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.

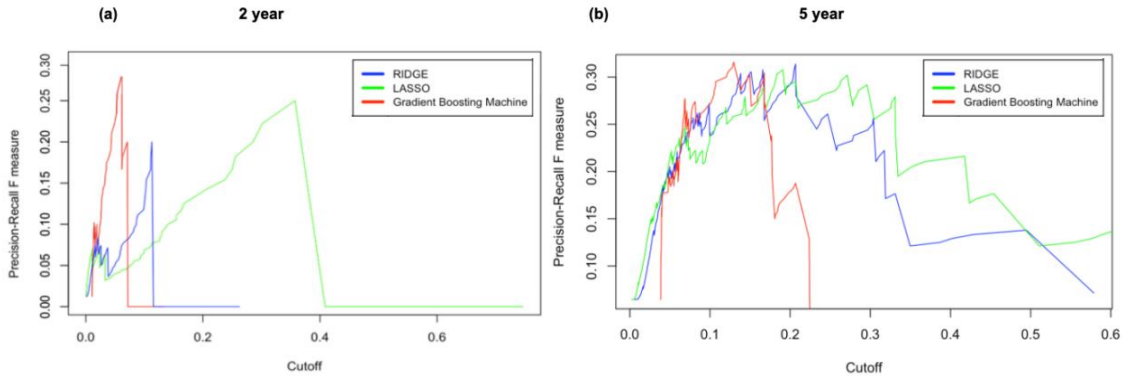
For RF, optimal tuning of parameters was manually performed, where the model's out-of-bag error approximated the optimum values. In consideration of our data size and feature number, 500 trees were grown. Changes to the model architecture were also performed. This included the number of features randomly 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.

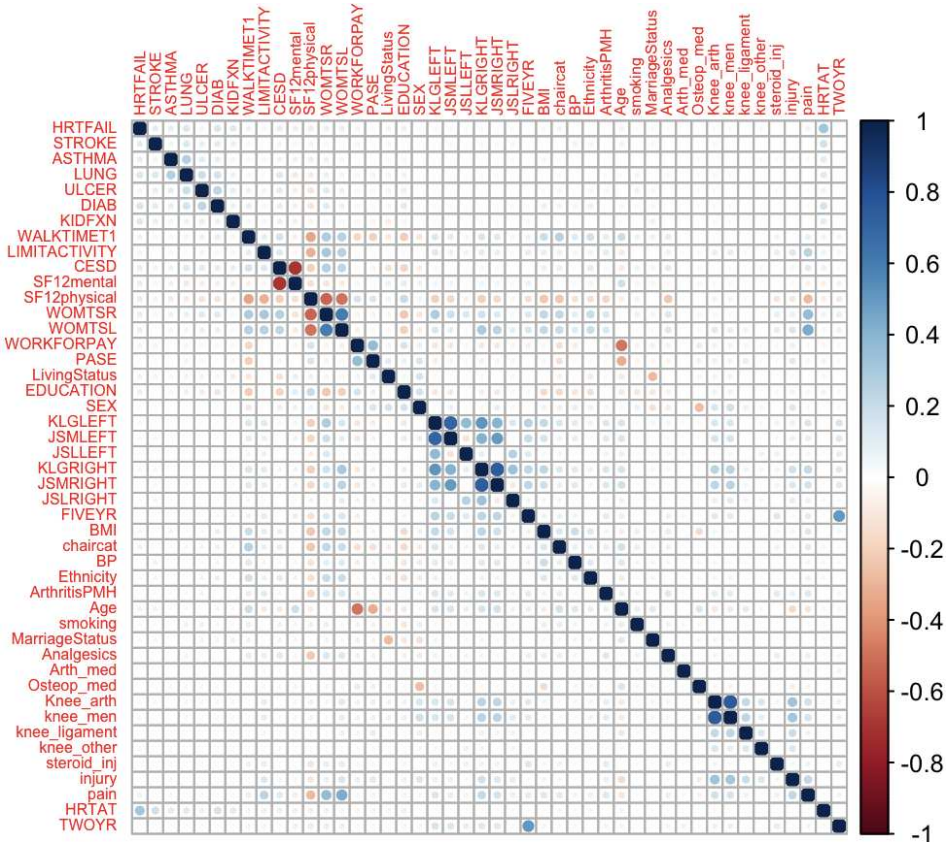
Supplementary figures

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).*