BMJ Surgery, Interventions, & Health Technologies

Predicting total knee replacement at 2 and 5 years in osteoarthritis patients using machine learning

Khadija Mahmoud ,[1] M Abdulhadi Alagha,[1,2] Zuzanna Nowinka,[1] Gareth Jones[1]

To cite: Mahmoud K, Alagha MA, Nowinka Z, *et al.* Predicting total knee replacement at 2 and 5 years in osteoarthritis patients using machine learning. *BMJ Surg Interv Health Technologies* 2023;**5**:e000141. doi:10.1136/bmjsit-2022-000141

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjsit-2022-000141).

Received 29 March 2022
Accepted 09 December 2022

[1]MSk Lab, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK
[2]Data Science Institute, The London School of Economics and Political Science, London, UK

Correspondence to
Dr M Abdulhadi Alagha;
h.alagha@imperial.ac.uk

ABSTRACT

Objectives Knee osteoarthritis is a major cause of physical disability and reduced quality of life, with end-stage disease often treated by total knee replacement (TKR). We set out to develop and externally validate a machine learning model capable of predicting the need for a TKR in 2 and 5 years time using routinely collected health data.

Design A prospective study using datasets Osteoarthritis Initiative (OAI) and the Multicentre Osteoarthritis Study (MOST). OAI data were used to train the models while MOST data formed the external test set. The data were preprocessed using feature selection to curate 45 candidate features including demographics, medical history, imaging assessments, history of intervention and outcome.

Setting The study was conducted using two multicentre USA-based datasets of participants with or at high risk of knee OA.

Participants The study excluded participants with at least one existing TKR. OAI dataset included participants aged 45–79 years of which 3234 were used for training and 809 for internal testing, while MOST involved participants aged 50–79 and 2248 were used for external testing.

Main outcome measures The primary outcome of this study was prediction of TKR onset at 2 and 5 years. Performance was evaluated using area under the curve (AUC) and F1-score and key predictors identified.

Results For the best performing model (gradient boosting machine), the AUC at 2 years was 0.913 (95% CI 0.876 to 0.951), and at 5 years 0.873 (95% CI 0.839 to 0.907). Radiographic-derived features, questionnaire-based assessments alongside the patient's educational attainment were key predictors for these models.

Conclusions Our approach suggests that routinely collected patient data are sufficient to drive a predictive model with a clinically acceptable level of accuracy (AUC>0.7) and is the first such tool to be externally validated. This level of accuracy is higher than previously published models utilising MRI data, which is not routinely collected.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The demand for total knee replacement (TKR) has increased exponentially in recent years, exerting a pressure on patients, surgeons and hospitals to decide on the timing of surgery.

⇒ Machine learning has the potential to forecast the need for TKR.

WHAT THIS STUDY ADDS

⇒ This study is the first to develop machine learning models using routinely collected accessible data and test these models using an external dataset and provides evidence that an externally validated machine learning model can predict the need for TKR with an acceptable level of accuracy.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Potential adoption of our tool provides early knee osteoarthritis patients with useful information regarding their likelihood of requiring TKR surgery over the next 2–5 years thus empowering them to make treatment decisions as well as lifestyle changes to reduce this risk.

⇒ The information would also assist health economists to understand and meet the future demand for knee replacement surgery.

INTRODUCTION

Osteoarthritis (OA) is the most common degenerative joint disease and a major cause of physical disability, pain and reduced quality of life (QOL) for patients, with increasing global prevalence due to ageing populations and obesity.[1] The resultant global socioeconomic burden of OA is estimated to cost in excess of £4.2 billion.[2] Total knee replacement (TKR) is an effective treatment for end-stage knee OA (KOA),[1] and in line with increasing disease prevalence, its use in the UK alone is expected to rise significantly from 70 000 per year at present, to at least 119 000 per year by 2035.[3]

A tool to evaluate the likelihood of a patient requiring a TKR over the next 5 years has much appeal. It would allow informed decision making by patients, both in terms of non-operative treatment such as lifestyle modification, and the timing of any surgical intervention. For clinicians and health economists, a better understanding of the likely case-load over a period of 2–5 years would allow for appropriate planning to meet demand.

Predictive modelling of the need for a TKR using machine learning (ML) has been explored. Among the earliest of TKR prediction tools was a population-based study using patient-reported risk factors to predict 10-year TKR risk.[4] The tool, however, was restricted to older female patients, limiting generalisability. Further studies have since been conducted using more complex ML strategies including deep learning. Studies exploring this pertain strong dependence on MRI image input and have previously predicted TKR risk at 2, 4 and 5 years[5 6] to predictive performances of up to area under the curve (AUC) 0.87±0.02.[7] Such studies have made strides in predicting TKR, however, dependence on MRI imaging is both costly and not routinely performed,[8] in addition to the use of deep learning strategies that are not very well understood and require significant computational power to analyse.[6 7] Additionally, despite promising predictive abilities none of the published ML models have been externally validated to date, which is a significant limitation to their general applicability.

To address these limitations, we set out to develop and validate a tool that predicts which patients, with or at high risk of KOA, will likely require a TKR in 2 and 5 years time, using patient information collected during routine clinical practice. Six different ML classification models were evaluated including multivariable logistic regression (LR), LASSO, RIDGE, decision tree (DT), random forest (RF) and gradient boosting machine (GBM). A number of factors may be considered when selecting ML models including understandability and complexity as while a complex model can identify more interesting patterns in the data, at the same time, it is harder to maintain and explain. Six of the simplest ML models that are best explained were thus selected.[9]

METHODS

A summary of the methodology is found in figure 1.

Data

Data source and exclusion criteria

This study used data from two multicentre USA-based prospective cohort studies of patients with, or at high risk, of KOA; the Osteoarthritis Initiative (OAI) and Multicentre Osteoarthritis Study (MOST).[10 11] The OAI study enrolled 4976 subjects (ages 45–79 years) between February 2004 and May 2006 at four clinical sites (Baltimore, Maryland; Columbus, Ohio; Pittsburgh, Pennsylvania; and Pawtucket, Rhode Island) and MOST enrolled 3026 subjects (ages 50–79) from April 2003 to April 2005 at two sites (Birmingham, Alabama and Iowa City, Iowa). Eligibility for OAI included subjects with, or at risk for, symptomatic femoral-tibial KOA, a cohort defined by the presence of both osteophytes and frequent symptoms in one or both knees, or frequent knee symptoms without radiographic changes, in one or both knees. For MOST, similar eligibility was used to select subjects but with a reliance on MRI rather than radiographs. Subjects with unilateral or bilateral TKR at baseline were excluded.

Data pre-processing

Feature selection

OAI and MOST databases included 96 and 103 features, respectively. Those representing possible risk factors for progression of KOA were identified based on literature and expert knowledge.[12 13] Forty-five relevant features present in both datasets were then selected (summarised in online supplemental table 1). Of note, the criteria for the feature 'steroid injection history' was different between the datasets, being recorded over the previous 12 months in OAI, and 6 months in MOST.

Feature extraction

Selected features were categorised into the following domains: demographic, medical history, imaging assessments, history of intervention and outcome, with 39 non-imaging features and six image-based features. Medical history comprised both clinical examination and patient-reported outcomes. Image-based variables were quantitative radiographic measures: Kellgren-Lawerence grade (KLG) and joint space narrowing (JSN) .

The MOST protocol imputed random numbers for missing feature responses, and we applied the same approach to any missing features in the OAI dataset (online supplemental table 2).

Data split

The dataset was divided into three for the purposes of analysis (figure 1): 80% of the OAI dataset was used to develop and optimise the models (training set) with the remaining 20% of the dataset used for internal evaluation (internal test set). The MOST dataset was used for external validation. The OAI training and test datasets were randomly stratified in R to contain similar proportions of positive (having had a TKR) and negative (no TKR) cases.

Data output

Our study outcome variable of TKR was a binary 'yes' or 'no' for each patient case at 2 and 5 years.

Models

Model development and training

Model configuration and optimisation

Supervised ML models were used to predict the outcome, categorising new probabilistic observations into the predefined categories of 'yes' or 'no' TKR at 2 and 5 years. ML software packages were used on R V.3.6.3 (packages used detailed in online supplemental table 3) for reproducibility). The following ML classification models were selected: multivariable LR, LASSO, RIDGE, DT, RF and GBM. For each model, a number of tuneable knobs (parameters and hyperparameters) were adjusted to optimise performance (see online supplemental material 'model optimisation').

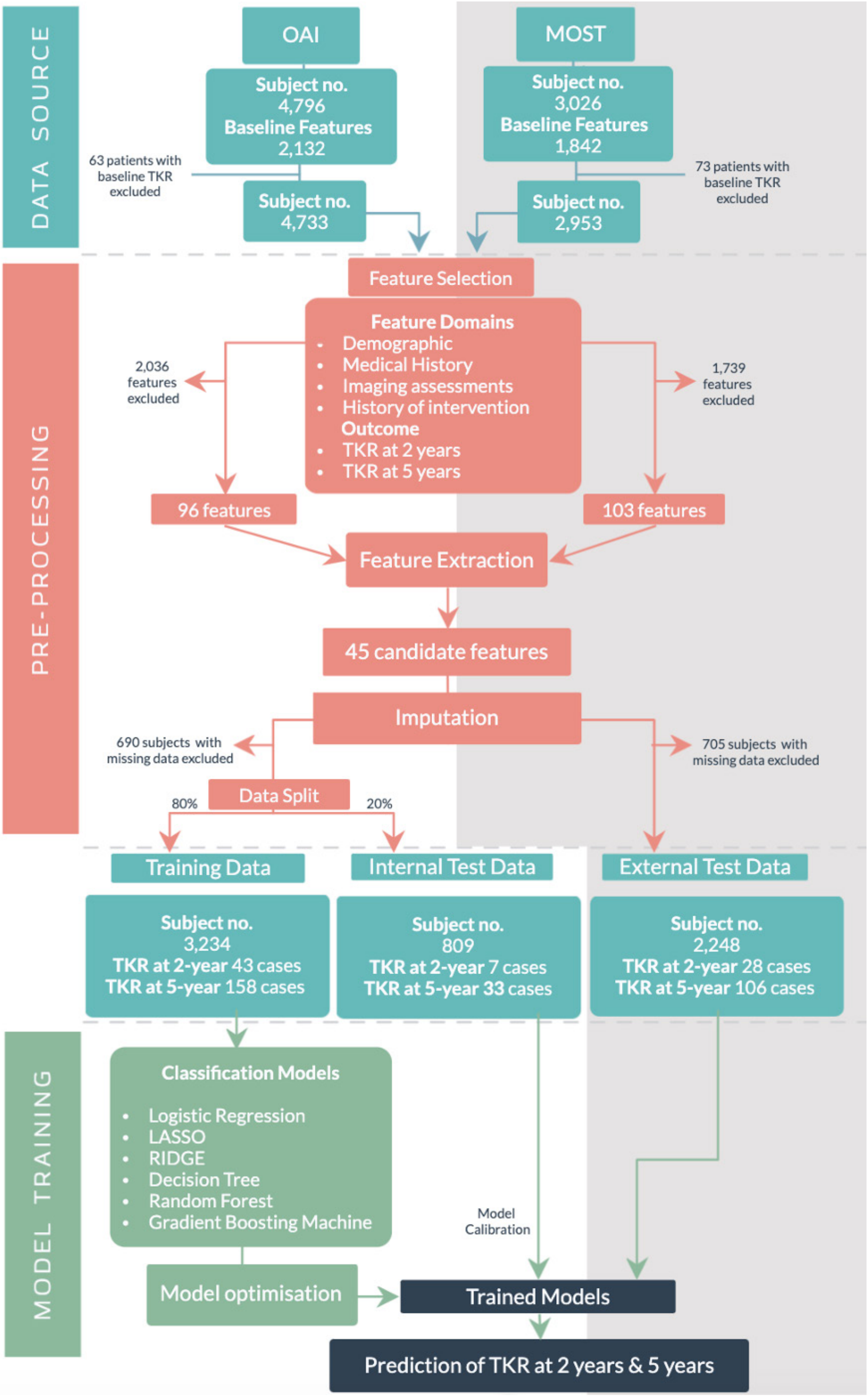


Figure 1 A Summary of the methodology, based on subject and feature disposition. The flow chart demonstrates the initial cohort, exclusion, approaches implemented at each stage, and resulting subjects and features included in analysis. The shaded section reflects the separation of the external dataset throughout. OAI, Osteoarthritis Initiative; TKR, total knee replacement.

Evaluation metrics

Model performance on all three data sets was evaluated with the area under the receiver operating characteristic (ROC) curve (AUC) for discrimination, with focused reporting on the internal test and external test sets. We considered AUC >0.7 to provide a clinically acceptable performance.[14] F1-scores were calculated for the best performing metrics as a harmonic mean of the precision and recall (sensitivity)[15 16] and a measure of positive predictive power. Key predictors in the best-performing model, at 2 and 5 years, were identified using variable importance evaluation functions of the ML models.

Model calibration

The optimal threshold for calibration, in line with the variation in numbers of positive and negative cases within datasets[12] was determined using F1-score, in order to optimise positive predictive ability (online supplemental figure 1).

Table 1 Data were split into training, internal test set and external test set as displayed

Feature domain	Training set	Internal test	External test
Sample size	N=3234	N=809	N=2248
Demographic			
Sex=male (%)	1345 (41.6)	344 (42.5)	947 (42.1)
Age (%)			
≤60 years	1512 (46.8)	391 (48.3)	929 (41.3)
60–70 years	1001 (31.0)	239 (29.5)	872 (38.8)
≥70 years	721 (22.3)	179 (22.1)	447 (19.9)
Body mass index (%)			
<18.5	7 (0.2)	3 (0.4)	6 (0.3)
18.5 to <25	787 (24.3)	204 (25.2)	357 (15.9)
25–30	1281 (39.6)	316 (39.1)	849 (37.8)
≥30	1159 (35.8)	286 (35.4)	1036 (46.1)
Ethnicity (%)			
White/Caucasian	2636 (81.5)	655 (81.0)	1923 (85.5)
Black/African	526 (16.3)	131 (16.2)	299 (13.3)
Hispanic/Latino	16 (0.5)	4 (0.5)	5 (0.2)
Other	56 (1.7)	19 (2.3)	21 (0.9)
Educational attainment (%)			
Less than high school graduate	87 (2.7)	27 (3.3)	67 (3.0)
High school graduate	386 (11.9)	108 (13.3)	505 (22.5)
College/associate degree/technical school after high school	762 (23.6)	192 (23.7)	601 (26.7)
College graduate	712 (22.0)	166 (20.5)	459 (20.4)
Some graduate school	272 (8.4)	63 (7.8)	199 (8.9)
Graduate degree	1015 (31.4)	253 (31.3)	417 (18.5)
Medical history			
Arthritis medical history (%)			
No arthritis history	1906 (58.9)	480 (59.3)	1103 (49.1)
At least one OA/degenerative disease	1027 (31.8)	262 (32.4)	758 (33.7)
Gout/other	140 (4.3)	32 (4.0)	124 (5.5)
OA/degenerative disease and gout/other	105 (3.2)	21 (2.6)	126 (5.6)
Unknown	56 (1.7)	14 (1.7)	137 (6.1)
Short-Form 12 Mental (mean (SD))	53.72 (7.92)	53.48 (7.98)	53.91 (8.94)
Short-Form 12 Physical (mean (SD))	49.54 (8.65)	49.63 (8.76)	46.43 (10.38)
Imaging—OA severity			
Kellgren and Lawrence Grade Left Knee (%)			
0: Normal	1243 (38.4)	311 (38.4)	1098 (48.8)
1: Minimal	592 (18.3)	133 (16.4)	379 (16.9)
2: Radiographic tibiofemoral knee OA	860 (26.6)	234 (28.9)	321 (14.3)
3: Moderate OA	433 (13.4)	102 (12.6)	320 (14.2)
4: Severe OA	106 (3.3)	29 (3.6)	130 (5.8)
Kellgren and Lawrence grade right knee (%)			
0	1289 (39.9)	342 (42.3)	1022 (45.5)
1	607 (18.8)	118 (14.6)	390 (17.3)
2	793 (24.5)	216 (26.7)	372 (16.5)
3	448 (13.9)	105 (13.0)	347 (15.4)
4	97 (3.0)	28 (3.5)	117 (5.2)

Continued

Table 1 Continued

Feature domain	Training set	Internal test	External test
Sample size	**N=3234**	**N=809**	**N=2248**
History of intervention			
No analgesics (%)	2478 (76.6)	624 (77.1)	448 (19.9)
No arthritis medication (%)	3195 (98.8)	799 (98.8)	1549 (68.9)
No osteoporosis medication (%)	2813 (87.0)	705 (87.1)	1829 (81.4)
No previous arthroscopy (%)	2690 (83.2)	656 (81.1)	1913 (85.1)
No previous meniscectomy (%)	2763 (85.4)	678 (83.8)	1920 (85.4)
No previous ligament repair surgery (%)	3127 (96.7)	781 (96.5)	2152 (95.7)
No previous other surgery(%)	3143 (97.2)	786 (97.2)	2192 (97.5)
Outcome			
TKR at 5 year (%)	158 (4.9)	33 (4.1)	106 (4.7)
TKR at 2 year (%)	43 (1.3)	7 (0.9)	28 (1.2)

To prevent data leakage, all entries from any given patient were only allowed to be in one of the three sets.
OA, osteoarthritis; TKR, total knee replacement.

RESULTS

Data distribution

The distribution of key candidate features is displayed in table 1. The training set comprised 3234 patients of which 41.6% were male, and 43.3% and 41.4% had radiographic, moderate or severe left KOA and right KOA, respectively. The internal test set consisted of 809 patients of which 42.5% were male, and 45.1% and 43.2% had radiographic, moderate or severe left KOA and right KOA, respectively. The external test set included 2248 patients of which 42.1% were male, and 34.3% and 37.1% had radiographic, moderate or severe left KOA and right KOA, respectively. Correlation between features within the primary dataset is visualised as a correlation heatmap (online supplemental figure 2).

Training and internal test performance

Optimised predictive abilities for each model applied to the training and internal test sets are detailed in table 2. The best performing model at 2 years was GBM AT 0.945 (95% CIs 0.901 to 0.988) and RIDGE at 5 years 0.869 (0.803 to 0.935). The worst performing model at 2 years was LR with an AUC of 0.730 (95% CI 0.496 to 0.965) and at 5 years DT at an AUC of 0.688 (95% CI 0.608 to 0.768). The DT model was unable to categorise any cases at 2 years because the uniform probability threshold selected for model calibration was not optimal. Performances on the external dataset (table 3) revealed that GBM models were best for both time points, with an AUC of 0.913 (95% CI 0.876 to 0.951) and 0.873 (95% CI 0.839 to 0.907) for 2 and 5 years, respectively. When applied to the external test set, low positive predictive ability is evident across both years as denoted by low F1-scores.

Overall, the best performing models, based on performance on the internal test set, were GBM, RIDGE and LASSO. TKR prediction at 2 years was also consistently more accurate than at 5 years for the three best performing models.

External test performance

ROC curves for the three best performing models are shown in figure 2A–C when applied to the internal test

Table 2 Displaying AUC for all five models predicting TKR at 2 years and 5 years when applied to training and internal test sets

	AUC			
	2 years		5 years	
Model type	Training set	Test set	Training set	Test set
Logistic regression	0.985 (0.974 to 0.996)	0.730 (0.496 to 0.965)	0.932 (0.915 to 0.950)	0.822 (0.745 to 0.898)
RIDGE	0.954 (0.933 to 0.974)	0.916 (0.865 to 0.967)	0.908 (0.886 to 0.929)	0.869 (0.803 to 0.935)
LASSO	0.966 (0.946 to 0.986)	0.901 (0.840 to 0.962)	0.907 (0.886 to 0.929)	0.864 (0.801 to 0.928)
Decision tree	–	–	0.696 (0.656 to 0.736)	0.688 (0.608 to 0.768)
Random forest	0.789 (0.713 to 0.864)	0.821 (0.690 to 0.952)	0.831 (0.795 to 0.867)	0.845 (0.799 to 0.910)
Gradient boosting machine	0.942 (0.914 to 0.970)	0.945 (0.901 to 0.988)	0.905 (0.883 to 0.927)	0.855 (0.794 to 0.915)

SEs were used to determine 95% CIs (shown in brackets).
*DT was unable to categorise cases at 2 years (Table 2)
AUC, area under the curve; DT, decision tree; TKR, total knee replacement.

Table 3 Displaying the performance of three models when applied to the external testset (MOST), as evaluated by AUC and F1-score

	External validation			
	2 years		5 years	
Model type	AUC	F1-score	AUC	F1-score
Gradient boosting machine	0.913 (0.876–0.951)	0.171	0.873 (0.839–0.907)	0.287
LASSO	0.805 (0.716–0.895)	0.118	0.841 (0.804–0.878)	0.267
RIDGE	0.820 (0.748–0.893)	0.0811	0.825 (0.785–0.864)	0.261

AUC, area under the curve; MOST, Multicentre Osteoarthritis Study .

set as well as figure 2B,C, when applied to the external test set at both time points. Performances across models are slightly reduced when applied to the external test set at both time points but remains within comparison to the internal test set, with the exception of GBM which exceeds its original performance when applied to the external test set at 5 years (AUC 0.855 compared with 0.873). GBM consistently forms the best performing model (AUC-2years=0.913, AUC-5years=0.873).

Model predictors

Relative influence is ranked to show order of the most important feature in training the model in table 4. For instance, 21.54 relative influence means it accounts for 21.54% of the reduction to the loss function given this set of features as opposed to 21.54% of variance. Radiographic features; KLG formed the highest predictor in the best performing model (GBM) across both prediction years, followed by less important features The Western

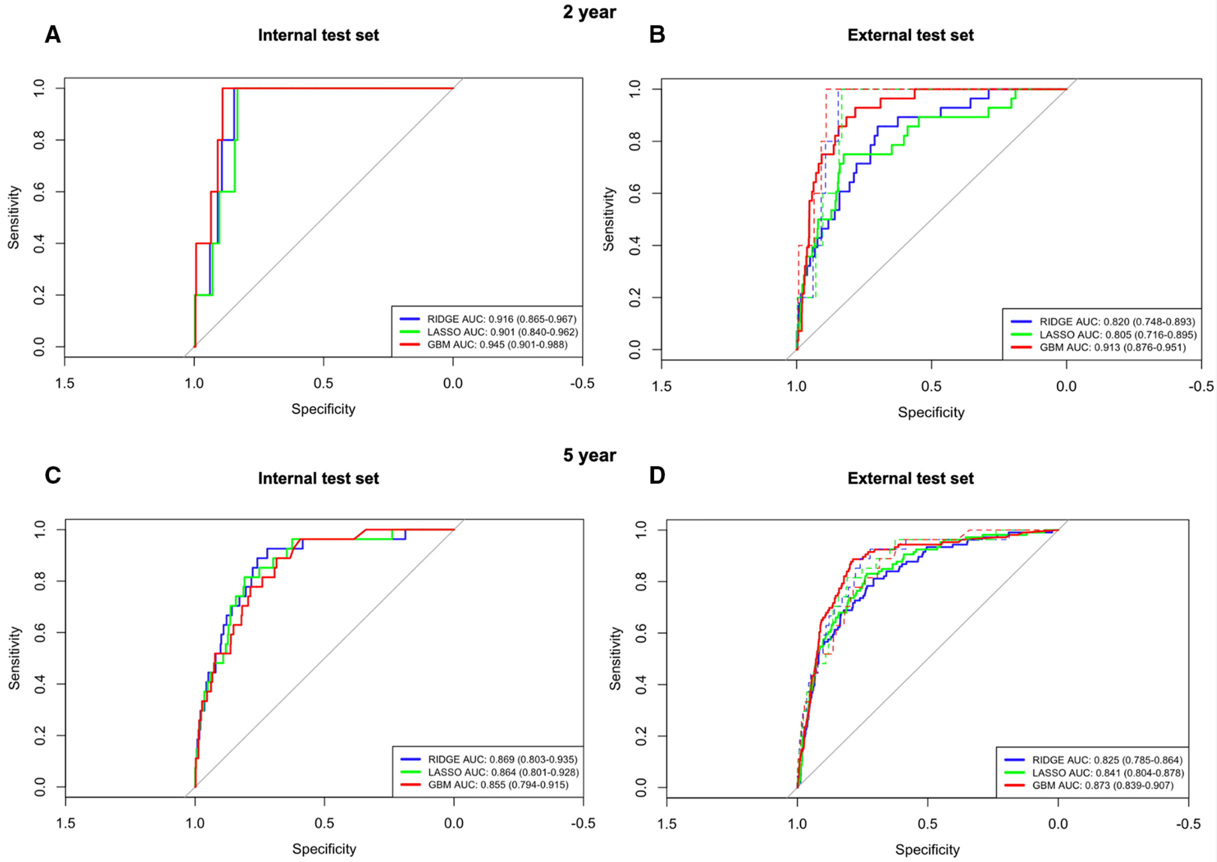


Figure 2 Comparison of the top three performing ml models' performance as receiver operating characteristic (ROC) curves for TKR prediction at 2 and 5 years. (A) and (C) demonstrate ROC curves on internal test set only (B) and (D) on external test set (MOST), with additional dashed lines that are the test set overlain to allow direct comparison. In all curves, the black line signifies the performance of a random classifier (area under the curve, AUC=0.500). The legends in the subplots indicate the AUC of the models with 95% CIs. GBM, gradient boosting machine; MOST, Multicentre Osteoarthritis Study; TKR, total knee replacement.

Table 4 Denotes the largest predictors for the best-performing model (GBM) alongside their relative influence at 2 years and 5 years. *The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) .

2 years		5 years	
Predictor	Relative influence	Predictor	Relative influence
KLG left	21.54	KLG left	31.65
WOMAC* score right (total)	9.36	JSN right	18.56
Short Form-12 Mental	6.48	KLG right	15.50
Short Form-12 Physical	6.39	WOMAC* score right (total)	5.91
KLG right	6.18	Short Form-12 Mental	4.56
Educational attainment	4.36	Educational attainment	4.29

Note: WOMAC score was assessed per leg, providing a separate total score for the right and left.
GBM, gradient boosting machine; JSN, joint space narrowing; KLG, Kellgren-Lawerence grade .

Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Short-Form-12 (Physical and mental components) and educational attainment (table 4). JSN also appeared to have a relatively high influence at 5 years despite being non-notable at 2 years. A number of remaining features or predictors were '0', and thus 'unnecessary' in predicting TKR at 2 and 5 years under the GBM model.

DISCUSSION

We set-out to predict the need for TKR at 2 and 5 years, using predictive variables that represent routinely collected data. With the exception of the DTs, the ML models produced were able to predict the need for a TKR with a clinically acceptable performance, using an independent external test set for validation. The GBM model achieved the highest predictive power at two (AUC 0.913 (95% CI 0.876 to 0.951)) and 5 years (0.873 (95% CI 0.839 to 0.907)). The key features driving the predictions of this best performing model were KLG, JSN, physical score features; WOMAC, SF-12 scores and educational attainment. This is the first study to validate predictive models externally, and the lack of reliance on MRI, with its associated costs and limited accessibility, facilitates the wider application of our tools through ease of interpretability, implementation and scalability to various clinical settings.

In terms of other non-imaging-reliant models, Wang *et al*[7] used OAI data to develop an LR model using selected demographic and clinical information, with an AUC of 0.77±0.02, which is lower than the internal and external dataset performance of all our top-performing models. Another study shared our novelty in using non-MRI-based features to predict TKR within 4 years[17] although their evaluation metrics did not include AUC but reported the total percentage of correctly predicted knees as 80% (69%–89%). However, this was not externally validated and conducted only on a sample of subjects as only the 165 patients receiving TKR were analysed. This study also used an artificial neural network, which carries advantages in terms of information processing, fault and noise tolerance compared with our ML models,[17] but they function as a 'black-box' and this lack of transparency may limit doctor and patient confidence in the model's predictions.[18 19]

In terms of imaging-dependent models, Tolpadi *et al* used direct imaging to predict TKR at 5 years for OAI subjects with varying OA severity.[5] The paper evaluated six models: raw imaging-based, non-imaging-based and integrated (both), for radiographic and MRI imaging, concluding that the model integrating MRI and non-imaging features outperformed the others. Interestingly, our AUC (GBM, internal test; 0.945 at 5 years) exceeded all six of their models performed on their internal test data: 0.868 (non-imaging), 0.848 (radiographic images only), 0.890 (integrated radiographic model), 0.886 (MRI images only) and 0.834 (integrated MRI model). Jamshidi *et al*,[6] a study that predicted TKR and time to TKR, also used MRI quantitative imaging data from OAI, developing a model with an AUC of 0.86, although this did not outperform our model. It should be noted that none of these previous studies validated their results using an external dataset, and so the real-world performance of their models remains uncertain.

Of note, Tolpadi *et al*'s model sensitivities exceeded that suggested by our F1-scores. This is important to consider as while the AUC considers the models' ability to assess both negative and positive cases, the F1-score considers precision; a measure of positive predictive power; the model's ability to predict TKR cases. Prediction of positive cases at both timepoints was <0.3 reflective of a lower sensitivity than Tolpadi *et al*'s. This suggests a bias of the AUC evaluation towards the majority class (negative cases), revealing that our models were better able to predict negative cases than positive. Explanation of our lower positive predicative abilities in comparison to Tolpadi *et al*'s potentially lie in their use of deep learning such as convolutional neural networks which use more advanced feature extraction to better manage the complex prognostic features that determine TKR risk,[5 20 21] thus, strengthening their positive predictive power. A distinct advantage of our models, however, was their simplicity and thus transparency as well as reliance on more obtainable data, particularly considering the higher costs and reduced availability of MRI.[8] Recent statistics estimate a single MRI scan to cost as much as US$1430 and £450 in the USA and UK, respectively.[22]

The transparency of our ML models also allowed us to examine the key predictors used by our most accurate model (GBM), and reassuringly they mostly align with previous literature findings.[12 23] A study which used RF modelling of the OAI dataset to explore TKR incidence over 2 years, selected the predictive variables used by our model that is, KLG, WOMAC and SF-12.[23] While this study was performed on OAI and thus, similarities with our findings are expected, the external validation of our study confirms the importance of these variables across different datasets. Elsewhere in the literature, a prospective Canadian population-based study identified WOMAC summary scores as key predictors for TKR risk, supporting our findings.[12] The other advantage of knowing which variables are most important is that data collection can be targeted, thus reducing the paperwork burden for both patients and physicians.

Interestingly, our models also identified education as a key predictor. While low socioeconomic status is well recognised as one of the strongest predictors of morbidity and mortality from many chronic diseases, there are little data regarding its impact within KOA.[24] One paper's analysis of the socioeconomic effect on KOA found educational attainment was associated with decreased KOA prevalence in their initial analyses, however, this association was lost after confounder adjustments.[24] Our finding may be a function of a correlation between higher rates of manual work, which are associated with increased risk of OA, among lower educational groups. Indeed, a study of pain disparities in underserved populations, within OAI, identified more severe OA in lower socioeconomic groups (inclusive of education) in addition to disparities in pain, and this was not accounted for by objective OA measures.[25] Alternatively, it may reflect the US insurance-based healthcare, with education serving as proxy for income and access to early healthcare intervention in a timely manner. Further exploration of educational attainment, in relation to OA and TKR may be merited.

The clinical relevance of our tool is dictated by its ability to use routinely collected data and transparent ML techniques to predict TKR with a clinically acceptable accuracy which surpasses previous models. Our model's independence from MRI scanning is important, because it resolves many of the issues of cost and accessibility and in doing so increases its potential for use in both the developed and developing countries. Our tool has the potential to facilitate targeted non-operative management efforts to modify risks for patients, particularly those predicted to require a TKR in 5 years time, with the aim of improving their QOL and potentially delaying the need for TKR. For patients predicted to require a TKR in 2 years, as well as modifying risk factors, this may assist with planning of care to closely monitor these patients and identify the ideal time to intervene surgically. Knowledge regarding the likelihood of requiring a TKR will empower and motivate patients, and facilitate informed shared decision making with their clinicians. It also has clear potential benefits for health economists tasked with planning future resource allocation.

A limitation of our study is the class imbalance in the dataset with the majority of patients included not progressing to have a TKR during the studied period. This is reflected in the low F1-scores, which suggest that our models were better at predicting negative cases, that is, patients not requiring a TKR at 2 or 5 years. Another limitation is the demographic imbalance in the OAI primary data, which has a bias towards older patients, as well as a higher proportion of female and white patients. Additionally, both datasets used were USA based, and further studies are required to confirm that the models are applicable outside of the USA.

This study presents the first externally validated ML model using simple and routinely available patient data, while delivering clinically acceptable levels of predictive power, to forecast a patient's need for TKR at 2 and 5 years. The simplicity and transparency of our models in terms of design and input, with no reliance on MRI, increases the likelihood of its adoption as a treatment decision aid, identifying patients who are more likely to benefit from non-operative management and risk factor modification. Sharing this information with patients would also be expected to facilitate shared decision making and empower them to play an active role in their KOA management. Future research will explore the accuracy of our models in non-US populations and the use of advanced sampling techniques to address the class distribution balance.

Acknowledgements This manuscript was prepared using OAI and MOST public use datasets which both received grants from National Institutes of Health (NIH), a branch of the Department of Health and Human Services. The OAI study is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262). The MOST study was supported by NIH-National Institute on Aging grants to participating university centres (AG18820; AG18832; AG18947; AG19069). MOST and OAI were both conducted by their respective study investigators.

Contributors KM, MAA, ZN and GJ were involved in setting out the project aim and methodology. KM conducted the literature search and wrote the original draft. KM, MAA and ZN contributed to data curation and analysis. MAA and GJ contributed to study design. GJ supervised the conduction of the study, and reviewed and edited the manuscript. All authors had access to the raw data and have approved the final manuscript. GJ acting as guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Disclaimer This manuscript is not necessarily reflective of the opinions of the respective investigators, the NIH, or the private funding partners.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Khadija Mahmoud http://orcid.org/0000-0002-2869-5778

REFERENCES

1 Glyn-Jones S, Palmer AJR, Agricola R, *et al*. Osteoarthritis. *The Lancet* 2015;386:376–87.
2 Chen A, Gupte C, Akhtar K, *et al*. The global economic cost of osteoarthritis: how the UK compares. *Arthritis* 2012;2012:1–6.
3 Culliford D, Maskell J, Judge A, *et al*. Future projections of total hip and knee arthroplasty in the UK: results from the UK clinical practice research datalink. *Osteoarthritis Cartilage* 2015;23:594–600.
4 Lewis JR, Dhaliwal SS, Zhu K, *et al*. A predictive model for knee joint replacement in older women. *PLoS One* 2013;8:e83665.
5 Tolpadi AA, Lee JJ, Pedoia V, *et al*. Deep learning predicts total knee replacement from magnetic resonance images. *Sci Rep* 2020;10:6371.
6 Jamshidi A, Pelletier J-P, Labbe A, *et al*. Machine learning-based individualized survival prediction model for total knee replacement in osteoarthritis: data from the osteoarthritis initiative. *Arthritis Care Res* 2021;73:1518–27.
7 Wang T, Leung K, Cho K. Total knee replacement prediction using structural MRIs and 3D convolutional neural networks. *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, 2019:79.
8 NICE. Recommendations | osteoarthritis: care and management | guidance, Nice.org.uk, 2020. Available: https://www.nice.org.uk/guidance/cg177/chapter/1-recommendations#referral-for-consideration-of-joint-surgery-2
9 Valdarrama S. Considerations when choosing a machine learning model. Available: https://towardsdatascience.com/considerations-when-choosing-a-machine-learning-model-aa31f52c27f3 [Accessed 03 Oct 2022].
10 Osteoarthritis initiative (OAI) study protocol. Available: https://nda.nih.gov/oai/study-details [Accessed 10 May 2021].
11 Multicenter osteoarthritis study (most) public data sharing. Available: https://most.ucsf.edu/ [Accessed 10 May 2021].
12 Hawker GA, Guan J, Croxford R, *et al*. A prospective population-based study of the predictors of undergoing total joint arthroplasty. *Arthritis Rheum* 2006;54:3212–20.
13 Heidari B. Knee osteoarthritis prevalence, risk factors, pathogenesis and features: part I. *Caspian J Intern Med* 2011;2:205–12.
14 Hosmer Jr D, Lemeshow S, Stardivant R. *Applied logistic regression*. 3rd ed. USA: Wiley, 2013.
15 Brownlee J. Classification accuracy is not enough: more performance measures you can use. -03-20T18:00:04+00:00, 2014. Machine learning mastery. Available: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/ [Accessed 18 May 2021].
16 Brownlee J. How to calibrate probabilities for imbalanced classification. -02-25T18:00:51+00:00, 2020. Machine learning mastery. Available: https://machinelearningmastery.com/probability-calibration-for-imbalanced-classification/[Accessed 18 May 2021].
17 Heisinger S, Hitzl W, Hobusch GM, *et al*. Predicting total knee replacement from Symptomology and radiographic structural change using artificial neural networks-data from the osteoarthritis initiative (OAI). *J Clin Med* 2020;9. doi:10.3390/jcm9051298. [Epub ahead of print: 01 05 2020].
18 Gunaratne R, Pratt DN, Banda J, *et al*. Patient dissatisfaction following total knee arthroplasty: a systematic review of the literature. *J Arthroplasty* 2017;32:3854–60.
19 Kelly CJ, Karthikesalingam A, Suleyman M, *et al*. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195.
20 Yu D, Jordan KP, Snell KIE, *et al*. Development and validation of prediction models to estimate risk of primary total hip and knee replacements using data from the UK: two prospective open cohorts using the UK clinical practice research datalink. *Ann Rheum Dis* 2019;78:91–9.
21 He K, Zhang X, Ren S. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016:770–8.
22 Statista. Average price MRI in selected countries 2017; 2017. https://www.statista.com/statistics/312020/price-of-mri-diagnostics-by-country/ [Accessed 16 May 2021].
23 Riddle DL, Kong X, Jiranek WA. Two-year incidence and predictors of future knee arthroplasty in persons with symptomatic knee osteoarthritis: preliminary analysis of longitudinal data from the osteoarthritis initiative. *Knee* 2009;16:494–500.
24 Shirinsky I, Shirinsky V. SAT0564 effects of education and income on prevalence, incidence, and progression of radiographic knee osteoarthritis: an analysis of the osteoarthritis initiative data. *Annals of the Rheumatic Diseases* 2018;77:1135.
25 Pierson E, Cutler DM, Leskovec J, *et al*. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med* 2021;27:136–40.

Supplementary methodology & tables

Domain		**Features and (input name)**
Demographic	Demographic	• Sex (SEX) • Body Mass Index (BMI) • Age (Age) • Blood pressure (BP)
	Socioeconomic	• Smoking by pack years (Smoking) • Ethnicity (Ethnicity) • Marital status (MarriageStatus) • Living alone/with others (LivingStatus) • Paid employment (WORKFORPAY) • Educational attainment (EDUCATION)
Medical History	Comorbidities	• Heart failure (HRTFAIL) • Heart Attack (HRTAT) • Stroke (STROKE) • Asthma (ASTHMA) • Emphysema, COPD, chronic bronchitis (LUNG) • Stomach Ulcer (ULCER) • Diabetes (DIAB) • Kidney problems(KIDFXN)
	Arthritis-specific	• Arthritis past-medical history (ArthritisPMH) • Either knee, ever injured badly enough to limit ability to walk for at least two days (Injury) • Either knee, pain, aching or stiffness: ever had more than half the days of a month (pain) • Either knee, limit activities due to pain, aching or stiffness, past 30 days (LIMITACTIVITY)
	Scoring systems	Mental • Center for Epidemiological Studies Depression Score(CESD) • Short-Form 12 Mental Component (SF12mental) Physical • Short Form 12 Physical Component (SF12physical) • Total Western Ontario and McMaster Universities Osteoarthritis Index Right Knee (WOMTSR) • Total Western Ontario and McMaster Universities Osteoarthritis Index Left Knee (WOMTSL) • Physical Activity Scale for the Elederly Score (PASE)
	Clinical examination	• Clinic 20-meter walk assessment (WALKTIMET1) • Timed chair stands (chaircat)
Osteoarthritis severity	Imaging Assessments	• Baseline Kellgren and Lawrence Grade on PA view (KLGLEFT, KLGRIGHT) • Baseline joint space narrowing medial/lateral TibioFemoral (JSMLEFT, JSLLEFT, JSMRIGHT, JSLRIGHT)
History of Intervention	Medication	• Osteoporosis medication; Vitamin /Bisphosphonate/Estrogen or Raloxifene/Calcitonin or Teriparatide (Osteop_med) • Analgesic medication; Salicylates/NSAIDs / COX2/Opioids/Combination/Other (Analgesics) • Arthritis medication; Oral corticosteroids/Supplements (SAMe, MSM, Fluorides, Glucosamine) (Arth_med) • Steroid Injection, past 12M (OAI) and past 6M (MOST) (steroid_inj)
	Knee-related surgical intervention	• Either knee, ever have arthroscopy (Knee_arth) • Either knee, ever have meniscectomy (knee_men) • Either knee, ever have ligament repair surgery (knee_ligament) • Either knee, ever have any other kind of surgery (knee_other)

Outcome	Total Knee Replacement (TKR)	• TKR within 5 years from baselines (FIVEYR) • TKR within 2 years from baseline (TWOYR)

Supplementary Table 1. *Finalised list of candidate features input into models, alongside outcome features.*

Feature extraction

The feature 'analgesics' was created based on a combination of 4 features (use of; Salicylates, Non-steroidal anti-inflammatory drugs, Opioids, No analgesics) forming an ordered ladder, in line with the WHO analgesic ladder[29]. Blood pressure was also categorised in accordance with the American Heart Association's guidelines[30]. Chair stands were categorised using a cut off point of 10 seconds, where 0 indicated no risk, whilst >=10 seconds was assigned 1 to indicate potential risk. This was in accordance with a large study (n= 4,335 community-dwelling adults) which suggested optimal cut-off points[31]. Other categorisations were based on combining multiple features into one to produce the finalised candidate features.

Feature Imputed	Number of cases replaced with '8' where missing
Smoking	*257*
Arthritis Past Medical History	*89*
Analgesic medication	*13*
Arthritis medication	*13*
Osteoporosis medication	*11*
Clinic 20-meter walk assessment	*21*
Timed chair stands	*247*
Blood Pressure	*1*

Supplementary Table 2. *Features imputed in OAI dataset, in addition to number of cases affected.*

Package	Model applied to
glm (version 3.6.2)	Logistic Regression (fitting generalised linear models)
glmnet (version 4.1-1)	LASSO & RIDGE (fitting a generalised linear model with regularisation)
rpart (version 4.1-15) rpart.plot (version 3.0.9)	Decision Tree (Recursive Partitioning and Regression Trees) Plot an rpart model
randomForest (version 4.6-14)	Classification and Regression with Random Forest
gbm (version 2.1.8)	Gradient Boosting Machine
Data visualisations	
pROC	To display and analyse ROC curves.
corrplot (version 0.88)	Correlations heatmap: A visualization of a correlation matrix.

ROCR version (1.0-11)	F1-score variation with threshold, used in model calibration

Supplementary Table 3. *Software packages used on R version 3.6.3*

Model optimisation

For both RIDGE and LASSO, hyperparameter tuning involved deciding the parameter (lambda) that controls the overall strength of the penalty. In both models, cross validation was used to determine the value of lambda that gave the minimum mean cross-validated error.

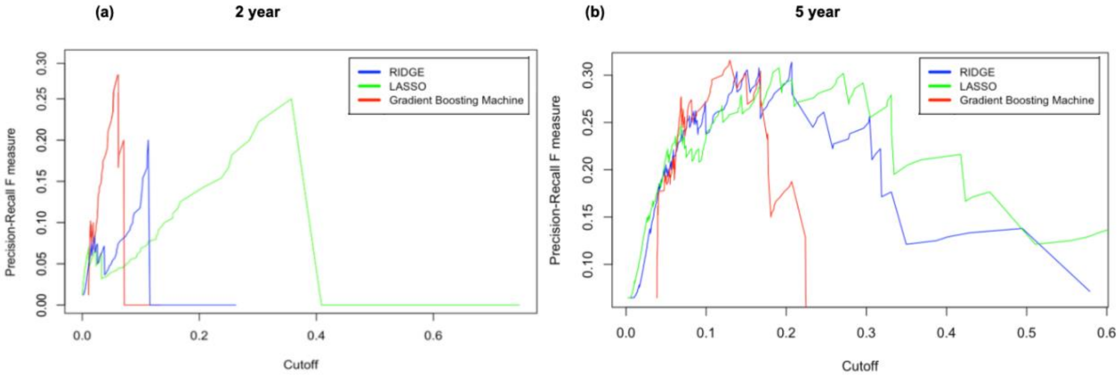
For RF, optimal tuning of parameters was manually performed, where the model's out-of-bag error approximated the optimum values. In consideration of our data size and feature number, 500 trees were grown. Changes to the model architecture were also performed. This included the number of features randomly sampled as candidates at each split which was determined optimal at 10. Similarly, a larger node size was selected, which specifies the minimum number of observations in a terminal node. This adjustment decreased tree depth to enable fewer splits to be performed until the terminal nodes. The maximum number of leaf nodes was capped to reduce overfitting by reducing the possible number of paths to leaf nodes.

For GBM, an identical forest size of 500 trees was chosen. An optimal interaction depth (maximum nodes per tree) was tuned alongside a low learning rate (shrinkage) to improve the model's generalisation.

For the base models, LR and DT, no parameter adjustments outside of default were performed. Similarly, where additional hyperparameters are unspecified across all models, default mode for each package was selected.

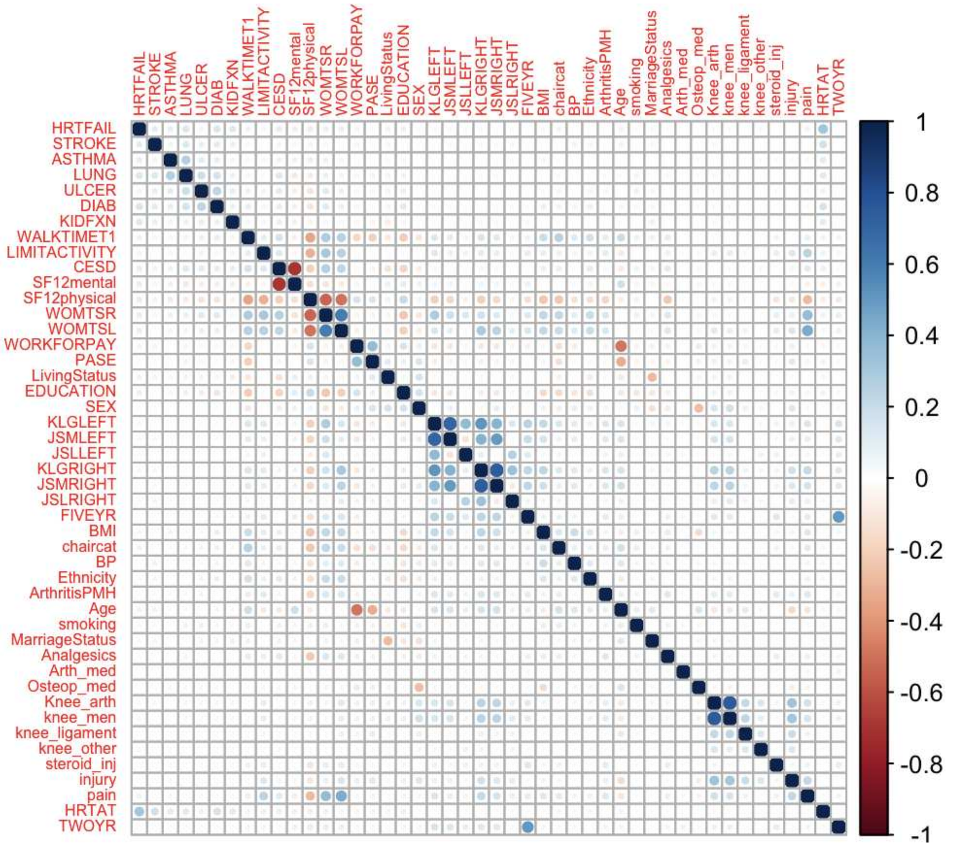
Supplementary figures

Each of the ML models generate probabilities as a basis for classification and consequently require the selection of an optimal threshold as part of calibration, in line with the variation in numbers of positive and negative cases within datasets . Given the class imbalance in our dataset, owing to the small number of positive cases, an optimal threshold was determined using F1-score, in order to optimise positive predictive ability. This was calculated at the optimal F1 threshold for the selected models using Figure 3. Approximately, at 2 and 5 years these were; <0.2 for RIDGE and GBM. For LASSO, these were around 0.4 at 2 years and 0.2 at 5 years.



Supplementary Figure 1. *The effect of varying threshold (Cut-off) on Precision-Recall F measure (F1 Score) for (a) 2 year prediction and (b) 5 year prediction for internal test-sets, where threshold is denoted as 'Cut Off' on the x-axes. Across both graphs, higher F1-Scores are shown to be achieved at lower thresholds than the default threshold of 0.5.*

Correlation heatmap was used to understand feature interactions before inputting features into the models. No consequent adjustments to candidate features were made as multicollinearity affects only the specific independent variables that are correlated and thus, given that there was no high correlation with our outcome features (TKR at 2 and 5 years), it does not affect model predictive ability and interpretation.



Supplementary figure 2. *Correlation heatmap as applied to the primary dataset (OAI) to display relationships between features. Correlation ranges from -1 to +1. Values closer to zero indicate no linear trend between the two features. Colour scale indicates strength of correlation, where 1 is perfect positive correlation and -1 is perfect negative correlation. Full feature names are detailed in Supplementary Table 1 to aid interpretation (where the feature is not clear from its input name above).*