated as a separate unilateral procedure, which has been shown to have a negligible effect on model estimates(18). Seven knots were used for modelling the log cumulative hazard in the flexible parametric model, with knots placed at the default location of equally spaced quantiles of the log uncensored survival times.

Random survival forests were grown using log-rank splitting with 300 trees. Two parameters were tuned: the terminal node size and the number of variables considered for splits when growing the survival trees. Full details of the tuning process are provided in Additional File 3.

Model Performance

The eight-year prediction performance of the four modelling approaches was averaged across 10 repetitions of 10-fold cross-validation. 95% confidence intervals for performance metrics were calculated by computing a standard normalised interval around the mean using the different values estimated within each fold. Normality was assessed using quantile-quantile plots and found to be a reasonable assumption for all performance metrics.

Model discrimination was assessed using Harrell's concordance index (c-index). The c-index estimates the probability that, for a randomly selected pair of patients, the patient with highest predicted risk fails first. The value of the c-index ranges from 0.5 to 1, with a value of 1 implying perfect discrimination and 0.5 representing a model that is no better than random guessing.

Calibration was assessed using smoothed calibration curves to compare the proportion of observed and predicted events at eight years(19). A calibration curve that closely follows the 45 degree identity line indicates a good match between predicted and observed values. The smoothed plots were generated using a Cox model with predicted probabilities modelled using a restricted cubic spline with four knots. Calibration was assessed for predicted probabilities ranging from the first percentile to the 99th percentile. A numeric summary of the calibration curve, the Integrated Calibration Index (ICI), was also calculated, with lower values implying a smaller average difference between the observed and predicted probabilities(20).

The overall performance of each model was assessed using the Index of Prediction Accuracy (IPA), derived as $1 - (\text{model Brier score} / \text{null model Brier score})$, where the null model is the Kaplan-Meier estimator(21). The Brier score measures the average squared distance between the observed event status and predicted event probability for each individual at a single point in time, thereby providing a combined measure of discrimination and calibration. Higher IPA values imply better model fit, with 100% representing a perfect model. Values ≤ 0 indicate the model performs no better than the population-level estimate. The IPA was integrated over eight years to summarise model performance in a single numeric value, as well as calculated at several time points and presented graphically to show the predictive performance of the modelling approaches over the eight-year period.

The c-index and IPA were weighted using inverse probability of censoring to correct for bias introduced by censoring (22–24).

Variable Importance

Two methods for determining the most important predictors of revision risk were compared: backwards elimination in the Cox model and minimal depth from the random survival forest.

Minimal depth in the random survival forest is defined as the shortest distance between the variable and the root node of the tree(16). A small minimal depth implies that the variable was chosen early in the splitting process, which implies the variable has a strong influence in determining the rates of revision for joint replacement. Minimal depth for each variable was averaged across 500 trees grown from a tuned random survival forest. Backwards elimination in the Cox model was performed with no stopping criterion. The order in which predictors were sequentially removed from the model was used to rank their importance. This process was repeated on 500 bootstrap samples of the data as variable selection from backward elimination is notoriously unstable(25). Ranks were averaged across bootstrap samples and 95% confidence intervals for ranks were calculated assuming a normal distribution. This rank-based approach was used to allow more direct comparison to the minimal depth from the random survival forest. Backwards elimination was not performed for the parametric regression approaches, as model coefficients were nearly identical to the Cox model (as shown in Additional File 2).

Statistical analyses were performed using R version 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria) with packages survival (26), pec (27), riskRegression (28), flexsurv(29), rms(30), and randomForestSRC(31).

Results

For TKAs, the revision rate was 4.1% at eight years and the hazard function was non-monotonic; revision risk was high initially after surgery, decreased, then spiked approximately one year after surgery before decreasing again (Fig. 1a and 1b). For THAs, the revision rate was 3.9% at eight years and the hazard function monotonically decreased over time, with the risk of revision highest immediately after surgery (Fig. 1c and 1d).

For both TKA and THA, the discrimination of the four modelling approaches was virtually identical (c-index 0.64 for all four approaches for TKA revision and 0.59 for all four approaches for THA revision). (Table 3, Table 4).

For TKAs the random survival forest had worse calibration than the Cox, Weibull or flexible parametric models. The ICI showed that on average, predicted risks from the Cox model differed from actual risk by 0.16% (95% CI [0.15, 0.18]), but this difference was 0.27% (95% CI [0.25,0.29]) for the random survival forest (Table 3). The Cox, Weibull, and flexible parametric models were well calibrated across the range of possible risks, whereas the random survival forest overestimated the risk for lower risk patients and underestimated the risk for higher risk patients (Fig. 2a).

For THA, all modelling approaches had similar overall calibration according to the ICI (Table 4). The Cox, Weibull and flexible parametric models were well calibrated for low-risk patients but overestimated the revision risk for higher risk patients. Conversely, the random survival forests were well calibrated in those with high risk but underestimated the risk for lower risk patients (Fig. 2b).

When predicting TKA revision, the Cox and flexible parametric models returned the highest integrated IPA, each with a value of 0.056 (95% CI [0.054, 0.057]) while the Weibull model had the lowest IPA of 0.054 (95% CI [0.053, 0.056]) (Table 3). All models performed similarly in the later follow-up period, with the Weibull and random survival forest slightly worse (Fig. 3a). Within the first year of TKA, the random survival forest was the best performing approach for prediction of revision. In this earlier time period, the Weibull model had negative IPA, implying it performed worse than the null model.

When predicting THA revision, the Cox and flexible parametric models had the highest integrated IPA (0.029 95% CI [0.027, 0.030] compared to 0.027 (95% CI [0.025, 0.028]) for the random survival forest) (Table 4). The random survival forest had the highest IPA for revisions within the first two years but showed poorer performance than the other modelling approaches for later time periods. The Weibull model had slightly worse performance than the Cox and flexible parametric models over the entire eight-year period (Fig. 3b).

Table 3

Performance metrics for predicting revision of TKA using Cox, Weibull parametric, flexible parametric, and random survival forest models

TKA			
Modelling Approach	c-index	Integrated Index of Prediction Accuracy	Integrated calibration index (×100)
Cox	0.643 (0.641, 0.645)	0.056 (0.054, 0.057)	0.16 (0.15, 0.18)
Weibull	0.642 (0.641, 0.644)	0.054 (0.053, 0.056)	0.17 (0.16, 0.19)
Flexible parametric	0.643 (0.641, 0.645)	0.056 (0.054, 0.057)	0.17 (0.15, 0.18)
Random survival forest	0.643 (0.642, 0.645)	0.055 (0.054, 0.056)	0.27 (0.25, 0.29)

Table 4

Performance metrics for predicting revision of THA using Cox, Weibull parametric, flexible parametric, and random survival forest models

THA			
	c-index	Index of Prediction Accuracy	Integrated calibration index (×100)
Cox	0.591 (0.589, 0.594)	0.029 (0.027, 0.03)	0.27 (0.25, 0.3)
Weibull	0.591 (0.588, 0.594)	0.028 (0.026, 0.029)	0.29 (0.26, 0.31)
Flexible parametric	0.591 (0.588, 0.594)	0.029 (0.027, 0.030)	0.28 (0.25, 0.3)
Random survival forest	0.59 (0.587, 0.592)	0.027 (0.025, 0.028)	0.28 (0.25, 0.3)

For TKA, rankings of variable importance from backwards elimination and random survival forest minimal depth identified the same three most important predictors of revision (age, use of pain medication (opioids), and use of patella resurfacing). Both selection methods ranked prosthesis stability, prosthesis bearing surface and patient depression as the three next most important predictors, with the order differing between methods (Table 5).

For THA, rankings of variable importance from backwards elimination and random survival forest minimal depth identified the same five most important predictors of revision in the same order: femoral cement, patient depression, use of pain medication (opioids), gastro-oesophageal reflux disease, and sex. Patient age and steroid responsive diseases were the next most important predictors, with the ordering swapped between methods (Table 6).

Table 5

Ranked importance of variables from backward elimination in Cox model compared to minimal depth from random survival forest, for prediction of TKA revision. Variables are displayed in decreasing order of importance.

Bootstrap backwards elimination in Cox model (95% CI)		Minimal depth from random survival forest (95% CI)	
Age	1 (1,1)	Age	1.54 (1.45,1.64)
Pain	2.26 (2.21,2.3)	Pain	1.91 (1.8,2.02)
Patella usage	2.92 (2.86,2.97)	Patella usage	2.04 (1.93,2.15)
Stability	4.39 (4.34,4.44)	Depression	2.2 (2.08,2.32)
Bearing surface	4.51 (4.43,4.59)	Bearing surface	2.46 (2.34,2.58)
Depression	6.22 (6.17,6.28)	Stability	2.79 (2.68,2.9)
Sex	7.39 (7.32,7.46)	Mobility	3.05 (2.91,3.18)
Tibial cement	8.8 (8.62,8.98)	Anxiety	3.05 (2.91,3.19)
Anxiety	10.6 (10.44,10.77)	Sex	3.24 (3.13,3.35)
Ischaemic heart disease angina	10.68 (10.48,10.87)	Tibial cement	3.49 (3.36,3.62)
Mobility	10.77 (10.6,10.95)	Inflammation pain	4.2 (4.07,4.33)
Gastro-oesophageal reflux disease	11.02 (10.81,11.23)	Ischaemic heart disease angina	4.2 (4.09,4.32)
Steroid responsive diseases	12.48 (12.27,12.7)	Steroid responsive diseases	4.23 (4.11,4.35)
Computer navigation	16.48 (16.22,16.74)	Gastro-oesophageal reflux disease	4.28 (4.15,4.41)
congestive heart failure	16.89 (16.5,17.28)	Hypertension	4.47 (4.35,4.6)
Arrhythmia	17.1 (16.74,17.46)	Ischaemic heart disease hypertension	4.68 (4.56,4.8)
Ischaemic heart disease hypertension	18.62 (18.29,18.95)	Computer navigation	4.73 (4.62,4.83)
Hypothyroidism	19.16 (18.82,19.5)	congestive heart failure	4.74 (4.63,4.84)
Hypertension	19.39 (19.02,19.76)	Gout	4.74 (4.64,4.84)
Osteoporosis/Paget's	19.88 (19.53,20.23)	Chronic Airways Disease	4.78 (4.66,4.9)
Anticoagulants	21.74 (21.37,22.12)	Femoral cement	4.86 (4.74,4.98)
Inflammation pain	22.09 (21.76,22.42)	Glaucoma	4.93 (4.81,5.05)

Bootstrap backwards elimination in Cox model (95% CI)		Minimal depth from random survival forest (95% CI)	
Chronic Airways Disease	22.96 (22.61,23.31)	Anticoagulants	4.93 (4.82,5.04)
Diabetes	23.9 (23.58,24.22)	Osteoporosis/Paget's	5.1 (4.99,5.21)
Hyperlipidaemia	24.12 (23.82,24.41)	Arrhythmia	5.11 (5,5.21)
Antiplatelet medication	24.28 (23.96,24.6)	Diabetes	5.17 (5.07,5.27)
Gout	24.86 (24.59,25.14)	Antiplatelet medication	5.18 (5.07,5.29)
Glaucoma	25.23 (24.95,25.52)	Hypothyroidism	5.19 (5.1,5.29)
Femoral cement	25.24 (24.97,25.51)	Hyperlipidaemia	5.26 (5.16,5.36)

Table 6

Ranked importance of variables from backward elimination in Cox model compared to minimal depth from random survival forest, for prediction of THA revision. Variables are displayed in decreasing order of importance.

Bootstrap backwards elimination in Cox model (95% CI)		Minimal depth from random survival forest (95% CI)	
Femoral cement	1.46 (1.43,1.49)	Femoral cement	0.66 (0.62,0.7)
Depression	1.92 (1.87,1.98)	Depression	0.99 (0.93,1.04)
Pain	2.82 (2.76,2.89)	Pain	1.5 (1.44,1.55)
Gastro-oesophageal reflux disease	5.33 (5.19,5.46)	Gastro-oesophageal reflux disease	2.44 (2.36,2.51)
Sex	5.58 (5.49,5.67)	Sex	2.7 (2.65,2.76)
Age	7.49 (7.32,7.67)	Steroid responsive diseases	2.77 (2.69,2.85)
Steroid responsive diseases	8.29 (8.06,8.52)	Age	3.1 (3.05,3.15)
Inflammation pain	8.96 (8.79,9.13)	Congestive heart failure	3.25 (3.18,3.32)
Anxiety	9.74 (9.53,9.95)	Anxiety	3.64 (3.55,3.74)
Congestive heart failure	12.22 (11.93,12.5)	Bearing Surface	3.74 (3.67,3.81)
Osteoporosis/Paget's	12.22 (11.98,12.46)	Head size	3.88 (3.83,3.94)
Head size	12.23 (11.97,12.49)	Osteoporosis/Paget's	3.93 (3.84,4.02)
Bearing surface	12.32 (12.1,12.55)	Gout	4.18 (4.1,4.26)
Hyperlipidaemia	12.37 (12.19,12.54)	Inflammation pain	4.32 (4.23,4.41)
Anticoagulant	15.47 (15.18,15.76)	Acetabular cement	4.33 (4.26,4.41)
Hypothyroidism	16.74 (16.48,17)	Hyperlipidaemia	4.6 (4.51,4.69)
Chronic Airways Disease	17.61 (17.34,17.88)	Chronic Airways Disease	4.78 (4.69,4.87)
Gout	19.13 (18.86,19.4)	Anticoagulants	4.95 (4.85,5.04)
Arrhythmia	19.92 (19.64,20.2)	Arrhythmia	5.05 (4.95,5.15)
Ischaemic heart disease hypertension	20 (19.76,20.23)	Hypothyroidism	5.28 (5.19,5.38)
Antiplatelet medications	21.3 (21.09,21.52)	Ischaemic heart disease angina	5.32 (5.22,5.43)
Glaucoma	21.4 (21.2,21.6)	Hypertension	5.48 (5.38,5.58)
Diabetes	21.53 (21.33,21.74)	Ischaemic heart disease hypertension	5.55 (5.46,5.64)

Bootstrap backwards elimination in Cox model (95% CI)		Minimal depth from random survival forest (95% CI)	
Ischaemic heart disease angina	21.62 (21.42,21.82)	Antiplatelet medications	5.57 (5.47,5.67)
Acetabular cement	21.64 (21.45,21.84)	Diabetes	5.74 (5.64,5.85)
Hypertension	21.69 (21.48,21.89)	Glaucoma	6.11 (6,6.21)

Discussion

This study found that the Cox and flexible parametric models outperformed the Weibull parametric model and random survival forest in the prediction of time-to-revision following either THA or TKA.

Unsurprisingly, the flexible parametric model always outperformed the simpler Weibull model, particularly for TKA revision where the hazard function was complex and non-monotonic.

Random survival forests did not outperform carefully constructed regression models, despite being optimised in a large training set. This result is consistent with the findings of a systematic review that found no evidence machine learning provides improved performance over logistic regression in the binary outcome setting(32). However, a review has not yet been conducted for time-to-event outcomes, where machine learning approaches have the additional advantage of not being constrained by the proportional hazards assumption.

A recent review highlighted the need for more studies comparing the prediction accuracy of the Royston-Parmar flexible parametric model to that of the Cox model.(14) Our study demonstrated that the flexible parametric approach had near-identical prediction accuracy to the Cox model. However, our results are in contrast to those of Aram et al., who found the flexible parametric model outperformed both Cox regression and random survival forests for prediction of eight-year revision of TKA(33), albeit with fewer predictors considered in that study.

Our results identified that ranking of important predictors was similar when using backwards elimination in the Cox model and minimal depth in random survival forests, suggesting that the important predictors identified are relatively robust to selection method. Many of the variables important in predicting revision risk were prosthesis-related, rather than patient-related. This was particularly true for TKAs, where patella resurfacing, prosthesis stability and bearing surface were among the six most important revision factors. However, for THA, use of femoral cement and several comorbid conditions were identified as important risk factors of revision. The presence of pain, identified by opioid usage, was predictive of revision risk in both THA and TKA patients, consistent with the association between pre-operative opioid use and increased revision risk documented in other studies(34–38). Depression was also identified as an important risk factor in both TKA and THA revision. In a study of the effect of 26 comorbidities on revision rates, depression was found to have the strongest effect on revision risk(38). Gastro-

oesophageal reflux disease was identified as an important predictor of THA revision, possibly reflecting the association between the use of proton pump inhibitors and increased risk of hip fracture(39, 40).

A limitation of this study was the relatively low prediction accuracy (c-index of 0.64 for TKA and 0.59 for THA), which may have been due to the absence of certain patient and surgical factors from the dataset, such as patient body mass index, frailty, socioeconomic measures, lifestyle factors, and comorbidities not treated with indicative prescription medication. However, the models reported here performed similarly to other prediction models for TKA revision developed using registry data(33, 41), and all models outperformed population level estimates, indicating that there is value in developing a predictive model.

This study also did not consider interaction terms nor time-varying coefficients in the regression models. Flexible parametric survival models can easily incorporate time-varying effects in auxiliary parameters. This could be explored in future research. However, given that we did not see improved performance from the random survival forest, which automatically models interactions and time varying effects, this may indicate that limited performance gain will be realized in this setting. Future work could also compare the performance of other machine learning approaches available for time-to-event data, including support vector machines(42), neural nets(43–45), and gradient boosting(46, 47).

Conclusion

The Cox and flexible parametric models were shown to have superior accuracy for predicting time-to-revision risk following TKA and THA compared to random survival forests. The Cox model and random survival forest also identified similar predictors as being the most important for revision risk. Our findings suggest that random survival forests for risk prediction models in the joint replacement setting offer no benefit over regression approaches in terms of prediction accuracy and give broadly similar conclusions regarding variable importance.

Abbreviations

THA – Total Hip Arthroplasty

TKA – Total Knee Arthroplasty

AOANJRR – Australian Orthopaedic Association National Joint Replacement Registry

CI – confidence interval

IPA – index of prediction accuracy

ICI – integrated calibration index

Declarations

Ethics approval

Ethics approval was granted by the University of South Australia Human Research Ethics Committee (0000035831) and the Australian Institute of Health and Welfare Ethics Committee (EO2016/4/316). All investigations were conducted in accordance with ethical principles of research (the Helsinki Declaration II).

Consent for publication

Not applicable

Availability of data and materials

The data for this study was creating from pre-existing datasets (Australian Orthopaedic Association National Joint Replacement Registry (AOANJRR) and routinely-collected Australian Government Medicare Benefits Schedule and Pharmaceutical Benefits Scheme data), with data custodian permissions and specific ethical approval. The data may potentially be made available to other researchers if they obtain the necessary approvals. Further information on the process is available from the following sources: Population Health and Research Network (<https://www.phrn.org.au/for-researchers/data-access/cross-and-multi-jurisdictional-application-process/>), AOANJRR (<https://aoanjrr.sahmri.com/aoanjrr-data-linkage>, admin@aoanjrr.org.au) and the Australian Institute of Health and Welfare (<https://www.aihw.gov.au/our-services/data-linkage>, linkage@aihw.gov.au).

Competing Interests

The authors declare they have no competing interests

Funding:

This work was supported by National Health and Medical Research Council Project Grant (GNT1148106). AC is supported by an Australian Government Research Training Program Scholarship and scholarship from Quality Use of Medicines and Pharmacy Research Centre, University of South Australia.

Authors' contributions

AC designed the study, performed all data analysis, interpreted the results and drafted the manuscript. LC, GG, LK and NP contributed to the design of the study, interpreted results and critically revised the manuscript. All authors have read and approve the manuscript.

Acknowledgements

Not applicable

References

1. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Medicine*. 2013;10(2):e1001381.
2. Wang P, Tech V, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys*. 2019;51(6):1–43.
3. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC medical research methodology*. 2014;14:137-.
4. Towards trustable machine learning. *Nature Biomedical Engineering*. 2018;2(10):709–10.
5. Boulesteix AL, Binder H, Abrahamowicz M, Sauerbrei W. On the necessity and design of studies comparing statistical methods. *Biometrical Journal*. 2018;60(1):216–8.
6. Cross M, Smith E, Hoy D, Nolte S, Ackerman I, Fransen M, et al. The global burden of hip and knee osteoarthritis: estimates from the Global Burden of Disease 2010 study. *Annals of the Rheumatic Diseases*. 2014;73(7):1323–30.
7. Evans JT, Walker RW, Evans JP, Blom AW, Sayers A, Whitehouse MR. How long does a knee replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. *Lancet*. 2019;393:655–63.
8. Evans JT, Evans JP, Walker RW, Blom AW, Whitehouse MR, Sayers A. How long does a hip replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. *Lancet*. 2019;393(10172):647–54.
9. Bhandari M, Smith J, Miller LE, Block JE. Clinical and economic burden of revision knee arthroplasty. *Clinical medicine insights Arthritis and musculoskeletal disorders*. 2012;5:89–94.
10. Australian Orthopaedic Association National Joint Replacement Registry. Hip, Knee & Shoulder Arthroplasty: 2018 Annual Report. 2018.
11. Zotov E, Hills AF, de Mello FL, Aram P, Sayers A, Blom AW, et al. JointCalc: A web-based personalised patient decision support tool for joint replacement. *Int J Med Inform*. 2020;142:104217-.
12. Pratt NL, Kerr M, Barratt JD, Kemp-Casey A, Kalisch Ellett LM, Ramsay E, et al. The validity of the Rx-Risk Comorbidity Index using medicines mapped to the Anatomical Therapeutic Chemical (ATC) Classification System. *BMJ Open*. 2018;8(4):e021122-e.
13. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*. 2002;21(15):2175–97.

14. Ng R, Kornas K, Sutradhar R, Wodchis WP, Rosella LC. The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review. *Diagnostic and Prognostic Research*. 2018;2(1):4-.
15. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. *The annals of applied statistics*. 2008;2:841–60.
16. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*. 2010;105:205–17.
17. Cuthbert AR, Graves SE, Giles LC, Glonek G, Pratt N. What Is the Effect of Using a Competing-risks Estimator when Predicting Survivorship After Joint Arthroplasty: A Comparison of Approaches to Survivorship Estimation in a Large Registry. *Clinical Orthopaedics and Related Research*. 2021;479:392–403.
18. Lie SA, Engesæter LB, Havelin LI, Gjessing HK, Vollset SE. Dependency issues in survival analyses of 55 782 primary hip replacements from 47 355 patients. *Statistics in Medicine*. 2004;23(20):3227–40.
19. Austin PC, Harrell FE, Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Statistics in Medicine*. 2020;39:2714–42.
20. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*. 2019;38(21):4051–65.
21. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and Prognostic Research*. 2018;2(1):7.
22. Gerds TA, Schumacher M. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*. 2006;48(6):1029–40.
23. Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*. 2013;32(13):2173–84.
24. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*. 1999;18(17–18):2529–45.
25. Harrell FE. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. 2015. ed. Cham: Springer International Publishing; 2015.
26. Therneau TM, Lumley, Thomas, Atkinson, Elizabeth, Crowson, Cynthia. survival: survival analysis. R package version 3.1–12 [Available from: <https://cran.r-project.org/package=survival>].
27. Gerds TA. pec: prediction error curves for risk prediction models in survival analysis. R package version 2019.11.03 [Available from: <https://CRAN.R-project.org/package=pec>].
28. Gerds TA, Blanche P, Mortensen R, Tollenaar N, Moensen UB, Ozenne B. riskRegression: risk regression models and prediction scores for survival analysis with competing risks. R package version 2020.02.05. [Available from: <https://cran.r-project.org/package=riskRegression>].
29. Jackson C, Metcalfe P, Amdahl J. flexsurv: flexible parametric survival and multi-state models. R package version 1.1.1 [Available from: <https://CRAN.R-project.org/package=flexsurv>].

30. Harrell FE. rms: Regression Modeling Strategies. R package version 6.1-1 [Available from: <https://CRAN.R-project.org/package=rms>.
31. Ishwaran H, Kogalur UB. randomForestSRC: fast unified random forests for survival, regression, and classification. R package version 2.9.3 [Available from: <https://CRAN.R-project.org/package=randomForestSRC>.
32. 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. *Journal of Clinical Epidemiology*. 2019;110:12–22.
33. Aram P, Trela-Larsen L, Sayers A, Hills AF, Blom AW, McCloskey EV, et al. Estimating an individual's probability of revision surgery after knee replacement: a comparison of modeling approaches using a national dataset. *American Journal of Epidemiology*. 2018;187:2252–62.
34. Shadbolt C, Schilling C, Inacio MC, Abbott JH, Prymachenko Y, Wilson R, et al. Opioid Use and Total Joint Replacement. *Current Rheumatology Reports*. 2020;22:58.
35. Bedard NA, DeMik DE, Dowdle SB, Owens JM, Liu SS, Callaghan JJ. Preoperative Opioid Use and Its Association With Early Revision of Total Knee Arthroplasty. *Journal of Arthroplasty*. 2018;33(11):3520–3.
36. Bedard NA, DeMik DE, Dowdle SB, Owens JM, Liu SS, Callaghan JJ. Does Preoperative Opioid Use Increase the Risk of Early Revision Total Hip Arthroplasty? *Journal of Arthroplasty*. 2018;33(7):S154–S6.
37. Weick J, Bawa H, Dirschl DR, Luu HH. Preoperative opioid use is associated with higher readmission and revision rates in total knee and total hip arthroplasty. *Journal of bone and joint surgery American volume*. 2018;100(14):1171–6.
38. Kim SC, Jin Y, Lee YC, Lii J, Franklin PD, Solomon DH, et al. Association of Preoperative Opioid Use With Mortality and Short-term Safety Outcomes After Total Knee Replacement. *JAMA network open*. 2019;2(7):e198061-e.
39. Nassar Y, Richter S. Proton-pump Inhibitor Use and Fracture Risk: An Updated Systematic Review and Meta-analysis. *Journal of Bone Metabolism*. 2018;25(3):141–51.
40. Poly TN, Islam MM, Yang HC, Wu CC, Li YC. Proton pump inhibitors and risk of hip fracture: a meta-analysis of observational studies. *Osteoporosis International*. 2019;30(1):103–14.
41. Cafri G, Li L, Paxton EW, Fan J. Predicting risk for adverse health events using random forest. *Journal of Applied Statistics*. 2017;45(12):2279–94.
42. van Belle V, Pelckmans K, van Huffel S, Suykens JAK. Improved performance on high-dimensional survival data by application of survival-SVM. *Bioinformatics*. 2011;27(1):87–94.
43. Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*. 1996;17:1169–86.
44. Faraggi D, Simon R. A neural network model for survival data. *Statistics in Medicine*. 1995;14(1):73–82.

45. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018;18:24.

46. Bühlmann P, Yu B. Boosting with the L2 loss: Regression and classification. Journal of the American Statistical Association. 2003;98(462):324–39.

47. Tutz G, Binder H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. Biometrics. 2006;62(4):961–71.

Figures

Figure 1

a) The rate of TKA revision over 8 years is shown using 1 - Kaplan-Meier survivorship estimates b) Non-parametric estimate of the baseline hazard of TKA revision over 8 years c) The rate of THA revision over 8 years is shown using 1 - Kaplan-Meier survivorship estimates d) Non-parametric estimate of the baseline hazard of THA revision over 8 years

Figure 2

The calibration of models predicting eight-year risk of revision are compared using smoothed calibration curves, with black diagonal line denoting line of perfect calibration for a) TKA and b) THA

Figure 3

The Index of Prediction Accuracy is used to compare prediction accuracy of Cox, Weibull, flexible parametric and random survival forest for prediction of revision over eight-year time period for a) TKA

and b) THA

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFile1.docx](#)
- [AdditionalFile2.docx](#)
- [AdditionalFile3.docx](#)