Predicting the onset of end-stage knee osteoarthritis over two- and five-years using machine learning

Salis, Zubeyir; Driban, Jeffrey B.; McAlindon, Timothy E.

How to cite

SALIS, Zubeyir, DRIBAN, Jeffrey B., MCALINDON, Timothy E. Predicting the onset of end-stage knee osteoarthritis over two- and five-years using machine learning. In: Seminars in arthritis and rheumatism, 2024, vol. 66, p. 152433. doi: 10.1016/j.semarthrit.2024.152433

This publication URL: https://archive-ouverte.unige.ch/unige:179069
Publication DOI: 10.1016/j.semarthrit.2024.152433

Seminars in Arthritis and Rheumatism 66 (2024) 152433

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

Predicting the onset of end-stage knee osteoarthritis over two- and five-years using machine learning

Zubeyir Salis [a,b,c,*], Jeffrey B. Driban [d], Timothy E. McAlindon [e]

[a] Division of Rheumatology, Geneva University Hospitals and Faculty of Medicine, University of Geneva, Geneva, Switzerland
[b] School of Human Sciences, the University of Western Australia, Perth, WA, Australia
[c] Centre for Big Data Research in Health, the University of New South Wales, Kensington, NSW, Australia
[d] UMass Chan Medical School, Department of Population and Quantitative Health Sciences, Worcester, MA, USA
[e] Division of Rheumatology, Allergy, and Immunology; Tufts Medical Center, Boston, MA, USA

ARTICLE INFO

Keywords:
Health care quality, access, and evaluation
Learning curve
Patient outcome assessment
Technology assessment, Biomedical

ABSTRACT

Objective: Identifying participants who will progress to advanced stage in knee osteoarthritis (KOA) trials remains a significant challenge. Current tools, relying on total knee replacements (TKR), fall short in reliability due to the extraneous factors influencing TKR decisions. Acknowledging these limitations, our study identifies a critical need for a more robust metric to assess severe KOA. The end-stage KOA (esKOA) measure, which combines symptomatic and radiographic criteria, serves as a solid indicator. To enhance future trials that use esKOA as an endpoint, our study focuses on developing and validating a machine-learning tool to identify individuals likely to develop esKOA within 2 to 5 years.

Design: Utilizing the Osteoarthritis Initiative (OAI) data, we trained models on 3,114 participants and validated them with 606 participants for the right knee, and similarly for the left knee, with external validation from the Multicentre Osteoarthritis Study (MOST) involving 1,602 participants. We aimed to predict esKOA onset at 2-to-2.5 years and 4-to-5 years, defining esKOA by severe radiographic KOA with moderate/severe symptoms or mild/moderate radiographic KOA with persistent/intense symptoms. Our analysis considered 51 candidate predictors, including demographics, clinical history, physical examination, and X-ray evaluations. An online tool predicting esKOA progression, based on models with ten and nine predictors for the right and left knees, respectively, was developed.

Results: External validation (MOST) for the right knee at 2.5 years yielded an Area Under Curve (AUC) of 0.847 (95 % CI 0.811 to 0.882), and at 5 years, 0.853 (95 % CI 0.823 to 0.881); for the left knee at 2.5 years, AUC was 0.824 (95 % CI 0.782 to 0.857), and at 5 years, 0.807 (95 % CI 0.768 to 0.843). Models with fewer predictors demonstrated comparable performance. The online tool is available at: https://eskoa.shinyapps.io/webapp/.

Conclusion: Our study unveils a robust, externally validated machine learning tool proficient in predicting the onset of esKOA over the next 2 to 5 years. Our tool can lead to more efficient KOA trials.

Introduction

Knee Osteoarthritis (KOA) is a prevalent musculoskeletal condition, affecting an estimated 654 million individuals over the age of 40 worldwide in 2020 [1]. This condition significantly impacts those affected, leading to substantial disability [2] and imposing considerable financial burdens on healthcare systems due to increased healthcare expenses [3,4]. Although various treatments, including total knee replacement (TKR), offer symptomatic relief, the quest for effectively halting, delaying, or reversing the progression of the disease continues [5].

KOA is a heterogeneous disease, characterized by a wide range of pathways and a slow progression course that can span several years, marked by periods of accelerated worsening and stability [5,6]. This variability presents a significant challenge in KOA research, especially in identifying who requires treatment and is likely to show progression. Traditional inclusion criteria in clinical trials are often inadequate in effectively selecting these individuals. As a result, the development of tools that can accurately predict the progression of KOA is critical.

These predictive tools have the potential to significantly enhance the

* Corresponding author at: HUG Av. de Beau-Séjour 26, 1206 Genève Suisse, Switzerland.
E-mail address: Zubeyir.Salis@etu.unige.ch (Z. Salis).

https://doi.org/10.1016/j.semarthrit.2024.152433

Available online 16 March 2024

efficacy of clinical trials by enabling the identification of candidates likely to experience clinical outcomes (e.g., end stage KOA). Ensuring that participants are those most likely to exhibit disease progression potentially reduces the required sample size and cost for those trials. Such precise selection is crucial to accurately assess the impact of various therapies on the trajectory of the disease [6].

Recent years have seen the development of prediction tools, especially with the advancement of machine learning techniques. These machine learning models have been used to predict structural outcomes, including those assessed by radiography (e.g., a decrease in joint space

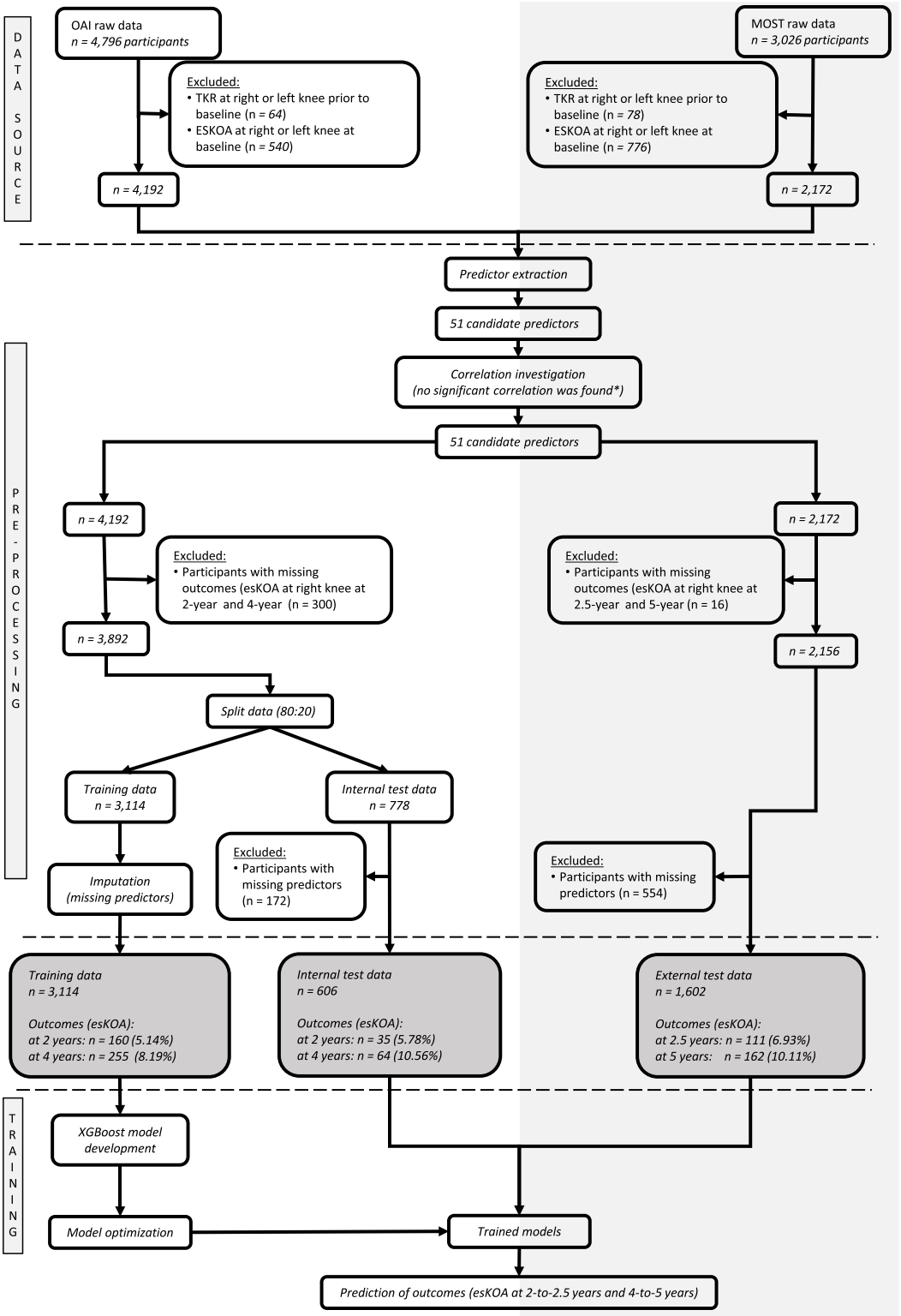

Fig. 1. A summary of the methodology throughout the process of creation of trained models for the prediction of the onset of esKOA at 2-to-2.5 years and 4-to-5 years, for the right knee analysis. The figure shows the data sources, selection of knees, and data pre-preprocessing for the creation of the Training Dataset, Validation Dataset, and External Validation Dataset. The shaded area indicates MOST data. EsKOA: End-stage Knee Osteoarthritis; MOST: Multicenter Osteoarthritis Study; OAI: Osteoarthritis Initiative; TKR: Total Knee Replacement.

[6,7], an increase in Kellgren-Lawrence (KL) grade [8]), and, predominantly, the incidence of TKR [8–15]. While TKR can signify a severe KOA stage, the determinants for receiving a TKR extend beyond disease progression and include an array of extraneous factors (e.g., education, surgery readiness, income, and health insurance) [16,17].

Given the complexities and potential pitfalls of the multifaceted decision-making behind TKR, a more comprehensive measure was sought. End-stage KOA (esKOA) emerged as this solution, integrating both symptomatic and radiographic criteria to robustly signify severe KOA [18,19]. A knee is classified as having esKOA under either of the following conditions: (1) it displays moderate to severe symptoms as assessed by pain and disability measurements, coupled with radiographically confirmed severe KOA; or (2) it exhibits intense symptoms alongside persistent knee pain, with radiography revealing mild to moderate KOA. Importantly, esKOA is not merely an alternative to the incidence of TKR. Instead, it is an outcome that denotes the advanced stage of KOA, uninfluenced by external factors that might drive TKR decisions.

Building on this foundation, our study aimed to develop and validate a machine learning tool to predict the onset of esKOA in 2-to-2.5 years and 4-to-5 years. Such a tool would be pivotal to improving efficiency in KOA clinical trials.

Materials and methods

Data sources

This research utilized data from two prominent multicenter USA-based prospective cohort studies focused on individuals with, or at high risk of, KOA: the Osteoarthritis Initiative (OAI) and the Multicentre Osteoarthritis Study (MOST). The OAI study enrolled 4976 participants aged between 45 and 79 from February 2004 to May 2006 across four clinical sites. The MOST study registered 3026 participants aged between 50 and 79 from April 2003 to April 2005 at two locations. Both studies aimed to recruit individuals with or at risk of symptomatic femoral-tibial KOA. The original OAI and MOST studies received ethical approval from their institutional boards. All participants in the original OAI and MOST studies provided informed written consent.

In this current study, we selected both the right and left knees of participants from the OAI and MOST datasets. We excluded the participants with TKR or esKOA at right or left knees at baseline (Figs. 1 and S1). After these exclusions, we had 4192 participants in the OAI cohort and 2172 participants in the MOST cohort.

Definition of esKOA

We used an esKOA definition developed and validated by Driban et al. [19], which was based on earlier works that determined criteria for TKR appropriateness by Escobar et al. [20] and then by Riddle et al. [21]. Driban et al. adapted these criteria to represent esKOA [18]. In their criteria for esKOA [18], Driban et al. defined and validated an esKOA definition that consists of pain and functional limitations, structural alterations of the knee joint assessed by radiography, or other clinical factors such as knee range of motion and instability. In subsequent research [19], Driban et al. eliminated the assessments for knee range of motion and instability from their earlier definition of esKOA [18]. Furthermore, using the OAI data, Driban et al. demonstrated that esKOA and changes in esKOA predict future TKR [22]. We adopted that later esKOA definition by Driban et al. [19], which integrates (1) the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) scores [23], evaluating KOA symptoms, (2) self-reported persistent knee pain, and (3) the KL grade [24], a radiographic assessment of knee structure. The esKOA definition uses the thresholds for interpreting knee symptoms based on Riddle et al. [21], classifying knee symptoms into four categories based on an aggregate score of WOMAC pain and function (aggregate score range=0 to 88): mild (< 11), moderate (12 to 22), severe (23 to 33), and intense (> 33) symptoms. Persistent knee pain in the esKOA definition was defined as frequent knee pain for at least three months in the past year, with knee symptoms on most days of the month. As the MOST dataset does not have pain chronicity information, we modified the definition of persistent knee pain in our current study as frequent knee pain for at least one or more months in the past year, with knee symptoms on most days of the month. Radiographic assessments in the esKOA definition were based on weight-bearing, bilateral, fixed-flexion, and posterior-anterior knee radiographs and were scored for KL grades (0 to 4) [24]. The esKOA definition used in this study was validated in a separate study that investigated the relationship between weight loss and esKOA, also using OAI and MOST data [25]. In that study [25], our esKOA definition was also shown to be a strong predictor of future TKR.

A knee was classified as having esKOA if it met either of the following criteria: (1) Displaying moderate to severe KOA symptoms (defined as a combined WOMAC pain and disability score of 12 or above) in conjunction with the most severe radiographic KOA (i.e., KL grade $= 4$, the maximum KL grade); (2) Exhibiting intense KOA symptoms (a combined WOMAC pain and disability score of 23 or more) alongside persistent knee pain and either mild or moderate radiographic KOA (i.e., KL grade $= 2$ or 3).

Outcomes and follow-up

Our primary outcomes were the occurrence of esKOA in the right knee and left knee (assessed separately) between the baseline and two distinct follow-up periods. The first follow-up was set at 2 years for the OAI dataset and 2.5 years for the MOST dataset. We refer to this follow-up time point as a '2-to-2.5-year' follow-up. The second follow-up occurred at 4 years for OAI and 5 years for MOST. We refer to this follow-up time point as a '4-to-5-year' follow-up. The outcomes were binary, indicating 'yes' or 'no' for any esKOA occurrence in the right knee and left knee during these periods.

It should be noted that based on the esKOA definition provided, an esKOA status might improve in later assessments due to symptom alleviation. However, in our study, if a knee was determined to have esKOA at any point between the baseline and a follow-up, we retained that classification for all future follow-ups. This approach ensured consistent and clear monitoring of esKOA occurrences throughout the study duration, not just those observed at specific follow-up periods (i.e., at 2-to-2.5-year and 4-to-5-year follow-up).

Data-preprocessing

We performed data preprocessing and subsequent analyses separately for the right and left knees. Fig. 1 depicts the summary of the methodology for creating trained models to predict the onset of esKOA at 2-to-2.5 years and 4-to-5 years for the right knee. Fig. S1 provides a corresponding summary for the left knee analysis.

We used 51 predictors, already identified in a previous study that investigated the prediction of TKR using the OAI and MOST data [12]. These 51 predictors were systematically organized into four domains: demography, intervention history, medical history, and radiographic assessment. Of the 51 predictors, six were related to radiographic assessment across both the right and left knee: (1) KL grade, (2) joint space narrowing (JSN) grade at the lateral tibiofemoral compartment, and (3) JSN grade at the medial compartment, assessed separately for each knee. While we used the right and left knee information for predictors, for limited activity predictor (limited activities due to pain, aching, or stiffness, past 30 days), we used the information for either knee as specific right or left knee information was not commonly available in both cohorts. Additionally, the predictor of steroid injection history was available for 12 months in OAI and 6 months in MOST. We investigated any correlation between the predictors, using the absolute correlation value 0.75 as the threshold. No correlation between the predictors was identified.

After establishing the 51 predictors, we removed the participants

with missing outcomes from the OAI and MOST datasets (Figs. 1 and S1). With that, we had 3892 right knees in the OAI dataset and 2156 in the MOST dataset (Fig. 1); and 3892 left knees in the OAI dataset and 2153 in the MOST dataset (Fig. S1). Additionally, we partitioned the OAI dataset, reserving 80 % for model development and optimization (n = 3114 for the right and left knee) and the remaining 20 % for validation (n = 778 for the right and left knee). In the split process, we used random stratification to ensure a balanced representation of positive (presence of esKOA) and negative (absence of esKOA) cases. The 80 % set aside for model development and optimization established our 'Training Dataset' (n = 3114 for the right and left knee) (Figs. 1 and S1). In the Training Dataset, missing continuous predictors were imputed using mean values, while categorical predictors used mode values. We excluded the participants with missing predictors for the 20 % set aside for validation (n = 172 for right knee and n = 158 for left knee) (Figs. 1 and S1). The remaining 606 right knees established our 'Validation Dataset' for right knee analysis (Fig. 1) and 620 left knees established our 'Validation Dataset' for left knee analysis (Fig. S1). The MOST dataset was exclusively used for external validation. From the MOST dataset, we excluded the participants with missing predictors (n = 554 for right knee and n = 551 for left knee). The remaining 1602 right knees in the MOST dataset established our 'External Validation Dataset' for right knee analysis (Fig. 1) and 1602 left knees established our 'External Validation Dataset' for left knee analysis (Fig. S1).

Machine learning model development and training

Model configuration and optimization

A supervised machine learning model, eXtreme Gradient Boosting (XGBoost) [26], was employed to predict the outcome of esKOA at 2-to-2.5-year and 4-to-5-year time points, for the right and left knees separately. The performances of the models were refined by adjusting a range of tuneable parameters and hyperparameters. Specifically, the maximum tree depth was varied over 3, 5, 7, and 9 values. The number of boosting rounds ranged from 50 to 500, in increments of 50. The learning rate was tested at 0.2, 0.3, and 0.4. The gamma, setting the minimum loss reduction required for further partitioning, was considered at levels 0, 1, and 2. The subsample ratio of columns when constructing each tree and subsample (the fraction of training samples used in any boosting round) were explored over values of 0.5, 0.8, and 1. Lastly, the minimum sum of instance weight needed in a child was adjusted over 1, 3, and 5 values. The model optimizations were performed using the Training Datasets for each knee.

Evaluation metrics

The performances of the models for each knee across the Training Dataset, Validation Dataset, and External Validation Dataset were assessed using the area under the receiver operating characteristic (ROC) curve (i.e., the area under curve [AUC]) for discrimination. An AUC exceeding 0.7 was deemed to offer clinically satisfactory performance [27]. Precision-Recall F measures (F1-scores) were computed as a harmonic mean of the precision and sensitivity, indicating positive predictive power.

We have extracted and reported the key predictors from the models. One of the advantageous attributes of XGBoost is its inherent ability to rank the importance of predictors within the training dataset. Predictor importance in XGBoost provides a score indicating how useful or valuable each predictor was in constructing the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. The resulting importance scores reflect the contribution of each predictor to the model, allowing us to rank and understand which predictors were most influential in predicting esKOA at the 2-to-2.5-year and 4-to-5-year models.

Model calibration

We set an 'optimal threshold' for our models, essentially a fine-tuned setting that helps us make the most accurate predictions possible. This optimal threshold helps us distinguish between positive and negative cases. Finding this 'sweet spot' is especially important because the number of positive and negative cases can vary greatly across different datasets. We used the F1-score as our guide to ensure the tool makes the best positive predictions. We determined this optimal threshold by testing it on the Validation Datasets separately for the right and left knees. Once confident in this setting, we applied it to the External Validation Datasets again separately for the right and left knees to confirm that it works effectively in different scenarios.

Models with fewer predictors and online tool

To provide practical benefit, we developed an online tool that predicts the probability of progression to esKOA for the right and left knees in 2-to-2.5 years and 4-to-5 years. For the online tool, we used models with fewer predictor variables. For that purpose, we selected the predictors above 4% on their importance in the models with 51 predictors.

Simulated trial participant selection and comparison against conventional trial participant selection

We conducted a simulation study to compare the selection of participants based on machine learning models and conventional selection for a trial. We only focused on the right knees for the simulated study. For the simulation of the conventional trial participant selection, we specified that a right knee must meet the following three conditions: (1) A modified version of the American College of Rheumatology (ACR) clinical classification criteria for knee OA [28]; (2) KL grade of 1, 2, or 3; and WOMAC pain score of 5 or more (on a scale of 20). The modified version of the ACR criteria was as follows: any pain in the right knee in the past 12 months, age over 50 years or older, or the presence of morning stiffness or osteophytes. We used the models with fewer predictors to simulate the selection by machine learning. For data in simulation, we used the OAI and MOST datasets, which were used to develop the models with fewer predictors. We reported the performance metrics of both selection methods (i.e., using machine learning models and conventional selection) for comparison.

Statistical analysis and software

We used STATA/BE 18.0 for Windows (64-bit x86–64) software for data preparation. All machine learning analysis, online tool development, and simulated trial selections were performed using R (version 4.3.1).

95 % Confidence Intervals (CI) for AUC were calculated using bootstrap resampling. We set the number of bootstrap samples as 1000. We used stratified bootstrapping to ensure that the proportion of the different classes (e.g., positive and negative cases) in each resampled dataset mirrors that in the original datasets.

We used the following packages in R: 1) For XGBoost, we used XGBoost (version 1.7.5.1); 2) for ROC curves, pROC (version 1.18.4), and 3) for correlations heatmap corrplot (version 0.92). The key predictors were derived using the varImp function from the caret package (version 6.0–94) in R by scaling to 100 (the original scale is 1). For the development of the online tool, we used the shiny (version 1.7.5) and shiny.js (version 2.1.0) packages.

Tool, data, and code availability

The online tool can be found at the following link: https://eskoa.shinyapps.io/webapp/. The datasets were derived from sources in the public domain: OAI public use data sets are available through the National Institute of Mental Health Data Archive, and MOST public use data sets are available through the NIA Aging Research Biobank. The R code for the models creation can be found at https://github.com/Zube-Geneve/predict_esKOA.git.

Results

Data distribution across the training dataset, validation dataset, and external validation dataset

Table 1 presents the distribution of predictors and outcomes for the Training Dataset, Validation Dataset, and External Validation Datasets for right knee, while Table S1 provides the corresponding information for the left knee. On average, participants fell into the overweight category for BMI and were predominantly female, white, married, employed, with no signs of KOA (KL grade 0) at baseline, in datasets for the right and left knees.

From the training data (OAI dataset) consisting of 3114 participants for each of the right and left knees, 160 participants (5.1 %) developed esKOA within 2 years and 255 participants (8.2 %) within 4 years for the right knee (Table 1). Similarly, for the left knee; 161 participants (5.2 %) developed esKOA within 2 years and 257 participants (8.3 %) within 4 years (Table S1). The Validation Dataset from OAI for the right knee included 606 participants, with 35 (5.8 %) developing esKOA within 2 years and 64 (10.6 %) within 4 years (Table 1). The left knee Validation Dataset: with 620 participants, showed 31 (5.0 %) developing esKOA within 2 years and 56 (9.0 %) within 4 years (Table S1), indicating similar trends to the right knee. In the External Validation Datasets from MOST, with 1602 participants for each knee, 111 (6.9 %) developed esKOA within 2.5 years and 162 (10.1 %) within 5 years for the right knee (Table 1). For the left knee, 99 (6.2 %) developed esKOA within 2.5 years and 157 (9.8 %) within 5 years (Table S1).

Model metrics from the training datasets

Right knee

Our model for the right knee had an AUC of 0.996 (95 % Confidence Intervals [CI]s 0.992 to 0.999) using the Training Dataset at 2 years. The optimized parameters were as follows: maximum tree depth: 9; the number of boosting rounds: 50; the learning rate 0.2; the gamma setting: 2; the subsample ratio of columns: 0.8; subsample: 0.5; and the minimum child weight: 1.

Our model at 4 years had an AUC of 0.959 (95 % CIs 0.949 to 0.968). The optimized parameters were as follows: maximum tree depth: 3; the number of boosting rounds: 50; the learning rate 0.2; the gamma setting: 1; the subsample ratio of columns: 1; subsample: 1; and the minimum child weight: 3.

Left knee

Our model for the left knee had an AUC of 0.973 (95 % CIs 0.964 to 0.982) using the Training Dataset at 2 years. The optimized parameters were as follows: maximum tree depth: 9; the number of boosting rounds: 50; the learning rate 0.2; the gamma setting: 0; the subsample ratio of columns: 0.5; subsample: 0.5; and the minimum child weight: 5.

Our model at 4 years had an AUC of 0.939 (95 % CIs 0.925 to 0.951). The optimized parameters were as follows: maximum tree depth: 3; the number of boosting rounds: 50; the learning rate 0.2; the gamma setting: 2; the subsample ratio of columns: 0.5; subsample: 0.5; and the minimum child weight: 5.

Performance of model with the Validation Datasets

Right knee

We determined the optimal threshold values from the Validation Dataset as 0.029 and 0.045 for the 2-to-2.5-year and 4-to-5-year models, respectively. Our model had an AUC of 0.894 (95 % CIs 0.851 to 0.933) at 2 years (Fig. 2a) and an AUC of 0.873 (95 % CIs 0.834 to 0.906) at 4 years (Fig. 2b). The F1 scores for the positive class were 0.898 at 2 years (Fig. 2a) and 0.849 at 4 years (Fig. 2b).

Table 1
Data distribution of predictors and outcomes in the Training Dataset, Validation Dataset, and External Validation Dataset (right knee analysis).

Predictor/outcome	Training data	Validation dataset	External Validation Dataset
Participants	***n* = 3114**	***n* = 606**	***n* = 1602**
Demographics	n (%) or Mean ± SD	n (%) or Mean ± SD	n (%) or Mean ± SD
Age, years	61.1 ± 9.2	60.3 ± 9.2	61.4 ± 7.9
Blood pressure			
Normal	1081 (34.7)	220 (36.3)	620 (38.7)
Elevated	479 (15.4)	96 (15.8)	289 (18.0)
Hypertension Stage 1	940 (30.2)	183 (30.2)	391 (24.4)
Hypertension Stage 2	600 (19.3)	107 (17.7)	294 (18.4)
Hypertensive crisis	14 (0.4)	0 (0.0)	8 (0.5)
Body Mass Index (BMI), kg/m^2	28.2 ± 4.6	27.9 ± 4.6	29.5 ± 5.1
Education			
Less than high school	72 (2.3)	16 (2.6)	34 (2.1)
High School	357 (11.5)	59 (9.7)	326 (20.3)
College/associate degree / technical school after high school	692 (22.2)	142 (23.4)	346 (21.6)
College Graduate	687 (22.1)	133 (21.9)	411 (25.7)
Some graduate degree	264 (8.5)	52 (8.6)	148 (9.2)
Graduate school	1042 (33.5)	204 (33.7)	337 (21.0)
Employment status			
Works for pay	1924 (61.8)	401 (66.2)	981 (61.2)
Ethnicity			
White/Caucasian	2595 (83.3)	498 (82.2)	1399 (87.3)
Black/African American	434 (13.9)	92 (15.2)	176 (11.0)
Hispanic/Latino	32 (1.0)	5 (0.8)	9 (0.6)
Other	53 (1.7)	11 (1.8)	18 (1.1)
Living status alone/with others			
Live with other(s)	2447 (78.6)	493 (81.4)	1353 (84.5)
Marital status			
Married	2142 (68.8)	441 (72.8)	1243 (77.6)
Widowed	238 (7.6)	35 (5.8)	114 (7.1)
Divorced	409 (13.1)	82 (13.5)	174 (10.9)
Separated	48 (1.5)	8 (1.3)	5 (0.3)
Never married	277 (8.9)	40 (6.6)	66 (4.1)
Gender			
Female	1801 (57.8)	337 (55.6)	905 (56.5)
Smoking, pack years	9.0 ± 15.1	8.8 ± 17.6	8.8 ± 16.3
History of intervention			
History of intervention - medication			
Analgesic medication (Salicylates, NSAIDs, COX2, Opioids, other)	717 (23.0)	131 (21.6)	1267 (79.1)
Arthritis medication (Oral corticosteroids, supplements (SAMe, MSM, Fluorides, Glucosamine))	217 (7.0)	62 (10.2)	479 (29.9)
Osteoporosis medication (Vitamin, Bisphosphonate, Estrogen, Raloxifene, Calcitonin, Teriparatide)	432 (13.9)	78 (12.9)	510 (31.8)
Steroid injection in right knee (in past 12 months in OAI, in 6 months in MOST)	27 (0.9)	3 (0.5)	21 (1.3)
Steroid injection in left knee (in past 12 months in OAI, in 6 months in MOST)	25 (0.8)	3 (0.5)	16 (1.0)
History of intervention – knee-related surgery			
Arthroscopy, ever, right knee	277 (8.9)	58 (9.6)	111 (6.9)
Arthroscopy, ever, left knee	254 (8.2)	56 (9.2)	110 (6.9)
Ligament repair surgery, right knee	41 (1.3)	9 (1.5)	21 (1.3)
Ligament repair surgery, left knee	51 (1.6)	15 (2.5)	23 (1.4)
Meniscectomy, ever, right knee	244 (7.8)	52 (8.6)	106 (6.6)
Meniscectomy, ever, left knee	220 (7.1)	49 (8.1)	97 (6.1)
Other kind of surgery, ever, right knee	42 (1.3)	10 (1.7)	20 (1.2)

(*continued on next page*)

Table 1 (*continued*)

Predictor/outcome	Training data	Validation dataset	External Validation Dataset
Participants	***n* = 3114**	***n* = 606**	***n* = 1602**
Other kind of surgery, ever, left knee	46 (1.5)	13 (2.1)	17 (1.1)
Medical history			
Medical history – arthritis specific			
Arthritis past medical history			
No arthritis history	1662 (53.4)	319 (52.6)	796 (49.7)
At least one OA/degenerative disease	1120 (36.0)	218 (36.0)	467 (29.2)
Gout/other	129 (4.1)	30 (5.0)	91 (5.7)
OA/degenerative disease and gout/other	90 (2.9)	15 (2.5)	70 (4.4)
Unknown	113 (3.6)	24 (4.0)	178 (11.1)
Injury (Right knee, ever injured badly enough to limit the ability to walk for at least two days)	836 (26.8)	175 (28.9)	402 (25.1)
Injury (Left knee, ever injured badly enough to limit the ability to walk for at least two days)	741 (23.8)	155 (25.6)	324 (20.2)
Limited activity (Either knee, limited activities due to pain, aching, or stiffness, past 30 days)	2434 (78.2)	454 (74.9)	1356 (84.6)
Symptoms (Right knee, pain, aching or stiffness, ever had more than half the days of a month)	1898 (61.0)	336 (55.4)	1151 (71.8)
Symptoms (Left knee, pain, aching or stiffness, ever had more than half the days of a month)	1893 (60.8)	360 (59.4)	1186 (74.0)
Medical history – clinical examination			
Clinic 20-meter walk assessment	15.5 ± 2.9	15.4 ± 3.0	16.3 ± 2.9
Timed chair stands (Potential risk)	2039 (65.5)	382 (63.0)	896 (55.9)
Medical history - comorbidities			
Asthma	257 (8.3)	47 (7.8)	114 (7.1)
Diabetes	200 (6.4)	37 (6.1)	124 (7.7)
Emphysema, Chronic Obstructive Pulmonary Disease (COPD), chronic bronchitis	58 (1.9)	15 (2.5)	50 (3.1)
Heart attack	59 (1.9)	8 (1.3)	47 (2.9)
Heart failure	53 (1.7)	9 (1.5)	38 (2.4)
Kidney problems	33 (1.1)	13 (2.1)	68 (4.2)
Stomach ulcer	69 (2.2)	10 (1.7)	63 (3.9)
Stroke	83 (2.7)	18 (3.0)	57 (3.6)
Medical history – questionnaires			
Center for Epidemiological Studies Depression (CESD) Score	248 (8.0)	49 (8.1)	137 (8.6)
Physical Activity Scale for the Elderly Score (PASE)	162.7 ± 80.6	169.5 ± 86.2	184.9 ± 88.6
Short-Form 12 Mental Component (SF12mental)	53.8 ± 7.6	53.8 ± 7.6	54.2 ± 8.4
Short-Form 12 Physical Component (SF12physical)	50.6 ± 8.0	51.0 ± 7.2	48.7 ± 9.4
Total Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) right knee	8.5 ± 10.8	8.9 ± 11.2	12.6 ± 12.9
Total Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) left knee	8.2 ± 11.8	8.2 ± 11.6	12.4 ± 12.8
Radiography			
Kellgren Lawrence (KL) grade right knee			
0	1393 (44.7)	270 (44.6)	890 (55.6)
1	615 (19.7)	112 (18.5)	310 (19.4)
2	702 (22.5)	144 (23.8)	219 (13.7)
3	348 (11.2)	66 (10.9)	170 (10.6)
4	56 (1.8)	14 (2.3)	13 (0.8)
Kellgren Lawrence (KL) grade left knee			
0	1338 (43.0)	250 (41.3)	963 (60.1)
1	615 (19.7)	110 (18.2)	287 (17.9)
2	755 (24.2)	152 (25.1)	197 (12.3)
3	335 (10.8)	76 (12.5)	139 (8.7)
4	71 (2.3)	18 (3.0)	16 (1.0)
Joint space narrowing (JSN) grade right knee-lateral			
0	2913 (93.5)	559 (92.2)	1508 (94.1)
1	121 (3.9)	29 (4.8)	53 (3.3)
2	62 (2.0)	16 (2.6)	36 (2.2)
3	18 (0.6)	2 (0.3)	5 (0.3)
Joint space narrowing (JSN) grade left knee-lateral			
0	2903 (93.2)	552 (91.1)	1535 (95.8)
1	104 (3.3)	28 (4.6)	42 (2.6)
2	82 (2.6)	20 (3.3)	21 (1.3)
3	25 (0.8)	6 (1.0)	4 (0.2)
Joint space narrowing (JSN) grade right knee-medial			
0	2177 (69.9)	416 (68.6)	1192 (74.4)
1	605 (19.4)	126 (20.8)	267 (16.7)
2	293 (9.4)	52 (8.6)	133 (8.3)
3	39 (1.3)	12 (2.0)	10 (0.6)
Joint space narrowing (JSN) grade left knee-medial			
0	2111 (67.8)	402 (66.3)	1231 (76.8)
1	696 (22.4)	134 (22.1)	236 (14.7)
2	261 (8.4)	58 (9.6)	121 (7.6)
3	46 (1.5)	12 (2.0)	14 (0.9)
Outcomes			
End-stage Knee Osteoarthritis within 2-to-2.5 years (right knee)	160 (5.1)	35 (5.8)	111 (6.9)
End-stage Knee Osteoarthritis within 4-to-5 years (right knee)	255 (8.2)	64 (10.6)	162 (10.1)

Data are presented as mean ± standard deviation or count (percentage).

Left knee

We determined the optimal threshold values from the Validation Dataset as 0.013 and 0.052 for the 2-to-2.5-year and 4-to-5-year models, respectively. Our model had an AUC of 0.841 (95 % CIs 0.767 to 0.903) at 2 years (Fig. S2a) and an AUC of 0.862 (95 % CIs 0.817 to 0.901) at 4 years (Fig. S2b). The F1 scores for the positive class were 0.758 at 2 years (Fig. S2a) and 0.857 at 4 years (Fig. S2b).

Performance of model with External Validation Dataset

Right knee

Performances on the External Validation Dataset for the right knee, using the threshold values obtained from the Validation Dataset, showed that our model yielded an AUC of 0.847 (95 % CIs 0.811 to 0.882) at 2.5 years (Fig. 2c) and 0.853 (95 % CIs 0.823 to 0.881) at 5 years (Fig. 2d). The F1 scores for the positive class were 0.896 at 2.5 years (Fig. 2c) and 0.851 at 5 years (Fig. 2d).

Left knee

Performances on the External Validation Dataset for the left knee, using the threshold values obtained from the Validation Dataset, showed

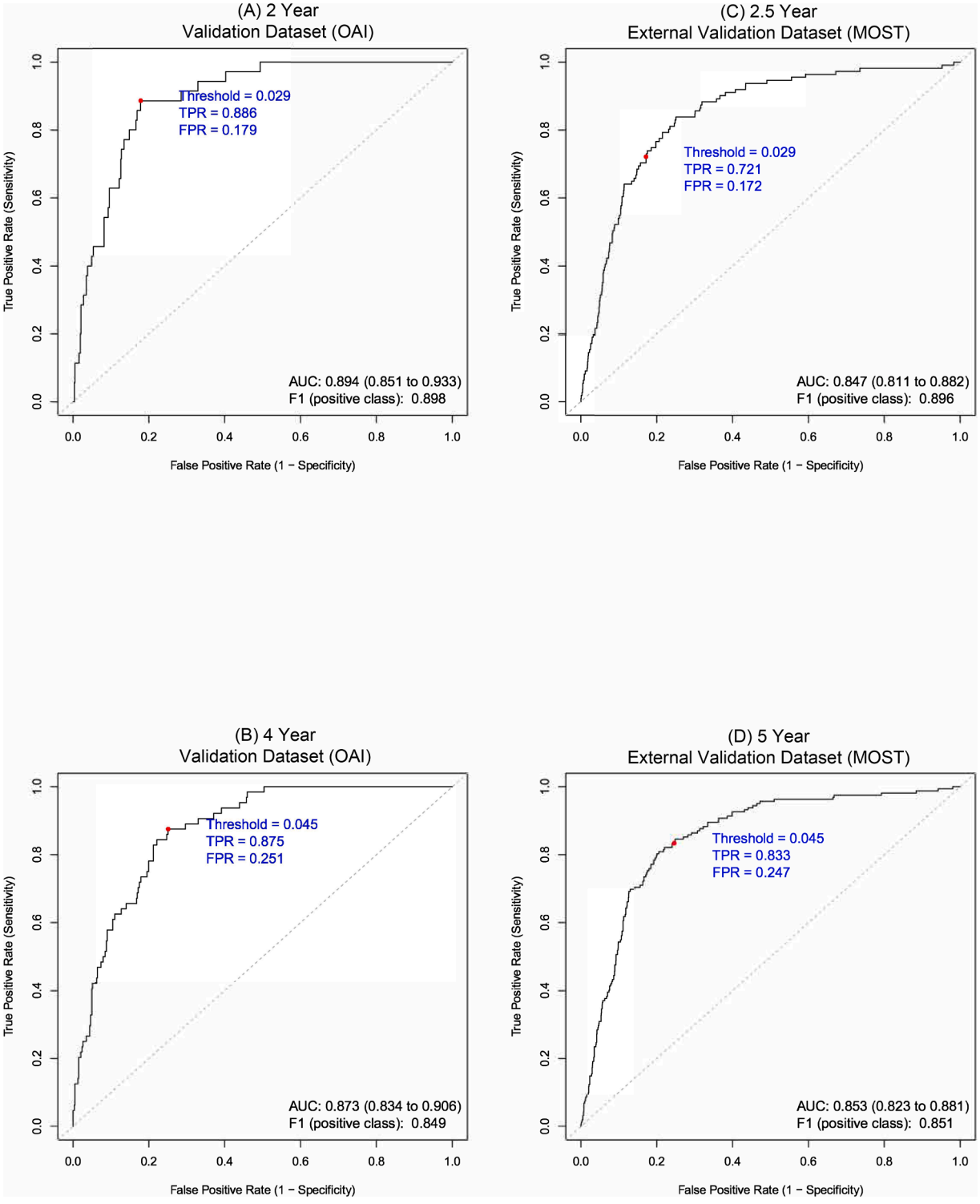

Fig. 2. Receiver operating characteristic (ROC) curves showing performance of models for prediction of esKOA at 2-to-2.5 years and 4-to-5 years, using 51 predictors, for the right knee. (A) Validation Dataset (OAI) at 2 years. (B) Validation Dataset (OAI) at 4 years. (C) External Validation Dataset (MOST) at 2.5 years. (D) External Validation Dataset (MOST) at 5 years. Red Points show the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for the selected thresholds. AUC: Area Under Curve; MOST: Multicenter Osteoarthritis Study; OAI: Osteoarthritis Initiative.

that our model yielded an AUC of 0.824 (95 % CIs 0.782 to 0.857) at 2.5 years (Fig. S2c) and 0.807 (95 % CIs 0.768 to 0.843) at 5 years (Fig. S2d). The F1 scores for thepositive class were 0.756 at 2.5 years (Fig. S2c) and 0.877 at 5 years (Fig. S2d).

Key predictors

Using the XGBoost algorithm, we ranked predictors based on their relative influence in predicting esKOA at 2-to-2.5 years and 4-to-5 years. The relative influence scores are scaled up to 100 for ease of interpretation. The relative influence scores indicate the contribution of each

predictor to the model. The higher the percentage score, the more impactful the predictor is in determining the outcome of esKOA. Table 2 lists the predictors with an importance score greater than zero for the 2-to-2.5-year and 4-to-5-year models for the right knee, while Table S2 presents the corresponding predictors for the left knee. Our top three predictors for the right knee were KL grade for right knee (relative influence 12.57 % and 35.30 %), Short-Form 12 (SF12) physical score (11.84 % and 8.85 %), and total WOMAC score for right knee (11.73 % and 18.82 %) for the 2-to-2.5-year model and 4-to-5-year model, respectively (Table 2). Our top three predictors for the left knee were total WOMAC score for left knee (relative influence 13.94 % and 20.08 %), KL grade for left knee (13.48 % and 20.08 %), and total WOMAC score for right knee (11.47 % and 11.70 %) for the 2-to-2.5-year model and 4-to-5-year model, respectively. (Table S2).

We have selected the predictor variables with an importance score greater than four percent whether they are for the 2-to-2.5-year and 4-to-5-year models for the online tool. For the right knee, there were ten predictors. Five of these predictors (KL grade right knee, short-form 12 physical component score, total WOMAC score right knee, clinic 20-meter walk assessment, and total WOMAC score left knee) were common to both the 2-to-2.5-year and 4-to-5-year models, four predictors (Physical Activity Scale for the Elderly Score (PASE) Body Mass Index (BMI), short-form 12 mental component score, and age) were unique to the 2-to-2.5-year model, and one predictor (KL grade left knee) was specific to the 4-to-5-year model (Table 2). For the left knee, there were nine predictors. Six of these predictors (total WOMAC left knee, KL grade left knee, total WOMAC right knee, short-form 12 physical component score, BMI, and PASE) were common to both the 2-to-2.5-year and 4-to-5-year models, two predictors (clinic 20-meter walk assessment and age) were unique to the 2-to-2.5-year model, and one predictor (KL grade right knee) was specific to the 4-to-5-year model (Table S2). Using these fewer predictor variables, we have trained our models and assessed their performance. Given that their performance was comparable to the models with 51 predictors, we used the models with fewer predictors to develop an online tool.

Performance of the models with fewer predictors

Right knee

The performance metrics obtained for our models with ten predictors were very close to those obtained from our model with 51 above-mentioned predictors. Our model with ten predictors had an AUC of 0.959 (95 % CIs 0.948 to 0.969) using the Training Dataset at 2 years. The optimized parameters were as follows: maximum tree depth: 3; the number of boosting rounds: 50; the learning rate 0.2; the gamma setting: 2; the subsample ratio of columns: 0.5; subsample: 1; and the minimum child weight: 5. The model at 4 years had an AUC of 0.945 (95 % CIs 0.932 to 0.957) at 4 years. The optimized parameters were as follows: maximum tree depth: 3; the number of boosting rounds: 50; the learning rate 0.2; the gamma setting: 2; the subsample ratio of columns: 0.8; subsample: 0.8; and the minimum child weight: 5.

When applied to the Validation Dataset, we determined the optimal threshold values from the Validation Dataset as 0.040 and 0.073 for the 2-to-2.5-year and 4-to-5-year models, respectively. Our model with ten predictors had an AUC of 0.887 (95 % CIs 0.848 to 0.923) at 2 years (Fig. 3a) and an AUC of 0.869 (95 % CIs 0.835 to 0.903) at 4 years (Fig. 3b). The F1 scores for the positive class were 0.872 at 2 years and 0.873 at 4 years.

Performances on the External Validation Dataset, using the threshold values obtained from the Validation Dataset, showed that our model with ten predictors yielded an AUC of 0.847 (95 % CIs 0.813 to 0.876) at 2.5 years (Fig. 3c) and 0.848 (95 % CIs 0.822 to 0.874) at 5 years (Fig. 3d). The F1 scores for the positive class were 0.887 at 2.5 years and 0.882 at 5 years.

Table 2
The predictors and their relative influence at 2-to-2.5 years and 4-to-5 years (right kne analysis).

No	2-to-2.5 years Predictor	Relative Influence (%)	4-to-5 years Predictor	Relative Influence (%)
1	Kellgren Lawrence (KL) grade right knee	12.57	Kellgren Lawrence (KL) grade right knee	35.30
2	Short-Form 12 Physical Component (SF12physical)	11.84	Total Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) right knee	18.82
3	Total Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) right knee	11.73	Short-Form 12 Physical Component (SF12physical)	8.85
4	Clinic 20-meter walk assessment	7.85	Kellgren Lawrence (KL) grade left knee	5.19
5	Total Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) left knee	7.18	Clinic 20-meter walk assessment	4.64
6	Physical Activity Scale for the Elderly Score (PASE)	5.83	Total Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) left knee	4.57
7	Body Mass Index (BMI)	5.82	Body Mass Index (BMI)	3.87
8	Short-Form 12 Mental Component (SF12mental)	5.14	Short-Form 12 Mental Component (SF12mental)	3.62
9	Age	4.23	Symptoms (Left knee, pain, aching or stiffness, ever had more than half the days of a month)	1.86
10	Smoking, pack years	3.24	Physical Activity Scale for the Elderly Score (PASE)	1.80
11	Kellgren Lawrence (KL) grade left knee	2.90	Smoking, pack years	1.52
12	Joint space narrowing (JSN) grade right knee-medial	2.71	Age	1.52
13	Education status	2.37	Center for Epidemiological Studies Depression (CESD) Score	1.12
14	Arthritis past medical history	2.22	Arthritis past medical history	1.05
15	Joint space narrowing (JSN) grade left knee-medial	1.52	Education status	0.91
16	Analgesic medication (Salicylates, NSAIDs, COX2, Opioids, other)	1.32	Meniscectomy, ever, right knee	0.79
17	Symptoms (Left knee, pain, aching or stiffness, ever had more than half the days of a month)	1.19	Blood pressure	0.69
18	Timed chair stands (Potential risk)	1.11	Joint space narrowing (JSN) grade right knee-medial	0.67
19	Gender	1.03	Joint space narrowing (JSN) grade right knee-lateral	0.65
20	Symptoms (Right knee, pain, aching or stiffness, ever had more than half the days of a month)	0.97	Living status alone/ with others	0.49

(*continued on next page*)

Table 2 (*continued*)

No	2-to-2.5 years Predictor	Relative Influence (%)	4-to-5 years Predictor	Relative Influence (%)
21	Blood pressure	0.91	Analgesic medication (Salicylates, NSAIDs, COX2, Opioids, other)	0.43
22	Joint space narrowing (JSN) grade left knee-lateral	0.74	Symptoms (Right knee, pain, aching or stiffness, ever had more than half the days of a month)	0.38
23	Employment status	0.65	Heart attack	0.28
24	Marital status	0.53	Timed chair stands (Potential risk)	0.25
25	Living status alone/ with others	0.51	Meniscectomy, ever, left knee	0.21
26	Heart attack	0.49	Medical history - Asthma	0.19
27	Limited activity (Either knee, limited activities due to pain, aching, or stiffness, past 30 days)	0.47	Gender	0.14
28	Medical history - Asthma	0.47	Employment status	0.11
29	Ethnicity	0.41	Stroke	0.08
30	Injury (Left knee, ever injured badly enough to limit the ability to walk for at least two days)	0.35	–	–
31	Arthroscopy, ever, left knee	0.34	–	–
32	Joint space narrowing (JSN) grade right knee-lateral	0.33	–	–
33	Osteoporosis medication (Vitamin, Bisphosphonate, Estrogen, Raloxifene, Calcitonin, Teriparatide)	0.23	–	–
34	Meniscectomy, ever, right knee	0.22	–	–
35	Kidney problems	0.17	–	–
36	Stroke	0.14	–	–
37	Medical history - Diabetes	0.10	–	–
38	Injury (Right knee, ever injured badly enough to limit the ability to walk for at least two days)	0.10	–	–
39	Center for Epidemiological Studies Depression (CESD) Score	0.08	–	–

Left knee

The performance metrics obtained for our models with nine predictors were very close to those obtained from our model with 51 above-mentioned predictors. Our model with nine predictors had an AUC of 0.951 (95 % CIs 0.935 to 0.964) using the Training Dataset at 2 years. The optimized parameters were as follows: maximum tree depth: 3; the number of boosting rounds: 50; the learning rate 0.2; the gamma setting: 1; the subsample ratio of columns: 1; subsample: 0.5; and the minimum child weight: 5. The model at 4 years had an AUC of 0.942 (95 % CIs 0.928 to 0.954) at 4 years. The optimized parameters were as follows: maximum tree depth: 3; the number of boosting rounds: 50; the learning rate 0.2; the gamma setting: 2; the subsample ratio of columns: 0.5; subsample: 1; and the minimum child weight: 3.

When applied to the Validation Dataset, we determined the optimal threshold values from the Validation Dataset as 0.045 and 0.085 for the 2-to-2.5-year and 4-to-5-year models, respectively. Our model with nine predictors had an AUC of 0.858 (95 % CIs 0.805 to 0.910) at 2 years (Fig. S3a) and an AUC of 0.870 (95 % CIs 0.831 to 0.904) at 4 years (Fig. S3b). The F1 scores for the positive class were 0.887 at 2 years and 0.885 at 4 years.

Performances on the External Validation Dataset, using the threshold values obtained from the Validation Dataset, showed that our model with nine predictors yielded an AUC of 0.822 (95 % CIs 0.781 to 0.860) at 2.5 years (Fig. S3c) and 0.811 (95 % CIs 0.780 to 0.842) at 5 years (Fig. S3d). The F1 scores for the positive class were 0.911 at 2.5 years and 0.890 at 5 years.

Online tool

Using our models with fewer predictors, we developed an online tool that predicts the development of esKOA at 2-to-2.5 years and 4-to-5 years, specifically for either the right or the left knee. On entering the inputs for those predictors, the tool then provides the probabilities of developing esKOA as a percentage and binary (i.e., high risk or low risk). The tool also allows to change the default threshold values (i.e., 0.040 and 0.073 for the 2-to-2.5-year and 4-to-5-year models for the right knee, respectively, and 0.045 and 0.085 for the 2-to-2.5-year and 4-to-5-year models for the left knee, respectively).

Comparison with conventional trial participant selection

Our results indicate that our machine learning models significantly outperform conventional trial participant selection methods in accurately identifying individuals at risk of developing esKOA (Table 3). The true positive rates (TPR) for the machine learning models are 2.7 to 3.6 times higher than those achieved by conventional methods across both the OAI and MOST datasets for the forecast periods of 2-to-2.5 years and 4-to-5 years. For instance, using the machine learning model, 75–90 % of participants who would develop esKOA within 4-to-5 years were correctly identified (based on the TPR), in stark contrast to only 24–25 % identified by conventional selection methods. Furthermore, the machine learning models demonstrated superior performance as evidenced by higher F1 scores, indicating enhanced precision and recall in identifying participants at risk.

Discussion

Our study successfully utilized a machine learning model to predict the onset of esKOA with notable accuracy. With AUCs exceeding the threshold of 0.7 which was deemed to offer clinically satisfactory performance at 2.5 and 5 years, and robust F1 scores at the same intervals, the high predictive capability of our model is evident. Moreover, our models with fewer predictors had similar performance metrics as models with 51 predictors. Except for two radiographic variables (KL grades for the right and left knees), other remaining variables (8 for the right and 7 for the left knees) can be obtained in a clinical setting. Using the models with fewer predictors, we created a practical online tool that predicts the development of esKOA in 2-to-2.5 years and 4-to-5 years. Notably, by moving away from the limitations and complexities of TKR-based decisions, our model introduces a transformative approach to predicting progression to severe KOA. Our simulated participant selection showed that our models were superior in identifying the individuals who will progress to esKOA compared to conventional trial selection. Our online tool (available at: https://eskoa.shinyapps.io/webapp/) could have important implications in future clinical trials by improving their efficiency in selecting participants.

Previously, Dunn et al. [29] developed a risk score algorithm to predict the progression of esKOA using data from the OAI. Their definition of esKOA, which was adopted from Driban et al. [18]. included information on knee range of motion and instability. They achieved a similar AUC (0.87) for both 2 and 4 years as our study, using internal

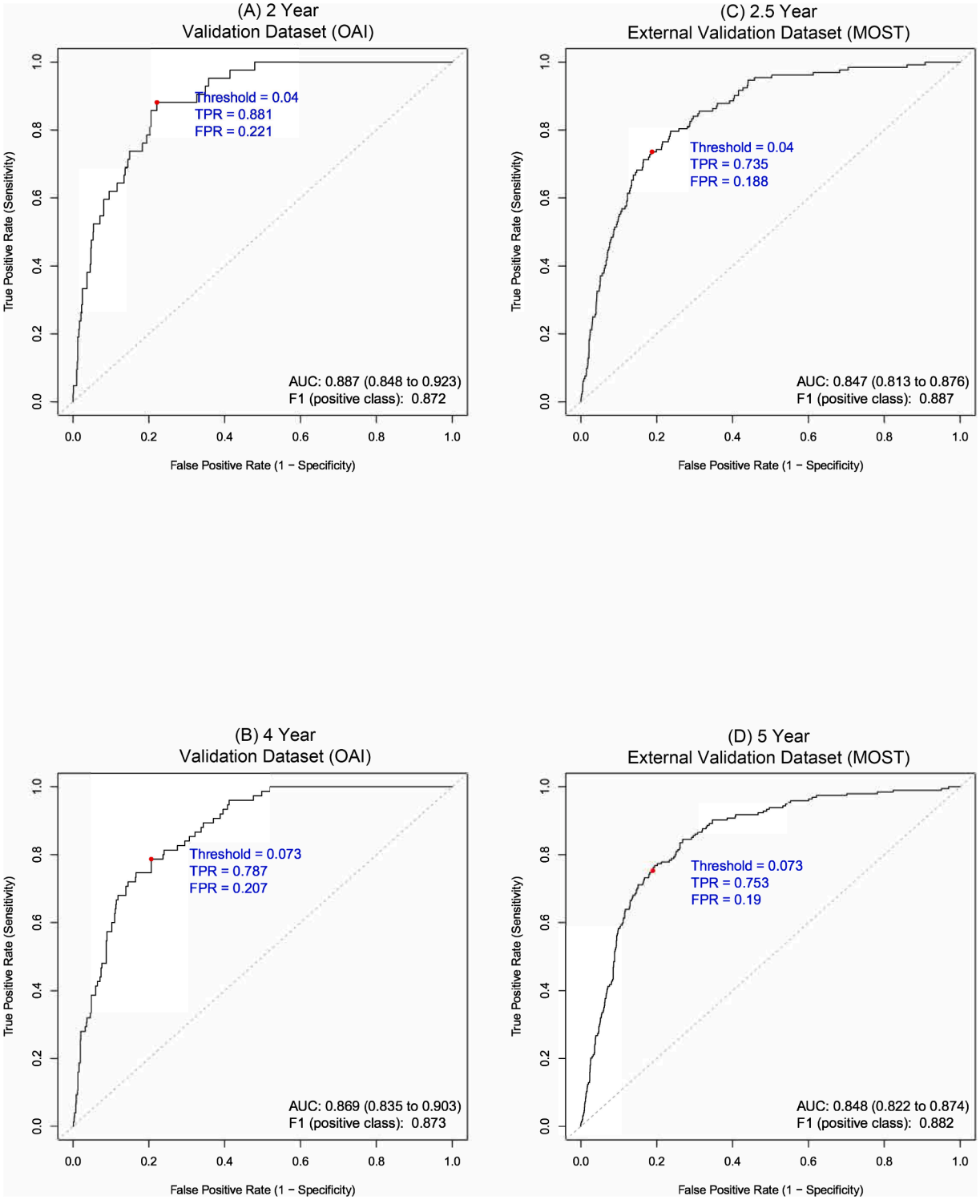

Fig. 3. Receiver operating characteristic (ROC) curves showing the performance of models for prediction of esKOA at 2-to-2.5 years and 4-to-5 years, for the right knee, using ten predictors. (A) Validation Dataset (OAI) at 2 years. (B) Validation Dataset (OAI) at 4 years. (C) External Validation Dataset (MOST) at 2.5 years. (D) External Validation Dataset (MOST) at 5 years. Red Points show the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for the selected thresholds. AUC: Area Under Curve; MOST: Multicenter Osteoarthritis Study; OAI: Osteoarthritis Initiative.

validation data. However, their study lacked external validation, unlike ours, which was validated using the MOST dataset. Additionally, their algorithm depended on knee range of motion and instability information, which OAI does not measure annually. Our study avoids this limitation by not relying on knee range of motion and instability information. Furthermore, while they utilized information from the opposite knee, their methodology was dependent on the selection of a target knee. In contrast, our study evaluates the progression of esKOA in both knees of an individual. To our knowledge, only one other study by Widera et al. [6] developed a model to predict the progression of KOA based on combining radiography and symptom outcomes, as in our study. While their performance, measured using the F1 score, was

Table 3
Simulated trial participant selection.

	2-to-2.5 years		4-to-5 years	
	Conventional	Machine Learning	Conventional	Machine Learning
OAI				
TPR	0.301	0.944	0.248	0.901
TNR	0.942	0.802	0.946	0.807
PPV	0.219	0.207	0.297	0.302
NPV	0.961	0.996	0.931	0.989
F1 Score	0.254	0.339	0.270	0.452
MOST				
TPR	0.272	0.735	0.242	0.753
TNR	0.924	0.812	0.927	0.810
PPV	0.206	0.220	0.269	0.304
NPV	0.946	0.976	0.917	0.967
F1 Score	0.235	0.339	0.255	0.433

The selections are made from the complete datasets of the OAI and MOST, which were also used to develop the models with fewer predictors. TPR: True Positive Rates (proportion of actual positives, i.e., individuals who develop esKOA, that are correctly identified); TNR: True Negative Rates (proportion of actual negatives, i.e., individuals who do not develop esKOA, that are correctly identified as not developing esKOA.); PPV: Positive Predictive Values Rates (proportion of positive identifications, i.e., predicted to develop esKOA, that were actually correct); NPR: Negative Preditive Values Rates (proportion of negative identifications, i.e., predicted not to develop esKOA, that were actually correct);; OAI: Osteoarthritis Initiative; MOST: Multicenter Osteoarthritis Study.

commendable (0.584 with Cohort Hip and Cohort Knee Study dataset and 0.689 with OAI), it still fell short of our results. A critical distinction is their reliance on data collection over extended periods, which introduces practical complexities due to logistical challenges such as limited equipment access and participant availability. In contrast, our esKOA model efficiently predicts progression individually, offering a more streamlined and practical approach. Additionally, the study by Widera et al. [6] applied one-hot encoding (a method of data transformation provided to machine learning algorithms to improve predictions) broadly, even to continuous attributes, which can potentially limit the adaptability of their decision trees to new data. Furthermore, they acknowledged certain issues of feature misuse in their methodology. On the other hand, our esKOA model emphasizes precise data encoding and judicious feature selection, enhancing its precision and potential applicability in clinical scenarios.

Our study differs from previous models that predicted the progression of KOA based on incidence of TKR in several ways. Several studies also investigated the prediction of TKR but were not externally validated and mainly used the OAI dataset [9–11,13,14,22], and some required magnetic resonance images (MRI) data [13]. One notable study by Mahmoud et al. [12] aimed at predicting the need for TKR within 2 and 5 years, using OAI for training and MOST data for external testing, similar to our study. They achieved AUCs of 0.913 and 0.873, respectively, using the Gradient Boosting Machine (GBM) model. While their AUC was higher than ours, their positive predictive ability was lower (F1-score: 0.171 and 0.287 for 2 and 5 years, respectively). Rajamohan et al. [15] also used OAI and MOST datasets for training and validation to predict future TKR, but they require MRI data for prediction.

Our tool has the potential to significantly impact clinical trials in KOA. Our models outperformed conventional selection methods by accurately identifying candidates who would develop esKOA with 2.7 to 3.6 times greater precision. This potentially equates to a reduction in sample sizes, compared to conventional selection methods. This enhanced predictive capability can streamline participant selection for trials focused on disease-modifying drugs or other interventions, ensuring that chosen participants are more likely to exhibit disease progression. Additionally, future studies should explore using this tool as an outcome measure in clinical trials with a shorter duration or participants with milder disease. For instance, the efficacy of a 3-month intervention could be gauged by observing changes in the likelihood of developing esKOA at 2.5 and 5-year intervals, measured both at baseline and trial conclusion. This allows for a quicker, yet still reliable, assessment of intervention efficacy, thereby accelerating clinical research. Given its versatility, our tool has the potential for broad applications in KOA research and clinical practice.

Despite the promising results, our study is not without limitations. The dominant demographic, leaning towards older, female, and white participants, may restrict the generalizability of our findings to a broader populace. Moreover, the United States of America (USA)-centric datasets utilized necessitate further investigations to validate the applicability of our models beyond the USA. We acknowledge the imbalance in our dataset, with fewer instances of esKOA. While we did explore balanced alternatives using techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) methods, we found that the original unbalanced dataset, with its naturally higher sample size, provided better predictive performance for esKOA. Finally, one limitation of our study is the lack of MRI features in our predictive models. Incorporating MRI data, which provides detailed insights into joint structures, could enhance the predictive accuracy of esKOA. We recommend that future studies consider including MRI data to improve the prediction of esKOA development.

In summary, our study unveils a robust, externally validated machine learning tool proficient in predicting the onset of esKOA using transparent and readily available data. Our tool has the potential to improve the efficiency of osteoarthritis trials. Subsequent research could focus on refining these models in diverse global populations.

Role of the funding sources

There were no funders for this study. Therefore, the study is free of funder involvement in study design, data collection, data analysis, data interpretation, or report writing. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

CRediT authorship contribution statement

Zubeyir Salis: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing, Validation. **Jeffrey B. Driban:** Writing – original draft, Writing – review & editing, Validation. **Timothy E. McAlindon:** Writing – original draft, Writing – review & editing, Validation.

Declaration of competing interest

ZS owns 50 % of the shares in Zuman International, which receives royalties and other payments for educational resources and services in adult weight management and research methodology. **Jeffrey B. Driban** declares that he is a consultant for Pfizer Inc. and Eli Lilly and Company and served on an advisory board for Novartis. **Timothy E. McAlindon** is a consultant for Remedium-Bio, Anika, Chemocentryx, Grunenthal, Kolon Tissue Gene, Novartis, BioSplice, Organogenesis, and Pfizer Inc.

Acknowledgments

We acknowledge the provision of datasets and/or research tools from two cohort studies: the Osteoarthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST).

The OAI is a collaborative informatics system created by the National Institute of Mental Health and the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS) to provide a worldwide resource to quicken the pace of biomarker identification, scientific investigation and osteoarthritis drug development. The OAI 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) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation; GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using OAI public-use data sets and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners. The OAI data repository is housed within the National Institute of Mental Health (NIMH) Data Archive (NDA).

For MOST, we wish to acknowledge the contributions of the study participants, investigators and research staff involved. MOST is comprised of four (4) cooperative grants: U01 AG18820 to David T Felson (Boston University); U01 AG18832 to James Torner (University of Iowa); U01 AG18947 to Cora E Lewis (University of Alabama at Birmingham); and U01 AG19069 to Michael C Nevitt (University of California, San Francisco), funded by the National Institutes of Health (NIH), a branch of the Department of Health and Human Services, and conducted by MOST investigators. This manuscript was prepared using MOST data and does not claim, infer, or imply endorsement by MOST, by the MOST investigators and their respective institutions, or by the University of California, of the Data Recipients' (i.e., our) use of the Data, of the entity or personnel conducting the research (i.e., for this paper), or of any results of the research (i.e., of the current study).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.semarthrit.2024.152433.

References

[1] Cui A, et al. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. EClinicalMedicine 2020;29-30:100587.
[2] Agaliotis M, et al. Burden of reduced work productivity among people with chronic knee pain: a systematic review. Occup Environ Med 2014;71(9):651–9.
[3] Hunter DJ, Schofield D, Callander E. The individual and socioeconomic impact of osteoarthritis. Nat Rev Rheumatol 2014;10(7):437–41.
[4] Leifer, V.P., J.N. Katz, and E. Losina, The burden of OA-health services and economics, in Osteoarthritis Cartilage. 2021.
[5] Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. Lancet 2019;393(10182):1745–59.
[6] Widera P, et al. Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data. Sci Rep 2020;10(1):8427.
[7] Hafezi-Nejad N, et al. Prediction of medial tibiofemoral compartment joint space loss progression using volumetric cartilage measurements: data from the FNIH OA biomarkers consortium. Eur Radiol 2017;27(2):464–73.
[8] Tiulpin A, et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci Rep 2019;9(1):20038.
[9] Jamshidi A, et al. Machine learning-based individualized survival prediction model for total knee replacement in osteoarthritis: data from the osteoarthritis initiative. Arthr Care Res (Hoboken) 2021;73(10):1518–27.
[10] Heisinger S, 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(5):1298.
[11] Leung K, et al. Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative. Radiology 2020;296(3):584–93.
[12] Mahmoud K, et al. Predicting total knee replacement at 2 and 5 years in osteoarthritis patients using machine learning. BMJ Surg Intervent Health Technol 2023;5(1):e000141.
[13] Tolpadi AA, et al. Deep learning predicts total knee replacement from magnetic resonance images. Sci Rep 2020;10(1):6371.
[14] Yoo HJ, et al. Prediction of progression rate and fate of osteoarthritis: comparison of machine learning algorithms. J Orthop Res 2023;41(3):583–90.
[15] Rajamohan HR, et al. Prediction of total knee replacement using deep learning analysis of knee MRI. Sci Rep 2023;13(1):6922.
[16] Mota RE, et al. Determinants of demand for total hip and knee arthroplasty: a systematic literature review. BMC Health Serv Res 2012;12:225.
[17] Hawker G, et al. Perspectives of Canadian stakeholders on criteria for appropriateness for total joint arthroplasty in patients with hip and knee osteoarthritis. Arthritis Rheumatol 2015;67(7):1806–15.
[18] Driban JB, et al. Defining and evaluating a novel outcome measure representing end-stage knee osteoarthritis: data from the osteoarthritis initiative. Clin Rheumatol 2016;35(10):2523–30.
[19] Driban JB, et al. The natural history of end-stage knee osteoarthritis: data from the osteoarthritis initiative. Semin Arthritis Rheum 2022;58:152148.
[20] Escobar A, et al. Development of explicit criteria for total knee replacement. Int J Technol Assess Health Care 2003;19(1):57–70.
[21] Riddle DL, Jiranek WA, Hayes CW. Use of a validated algorithm to judge the appropriateness of total knee arthroplasty in the United States: a multicenter longitudinal cohort study. Arthr Rheumatol 2014;66(8):2134–43.
[22] Driban JB, et al. The prognostic potential of end-stage knee osteoarthritis and its components to predict knee replacement: data from the osteoarthritis initiative. J Rheumatol 2023. jrheum.2023-0017.
[23] Mcconnell S, Kolopack P, Davis AM. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. Arthr Rheumatism 2001;45(5):453–61.
[24] Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16(4):494–502.
[25] Salis Z, et al. Evaluation of a measure of end-stage knee osteoarthritis compared to total knee replacement: an observational study using multicohort data. Semin Arthr Rheum. 2024;64:152336.
[26] Chen, T. and C. Guestrin. Xgboost: a scalable tree boosting system. in Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining. 2016.
[27] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression 2013.
[28] Altman R, et al. Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. Arthr Rheum 1986;29(8):1039–49.
[29] Dunn R, et al. Risk scoring for time to end-stage knee osteoarthritis: data from the Osteoarthritis Initiative. Osteoarthr Cartil 2020;28(8):1020–9.