Title
Deep Learning Predicts Total Knee Replacement from Magnetic Resonance Images.

Permalink
https://escholarship.org/uc/item/6h23t9p5

Authors
Tolpadi, Aniket A
Lee, Jinhee J
Pedoia, Valentina
et al.

SCIENTIFIC REPORTS
natureresearch

OPEN

Deep Learning Predicts Total Knee Replacement from Magnetic Resonance Images

Aniket A. Tolpadi [1,2], Jinhee J. Lee[2], Valentina Pedoia[2] & Sharmila Majumdar[2*]

Knee Osteoarthritis (OA) is a common musculoskeletal disorder in the United States. When diagnosed at early stages, lifestyle interventions such as exercise and weight loss can slow OA progression, but at later stages, only an invasive option is available: total knee replacement (TKR). Though a generally successful procedure, only 2/3 of patients who undergo the procedure report their knees feeling "normal" post-operation, and complications can arise that require revision. This necessitates a model to identify a population at higher risk of TKR, particularly at less advanced stages of OA, such that appropriate treatments can be implemented that slow OA progression and delay TKR. Here, we present a deep learning pipeline that leverages MRI images and clinical and demographic information to predict TKR with AUC 0.834 ± 0.036 ($p < 0.05$). Most notably, the pipeline predicts TKR with AUC 0.943 ± 0.057 ($p < 0.05$) for patients without OA. Furthermore, we develop occlusion maps for case-control pairs in test data and compare regions used by the model in both, thereby identifying TKR imaging biomarkers. As such, this work takes strides towards a pipeline with clinical utility, and the biomarkers identified further our understanding of OA progression and eventual TKR onset.

Knee Osteoarthritis (OA) is one of the most common musculoskeletal disorders in the United States, with estimates of its incidence rate ranging from 14 to 30 million[1,2]. Annual arthritis-related medical expenditures are nearly $140 million, and hip and knee OA together are the 11th highest contributor to global disability[3,4]. The propensity of knee OA to induce eventual disability can be attributed to structural changes in the joint that characterize the disease, as well as symptoms that can include inflammation, debilitating pain, and functional limitations[5,6]. Progression of the full-joint disease is typically assessed using the Kellgren-Lawrence (KL) scale, a 0-4 scale in which a higher score is associated with narrowing of the tibiofemoral joint (TFJ) space and other radiographic changes, and thus, a more advanced stage of knee OA[7]. When diagnosed at early stages (KL = 0, 1), knee OA can be managed through nonsurgical treatment options, including exercise and/or weight loss, oral medications such as acetaminophen or NSAIDs, or intra-articular injections such as corticosteroids and hyaluronic acid, all of which have varying degrees of success in reducing pain[8]. At late stages (KL = 4), however, no noninvasive option exists[9]; here, the only option is total knee replacement (TKR).

TKR is an elective procedure in which the knee joint is resurfaced with a metal or plastic implant intended to restore function, provide pain relief, and improve quality of life[10]. In the United States, estimates of TKR incidence lie at 400,000 each year, a figure expected to grow 143% by 2050 even through conservative projections[11]. While TKR is considered one of the most effective procedures in orthopedic surgery, electing for it is far from straightforward: noninvasive alternatives such as weight loss, physical therapy, and NSAIDs are first exhausted. If unsuccessful, a patient will undergo a thorough examination of clinical history and comprehensive imaging of the joint to determine if a TKR is feasible, and if so, the desired implant design and size[12,13]. The procedure is also imperfect: only 66% of patients report their knees feeling "normal," and 33% of patients report some degree of pain post-implant[14]. Furthermore, the implant can fail under some circumstances: periprosthetic joint infection and wound complications can be observed, and implant instability can occur due to aseptic loosening, malpositioning of the implant, and wear of joint components[15,16]. It is thus much preferable to prolong the good health of the knee, particularly in patients where OA has not advanced to the most severe stages, thereby delaying TKR as long as possible. This necessitates a model to identify patients at higher risk of TKR such that appropriate treatment options can be pursued.

[1]Department of Bioengineering, University of California, Berkeley, USA. [2]Department of Radiology and Biomedical Imaging, University of California, San Francisco, USA. *email: Sharmila.Majumdar@ucsf.edu
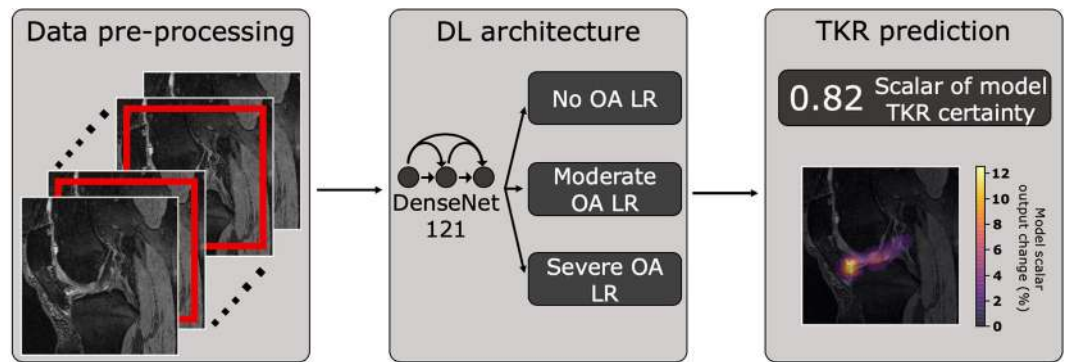
Figure 1. Pipeline predicting if patient will undergo TKR within 5 years from MRI/X-ray images and non-imaging variables. MRI and X-ray images are center-cropped and cropped to a region centered around the joint, respectively, and normalized. DenseNet-121 is pretrained to predict OA and fine-tuned to predict TKR. Image-based predictions and clinical information are fed to a logistic regression (LR) ensemble based on OA severity. Each ensemble, whose hyperparameters were optimized for Youden's index in a hyperparameter search, averages predictions of LR models in its OA severity for final TKR prediction. Pipeline is subsequently analyzed through occlusion map analysis to identify imaging biomarkers of TKR.

Given the multitude of factors on which a decision to pursue TKR is made, devising a model to predict if the invasive intervention will be necessary is a difficult task, but with obvious utility. For a patient in earlier stages of OA, a model predicting the patient to be at risk of TKR can be the impetus for a more aggressive nonsurgical treatment. Meanwhile, for a late-stage OA patient, a model predicting them to undergo TKR may facilitate a doctor and patient opting for the treatment earlier than they otherwise would, thereby reducing time spent pursuing nonsurgical alternatives with minimal probability of success while dealing with serious pain. Beyond this, if the model were to draw from medical images of the knee, it could identify anatomic regions most correlated with a TKR prediction. To this point, few studies have been conducted in this space, and those that have primarily investigate the importance of cartilage volume loss, subchondral bone defects, and bone marrow lesions[17–19]. An identification of more such biomarkers for TKR, however, could greatly improve understanding of both OA and TKR, and ultimately guide treatment strategies.

Predictive modeling of TKR, however, has a limited history, particularly with models that use medical images. A few studies have leveraged random forest regression, Cochran-Armitage tests for trend, and t-tests to identify demographic, general health, and physical examination measurements that most strongly correlate with TKR or total joint arthroplasty (TJA)[20,21]. Others have taken these efforts further, using techniques such as multiple regression and multivariate risk prediction models to predict TKR outright[22,23]. To our knowledge, only one group has developed a predictive model of TKR that accepts image inputs, attaining performance that surpasses that of models using only clinical and demographic information[24]. Notably, past TKR predictive models largely measure performance by evaluating the area under the receiver operating characteristic (ROC) curve, which plots true positive rate against false positive rate[25]. However, in most datasets used in this space, the number of patients who eventually undergo TKR is dramatically higher among those who have advanced OA as opposed to those with no or moderate OA. Consequently, this performance metric (AUC), while effectively capturing a model's combination of sensitivity and specificity, can be inflated for TKR prediction by indiscriminately predicting patients without OA not to undergo TKR, while more accurately predicting patients with severe OA to undergo TKR, the latter of which is easier. As a result, while past works have made clear progress in predicting TKR, none have overcome datasets imbalanced with respect to OA severity to report sensitive and specific prediction at these early stages, where a model would have the most utility.

One technique that has shown promise in delivering such performance is deep learning (DL). DL, especially convolutional neural networks (CNNs), has made strides in image classification tasks, attaining performances on the popular ImageNet classification challenge that approach or surpass human performance[26–28]. DL shines when afforded large datasets, as its automated feature extraction allows one to solve problems too complex for conventional approaches[29]. Given the complex prognostic features in TKR recommendation, CNNs become more promising for TKR prediction. In the past, DL had seen limited utility in OA and TKR prediction due to the large dataset requirement for efficacy; that limitation has been somewhat mitigated by the curation of large-sized cohort studies such as the Osteoarthritis Initiative (OAI)[30]. Consequently, DL has recently been applied for knee OA classification and progression prediction[9,31–33]. The success of these works further suggests the feasibility of leveraging DL to predict TKR.

In this study, we formulate a DL-based pipeline that incorporates knee joint images in addition to clinical and demographic information to predict the onset of TKR (Fig. 1). We demonstrate that the pipeline's predictions using solely Magnetic Resonance Imaging (MRI) images matches that of past work, while the integration of MRI image-based predictions with non-imaging variables facilitates TKR prediction with especially high sensitivity and specificity for patients without radiographic OA. Furthermore, we show the increase in pipeline performance when using 3D MRI images as opposed to 2D radiographs, suggesting MRI may have a role in TKR risk screening despite higher costs and more limited availability. And finally, we leverage occlusion maps to conduct a thorough

analysis of tissues that most significantly affect the output model metric associated with TKR prediction confidence, thereby identifying a set of imaging biomarkers for eventual TKR onset.

Novelty

This work reports a methodology and results that are novel in the following manners:

1. This model is the first to apply a 3-dimensional DenseNet CNN for prediction of TKR from MRI.
2. The TKR prediction model is evaluated for patients stratified by OA severity, which has not been reported in previous studies.
3. With the aim of improving model interpretability and clinical utility, we report the first comprehensive, case-control study to identify imaging biomarkers for TKR.

Materials and Methods

Data. Data was acquired from a prospective observational study conducted by OAI. The dataset followed 4,796 patients and acquired images including 2D posteroanterior radiographs and 3D Sagittal Double Echo Steady-State (DESS) MRI images over the course of 10 years. Details of data collection and study design have been previously reported[30]. The OAI study protocol was approved by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) and is registered on ClinicalTrials.gov as "Osteoarthritis Initiative (OAI): A Knee Health Study", NCT#00080171. The study was carried out in accordance with all pertinent guidelines and regulations, and written and informed consent was obtained from participants prior to each clinical visit in the study.

Both posteroanterior radiographs and DESS MRI images were evaluated as data sources for TKR prediction models. Patients for whom KL grade was not recorded at any point in the longitudinal study were excluded. To homogenize datasets, radiograph and MRI images were only taken from patients and time points at which both were available (n = 35,482). We labeled entries as cases if the patient underwent a first TKR within 5 years of the given time point (n = 1,043). We labeled entries as controls if patients did not undergo a TKR or eventually underwent one but the time to it was longer than 5 years (n = 34,439). Contralateral TKRs were not considered.

The radiographs and MRI images were preprocessed for training and model evaluation. Radiographs were cropped to a 500 × 500 region centered around the knee joint. Briefly, 2D cross-correlation template matching was used to identify a 500 × 500 bounding box centered around the knee joint in 450 joints, and these cases were used to train a U-Net architecture that identified this region for all posteroanterior radiographs from the OAI study[32]. DESS MRIs were center-cropped to a 120 × 320 × 320 region, after which both sets of cropped images were normalized. Normalized MRI pixel values were then rounded to nearest integers, compressing the MRI image to 14 possible pixel values. This rounding approach was initially tested as a strategy to accelerate training of a 3D CNN, given the large imaging volumes and large dataset on which it was being trained, believing the approach could suppress information extraneous to eventual TKR. Empirically, this approach yielded superior validation performance to leaving pixel values unrounded, so it was utilized. Examples of the results of this compression strategy are in Supplementary Fig. S1.

Non-imaging variables were screened for among studies and reviews detailing risk factors for knee OA progression and TKR onset[21–23,34–37]. Variables such as KL grade known to be deducible directly from MRI images and radiographs were not considered. From these studies, 40 non-imaging variables of interest were identified (see Supplementary Table S1). The OAI database was then parsed for corresponding variables, and these corresponding variables were added as potential non-imaging variables for our study, yielding 44 potential non-imaging variables. In some cases, multiple OAI metrics corresponded to non-imaging variables of interest, causing the number of OAI non-imaging variables to exceed what was identified from literature. Missing data points were imputed with k-nearest neighbors. These potential variables were used to train a random forest with 100 trees to predict onset of TKR within 5 years, and the minimum depth at which each feature was used across all trees in the forest was identified. Features whose minimum depth was below the average minimum depth of all features were preserved as non-imaging variables[38]. This yielded 27 non-imaging variables that are displayed in Table 1.

The data were then split into training, validation, and test with a 65%/20%/15% split, ensuring entries of any patient were only in one of the three datasets to prevent data leakage. Within the training set, imbalance between TKR and non-TKR cases was addressed with data augmentation, drawing bootstrap samples from the rare class with replacement[39]. A summary of the data prior to augmentation is provided in Table 2, detailing the number of cases and controls while showing descriptive statistics regarding demographics in each of the three datasets.

Pipeline architecture. The DL-based pipeline is based on a DenseNet-121 with the following parameters: 16 filters in initial layer, growth rate of 32, pooling block configuration of [6, 12, 24, 16], 4 bottleneck layers, 2 classes. The same architecture was used for the radiograph and MRI pipelines, but for the MRI pipeline, we modified the convolutional layers, batch normalization layers, pooling layers, and leaky rectified linear unit (ReLU) layers to allow for 3D image input[40]. The network yielded a scalar reflecting certainty of TKR within 5 years, which was added to the non-imaging variables. The 28 resulting variables were fed into one of three sets of Logistic Regression (LR) ensembles, with each ensemble optimized to maximize sensitivity and specificity in cases of no (KL = 0, 1), moderate (KL = 2, 3), and severe OA (KL = 4). Based on the KL grade of a sample, it was fed into an LR ensemble, yielding a prediction as to whether the patient will undergo a TKR within 5 years.

Training. A DenseNet-121 was initially pretrained to predict knee OA using the entire training set, assessing cross-entropy loss and accuracy on the validation set after completion of each epoch. The pre-train was stopped when validation loss began to increase. The pretrained model was subsequently fine-tuned to predict TKR. We utilized a random search to determine optimal learning rate, dropout rate, weights of the cross-entropy loss

Non-imaging variables used to augment image-based predictions	
Age	Comorbidity score
BMI	Injections to treat arthritis in previous 6 months
Education	Seen physician for arthritis in previous year
Ethnicity	Knee valgus negative alignment (degrees)
Income	Isometric leg strength
NSAID usage	Back pain in previous 30 days
Analgesics usage	Difficulty squatting in previous 7 days
Systolic BP	Difficulty kneeling in previous 7 days
Considering TKR	Baseline frequent knee pain status
PASE	Previous knee injury that limited walking
KOOS QOL	0-10 global rating assessing effect of knee pain
KOOS pain	SF-12 physical component score
WOMAC pain	SF-12 mental component score
WOMAC disability	

Table 1. List of non-imaging variables fed into logistic regression models to make predictions of whether a patient would undergo TKR within 5 years. Abbreviations used: Body Mass Index (BMI), Nonsteroidal Anti-inflammatory drugs (NSAIDS), Blood Pressure (BP), Physical Activity Scale for the Elderly (PASE), Knee Injury and Osteoarthritis Outcome Score (KOOS), Quality of Life (QOL), Western Ontario and McMaster Universities Arthritis Index (WOMAC), Short Form 12 (SF-12).

Dataset	Data type	Age	BMI	KOOS pain	Male	Female	OA status			Total entries	Unique patients
							None	Moderate	Severe		
Training	Control	62.5 ± 9.15	28.3 ± 4.75	87.2 ± 16.2	9,708	12,731	12,721	8,950	768	23,126	3,114
	Case	66.3 ± 8.38	29.6 ± 4.79	67.2 ± 19.6	291	396	41	357	289		
Validation	Control	62.4 ± 9.21	28.4 ± 4.64	87.6 ± 15.9	2,876	4,035	4,118	2,611	182	7,115	957
	Case	66.1 ± 8.76	29.8 ± 4.61	66.2 ± 19.1	70	134	13	93	98		
Test	Control	62.8 ± 9.55	28.4 ± 4.81	87.4 ± 16.5	2,126	2,963	2,892	2,056	141	5,241	719
	Case	66.4 ± 7.78	29.9 ± 3.96	68.7 ± 20.6	59	93	12	83	57		

Table 2. Data used to train 3D DESS MRI and 2D radiograph architectures. After exclusion criteria were applied, 35,482 qualifying entries were found in the OAI dataset across 4,790 unique patients, all of which were split into training, validation, and test sets as displayed in table. To prevent data leakage, all entries from any given patient were only allowed to be in one of the three sets. S.d. is reported for age, BMI, and KOOS pain score within the table.

function, and number of layers to freeze during fine-tuning. The search was carried out for 25 iterations, after which a set of parameters were selected that yielded the best combination of accuracy, sensitivity, and specificity on the validation set. Due to computational intensity, the hyperparameter search was not conducted on the entire dataset: for the 2D DenseNet-121, 10% of training and validation sets were used, whereas for the 3D DenseNet-121, 2.5% of both were used. After the search, the model fine-tuned using the subset of the training set was further fine-tuned on the entire training set using optimal parameters until validation loss began to increase. The test set was held out during training and predictions for it evaluated just once after fine-tuning, which marked the end of model optimization.

Integration of imaging and non-imaging data. Random forest regression, support vector machine, neural network, and LR architectures were assessed for efficacy of integrating imaging and non-imaging predictions, with LR providing best results on validation data. The LR architecture was thus used: all 28 imaging and non-imaging models were fed into an LR model, the optimal parameters of which were also identified through a random search. The search was conducted for 100 iterations, seeking to optimize the cross-entropy loss function weights afforded to both classes. For the cases of no, moderate, and severe OA, ideal parameters were identified by selecting those that maximized Youden's index within each OA classification in the search[41]. Predictions of the best few models in each classification were averaged to yield final TKR predictions. The number of predictions averaged in each classification was selected by finding a value that optimized validation accuracy, AUC, and Youden's index. The resulting LR models were ensembled and run on test data just once. Confidence intervals of accuracy, sensitivity, and specificity for each OA severity were obtained by bootstrapping, sampling 100% of test data with replacement (B = 100). Confidence intervals for AUC were calculated in the same manner. Results are reported on 3 versions of each model: the sole DenseNet-121 output (image only), output of a single LR model trained to predict TKR using solely the 27 non-imaging variables while not weighting the loss function class weights (non-imaging info. only), and output of the LR ensemble with image predictions (integrated model).

Statistical analysis.   The accuracies of X-ray and MRI pipeline performances within each OA classification and overall were compared using McNemar's test[42,43]. This test was appropriate because it specifically tests for differences in a dichotomous variable in matched groups. In our case, the variable was correct TKR prediction and the groups were the X-ray and MRI pipelines. Initially, the McNemar test statistic was modeled with a chi-squared distribution to test for significant differences between the pipelines, and if one existed, a binomial distribution was used to interrogate which pipeline yielded the significantly higher performance. All tests were carried out at $\alpha = 0.05$.

Relative sensitivity and specificity of the X-ray and MRI pipelines were assessed by comparing their AUCs within each OA classification and overall. This test is appropriate because the ROC curve plots true positive rate (sensitivity) against false positive rate (1 – specificity); consequently, the closer the AUC is to 1, the better the combination of sensitivity and specificity. 100% of test data was sampled with replacement (B = 100), and for each corresponding pair of X-ray and MRI pipelines (matched by OA classification and use of images only or both image and non-image information), AUCs were calculated. To test if one outperformed the other, differences in AUCs were calculated at each iteration, and the mean and standard deviation of the differences used to conduct a student's t-test with 99 degrees of freedom. This test is applicable on each matched pair of X-ray and MRI pipelines due to the number of iterations for which test data was sampled, allowing the central limit theorem to apply. For confidence intervals, mean and standard deviation of AUCs of individual models were calculated and used to report 95% intervals.

Imaging biomarker identification.   For all 124 true positives in the test data for the integrated MRI pipeline, corresponding controls were identified by randomly sampling from test data true negatives, keeping OA status distributions identical and using a student's t-test with 123 degrees of freedom to ensure no significant difference in KOOS pain scores across cases and corresponding controls at $\alpha = 0.05$. Occlusion maps were generated for all cases and controls using voxel size of $12 \times 32 \times 32$ and stride of 12. For each pixel, the value displayed represented the magnitude of change in the scalar pipeline output resulting when that pixel was occluded, averaged across all occlusions in which that pixel existed. Pixels for which scalar pipeline output change lied in the top 5% were designated as "hotspots." Anatomic regions of these hotspots were identified and odds ratios (OR) calculated to interrogate possible imaging biomarkers of TKR. 95% OR confidence intervals were calculated for each anatomic region investigated in this analysis using Cornfield's method, as this method performs well with relatively small sample sizes[44]. P values of ORs were calculated using a two-tailed Fisher's exact test[45]. Tissues where p values fell below the significance level of $\alpha = 0.05$ and in which 95% OR confidence intervals did not include 1 were deemed significant. These test selections were appropriate, as they allowed for direct comparison of the frequencies at which several tissues were hotspots across cases and controls, and as such, identified significant tissues with regards to TKR onset.

Results

OA pretrain utility in TKR prediction.
To test information learned from the OA pretrain, pretrained models themselves were used to predict TKR, with results depicted in Table 3. Predictably, the radiograph OA pretrain model had poor sensitivity for patients without OA, and poor specificity in moderate and severe cases of OA. While the MRI OA pretrain model expectedly yielded more balanced sensitivity and specificity across all OA stages, it too left room for improvement, particularly in sensitivity at no OA and specificity at severe OA. This confirmed the pretrain provided useful information to both architectures but fine-tuning and integration of non-imaging variables were necessary to attain desired TKR prediction performance.

X-Ray pipeline optimization and performance.
For the X-Ray model, hyperparameter tuning steps found the following to yield the best combination of validation accuracy, sensitivity, and specificity: learning rate of 3.981×10^{-6}, TKR class weight in cross-entropy loss function of 0.927 and non-TKR class weight of 0.073, dropout rate of 0.375, and only the last 2 layers fine-tuned after OA pretrain.

A radiograph model was fine-tuned to predict TKR with these parameters, and its predictions fed into an LR ensemble. Averaging predictions of the best 5 LR models found through random search in the 3 OA categories yielded best validation performance, so this ensemble was used on the test set. Test accuracy, sensitivity, and specificity are provided in Table 4, and ROC curves of all three versions of this pipeline are found in Fig. 2a. AUCs are as follows: 0.848 ± 0.039 (image only), 0.868 ± 0.028 (non-imaging info. only), 0.890 ± 0.021 (integrated model). Furthermore, AUCs for the image-only and combined versions of the pipeline at no OA are as follows: 0.514 ± 0.087 (image only); 0.799 ± 0.055 (integrated model). At moderate OA: 0.788 ± 0.025 (image only); 0.865 ± 0.016 (integrated model). At severe OA: 0.552 ± 0.040 (image only); 0.641 ± 0.044 (integrated model). All AUC intervals are calculated using standard deviation (s.d.), $p < 0.05$.

MRI pipeline optimization and performance.
Similarly, a hyperparameter search was carried out for the MRI pipeline to optimize parameters for eventual fine-tuning. The following hyperparameters were found optimal: learning rate of 1.906×10^{-2}, TKR class cross-entropy weight of 0.902 and non-TKR class weight of 0.098, dropout rate of 0.329, only last layer of model fine-tuned after OA pretrain.

An MRI-based model was fine-tuned from these parameters. The resulting predictions were fed into an LR ensemble, where averaging predictions of the best 4 models in each OA category optimized validation performance. Performance of the resulting architecture on test data is reported in the same manner as the radiograph pipeline, in Table 4 and Fig. 2b. AUCs are as follows: 0.886 ± 0.020 (image only), 0.868 ± 0.028 (non-imaging info. only), 0.834 ± 0.036 (integrated model). AUCs for the image-only and combined pipeline versions at no OA are as follows: 0.897 ± 0.039 (image only); 0.943 ± 0.029 (integrated model). At moderate OA: 0.764 ± 0.020

OA status	Model type	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Non-TKR cases	TKR cases
None	Radiograph	92.1 ± 0.083	25.2 ± 2.16	92.4 ± 0.081	2,892	12
	MRI	94.3 ± 0.070	48.7 ± 2.48	94.4 ± 0.070		
Moderate	Radiograph	29.3 ± 0.151	93.8 ± 0.439	26.7 ± 0.156	2,056	83
	MRI	65.4 ± 0.154	65.5 ± 0.848	65.4 ± 0.158		
Severe	Radiograph	29.7 ± 0.488	100.0 ± 0.000	1.4 ± 0.180	141	57
	MRI	33.4 ± 0.523	82.2 ± 0.824	14.0 ± 0.441		
All	Radiograph	64.2 ± 0.124	90.7 ± 0.378	63.4 ± 0.126	5,089	152
	MRI	80.2 ± 0.079	70.4 ± 0.595	80.5 ± 0.082		

Table 3. Performance in TKR prediction of OA pretrained models for radiographs and MRI, stratified by severity of OA. Pretraining strategy yields useful information to both models, but performance at no OA in particular leaves room for improvement, justifying subsequent model fine-tuning. Standard errors used to calculate confidence intervals.

OA status	Image source	Model type	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Non-TKR cases	TKR cases
None	X-ray	Non-imaging info. only	89.1 ± 0.139	49.7 ± 2.80	89.3 ± 0.140	2,892	12
		Image only	95.0 ± 0.089	7.8 ± 1.64	95.4 ± 0.089		
		Integrated model	95.4 ± 0.081	8.6 ± 1.95	95.8 ± 0.077		
	MRI	Non-imaging info. only	89.1 ± 0.139	49.7 ± 2.80	89.3 ± 0.140		
		Image only	95.2 ± 0.088	66.9 ± 3.23	95.3 ± 0.089		
		Integrated model	82.4 ± 0.171	92.2 ± 1.68	82.4 ± 0.173		
Moderate	X-ray	Non-imaging info. only	72.9 ± 0.208	70.0 ± 1.16	73.0 ± 0.212	2,056	83
		Image only	79.9 ± 0.196	66.7 ± 1.23	80.4 ± 0.195		
		Integrated model	81.4 ± 0.178	76.0 ± 1.12	81.6 ± 0.179		
	MRI	Non-imaging info. only	72.9 ± 0.208	70.0 ± 1.16	73.0 ± 0.212		
		Image only	68.8 ± 0.225	78.3 ± 0.952	68.4 ± 0.227		
		Integrated model	74.9 ± 0.216	78.9 ± 0.974	74.7 ± 0.228		
Severe	X-ray	Non-imaging info. only	51.3 ± 0.744	89.4 ± 0.864	35.8 ± 0.925	141	57
		Image only	32.1 ± 0.714	94.5 ± 0.735	7.2 ± 0.467		
		Integrated model	60.5 ± 0.775	64.0 ± 1.57	59.0 ± 0.959		
	MRI	Non-imaging info. only	51.3 ± 0.744	89.4 ± 0.864	35.8 ± 0.925		
		Image only	34.6 ± 0.775	98.3 ± 0.390	9.2 ± 0.632		
		Integrated model	59.6 ± 0.770	84.0 ± 1.03	49.6 ± 1.04		
All	X-ray	Non-imaging info. only	81.1 ± 0.118	75.6 ± 0.776	81.2 ± 0.122	5,089	152
		Image only	86.4 ± 0.095	72.5 ± 0.864	86.9 ± 0.095		
		Integrated model	88.4 ± 0.094	66.3 ± 0.924	89.1 ± 0.090		
	MRI	Non-imaging info. only	81.1 ± 0.118	75.6 ± 0.776	81.2 ± 0.122		
		Image only	82.1 ± 0.118	84.9 ± 0.636	82.1 ± 0.119		
		Integrated model	78.5 ± 0.134	81.8 ± 0.643	78.4 ± 0.138		

Table 4. Performance of X-ray and MRI architectures on test data. While integrated X-ray pipeline delivers higher accuracy than integrated MRI pipeline, integrated MRI pipeline yields improved sensitivity over integrated X-ray pipeline across all stages of OA, markedly so at no OA. Standard errors used to calculate confidence intervals.

(image only); 0.830 ± 0.024 (integrated model). At severe OA: 0.560 ± 0.042 (image only); 0.726 ± 0.038 (integrated model). Again, all AUC intervals are calculated using s.d., $p < 0.05$.

Comparison of MRI and radiograph pipeline performances. A comparison of overall AUCs attained by the integrated MRI and X-ray pipelines across OA grades and overall shows that at no OA and severe OA, the MRI pipeline outperformed the X-ray pipeline (No OA, B = 100: $p = 3.04 \times 10^{-2}$; Moderate OA, B = 100: $p = 9.55 \times 10^{-1}$; Severe OA, B = 100: $p = 4.57 \times 10^{-2}$; Overall, B = 100: $p = 9.94 \times 10^{-1}$). The MRI pipeline thus has a superior combination of sensitivity and specificity than does the X-ray pipeline for patients without OA and those with severe OA. The AUCs obtained by the image-only pipelines also were compared, and showed the MRI pipeline to outperform the X-ray pipeline for patients without OA and overall (No OA, B = 100: $p = 6.10 \times 10^{-5}$; Moderate OA, B = 100: $p = 7.58 \times 10^{-1}$; Severe OA, B = 100: $p = 4.37 \times 10^{-1}$; Overall, B = 100: $p = 1.16 \times 10^{-2}$). These results follow intuition: while radiographic imaging is primarily capable of illuminating bones in the joint, MRI can visualize soft tissues such as cartilage, muscle, and meniscus[46,47]. It follows that an MRI model will
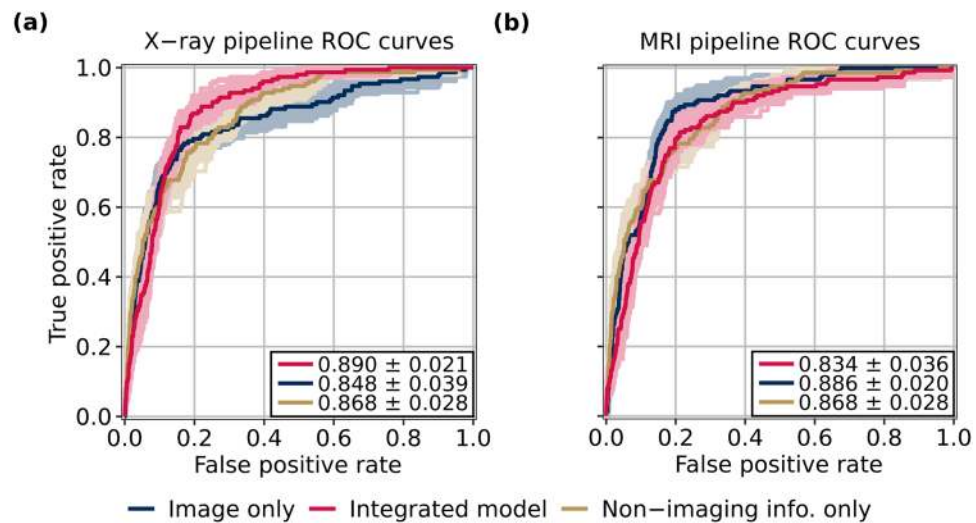
Figure 2. ROC curves for X-ray and MRI architectures on test data. X-ray pipeline ROC curves are shown in **(a)**, with AUCs as follows, $p < 0.05$: 0.848 ± 0.039 (image only), 0.868 ± 0.028 (non-imaging info. only), 0.890 ± 0.021 (integrated model). MRI pipeline ROC curves are shown in **(b)**, with AUCs as follows, $p < 0.05$: 0.886 ± 0.020 (image only), 0.868 ± 0.028 (non-imaging info. only), 0.834 ± 0.036 (integrated model). Standard deviations used to calculate confidence intervals. ROC curves with AUCs within 1 standard deviation of the mean for each model type during bootstrapping are also shown on plots.

exhibit a better combination of sensitivity and specificity, especially in early OA stages at which few radiographic changes in the knee have occurred. ROC curves for pipeline versions and OA classifications in which the MRI architecture yielded a significantly better AUC than its X-ray counterpart are shown in Fig. 3.

McNemar's test assessed relative accuracies of these pipelines. There was a statistically significant difference between the accuracies of the integrated X-ray and MRI pipelines for patients at no OA, moderate OA, and overall (No OA, n = 537: $p = 1.65 \times 10^{-59}$; Moderate OA, n = 521: $p = 1.13 \times 10^{-9}$; Severe OA, n = 47: $p = 8.84 \times 10^{-1}$; Overall, n = 1,105: $p = 1.52 \times 10^{-54}$), and in each of those 3 statistically significant cases, the X-ray pipeline outperformed the MRI pipeline (No OA, n = 537: $p = 1.11 \times 10^{-16}$; Moderate OA, n = 521: $p = 5.97 \times 10^{-10}$; Overall, n = 1,105: $p = 1.11 \times 10^{-16}$). In interpreting these tests and the AUC tests holistically, it is evident that the X-ray pipeline is able to attain superior accuracy in several OA classifications by compromising on its combination of sensitivity and specificity. This is further supported by the accuracies and sensitivities reported for the respective pipelines in Table 4, which show that while the X-ray pipeline is more accurate than its MRI counterpart at every OA classification, the opposite is true for sensitivity—drastically so for patients without OA. In the clinic, where sensitivity as to whether a patient is at risk of eventual TKR is paramount, these results would show the MRI pipeline to be the more useful model.

It is also worthy to note the improvement in performance that occurs for patients without OA when imaging predictions are added to non-imaging variables in both pipelines. In the X-ray pipeline, the model's AUC increased from 0.514 ± 0.087 to 0.799 ± 0.055 when non-imaging variables were added to the radiographs, a sizeable increase when compared to the MRI pipeline performance, which saw AUC increase from 0.897 ± 0.039 to 0.943 ± 0.029 ($p < 0.05$ for all). This demonstrates that non-imaging variables such as various pain scales seem to add critical information to the X-ray pipeline, while the same information is less important in the MRI pipeline.

Biomarker identification and analysis. Of the 152 patients in test data who underwent a TKR, 124 were detected by the MRI pipeline. Occlusion maps were generated for these cases and their corresponding true negative controls, an example of which is shown in Fig. 4. Tissues and their hotspot percentages across these true positives and corresponding true negative controls can be found in Supplementary Table S2 and S3, respectively. ORs, 95% confidence intervals, and associated p values for each tissue can be found in Table 5.

Three tissues saw ORs and 95% confidence intervals that lied above 1 and p values below $\alpha = 0.05$: the medial patellar retinaculum, gastrocnemius tendon, and plantaris muscle. Thus, we conclude there is a substantial and statistically significant difference in the risk of TKR within 5 years when these tissues are identified as hotspots by the pipeline. From the ORs, we see that the risk of TKR increases when any of the three are identified as hotspots: for the medial patellar retinaculum, the risk is 1.98 times higher with a 95% confidence interval from 1.02 to 3.99; for the gastrocnemius tendon, it is 2.97 times higher with a 95% confidence interval from 1.12 to 10.0; and for the plantaris muscle, it is 2.84 times higher with a 95% confidence interval from 1.47 to 5.82. As such, these results provide evidence that all are imaging biomarkers of TKR.

On the other hand, several tissues located within or near the tibiofemoral joint—namely, cartilage and bone in both medial and lateral locations of the joint, menisci in all tested regions, and the ACL—saw ORs and 95% confidence intervals entirely below 1 and p values below $\alpha = 0.05$. Consequently, for all of these tissues, we find a statistically significant difference in the risk of TKR within 5 years when these tissues are identified as hotspots. In the case of each, the risk of TKR appears to decrease when these tissues are identified as hotspots. Interestingly,
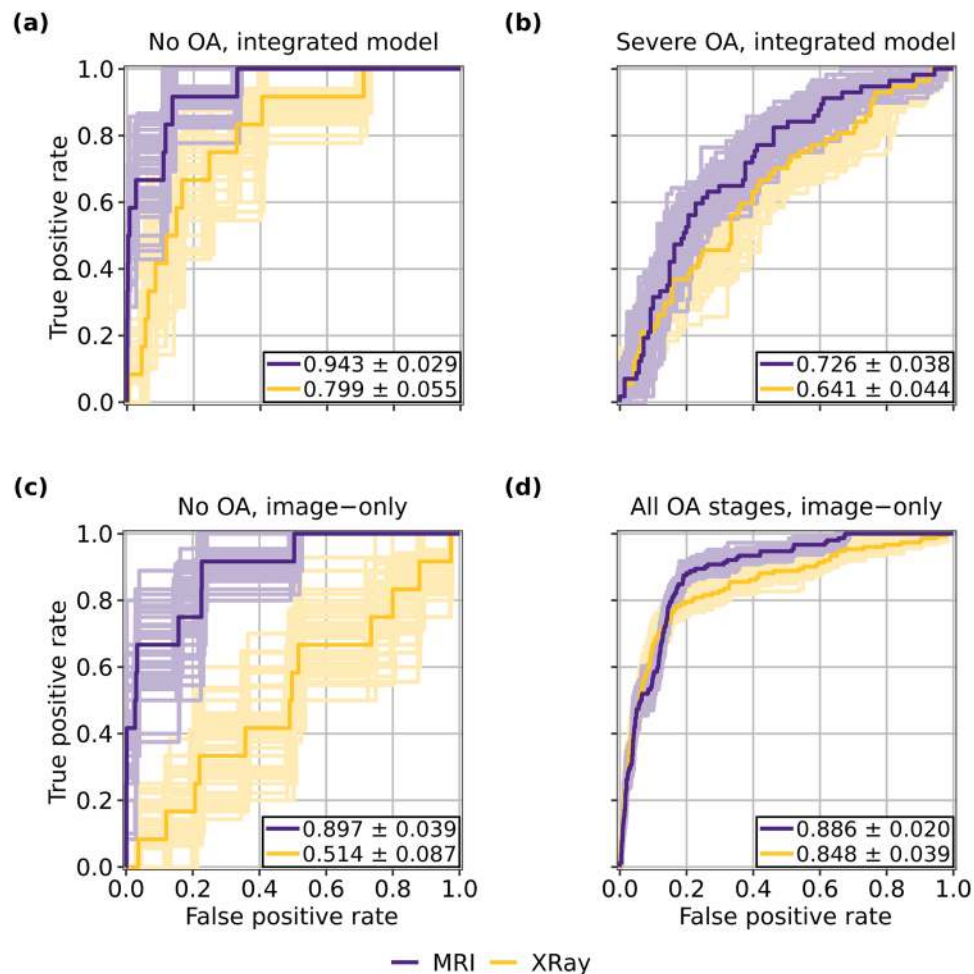
Figure 3. ROC curves for MRI and X-ray pipelines at selected OA classifications and pipeline versions in which MRI performance was significantly better than that of X-ray. MRI pipeline outperforms X-ray pipeline at no OA for both image-only and integrated models, as seen in **(a,c)**. As shown in **(b)**, integrated MRI pipeline also outperformed integrated X-ray pipeline for patients with severe OA, while **(d)** shows image-only MRI pipeline outperformed image-only X-ray pipeline across all OA stages. AUCs are displayed in the figure with p < 0.05. Standard deviations used to calculate confidence intervals. ROC curves with AUCs within 1 standard deviation of the mean for each pipeline version during bootstrapping are also shown on plots.

each of these tissues have either been implicated as imaging biomarkers of OA progression, or damage within them is associated with OA onset[48–50]. These results, in conjunction with the three tissues in which risk of TKR increased when identified as hotspots, suggest that compared to OA progression, TKR onset relies less on tissues in and around the tibiofemoral joint and more on tissues in other locations of the joint to make predictions. TKR has been considered an outcome of OA progression, but these results demonstrate in part how it is a more nuanced problem.

Discussion

In this work, we present a pipeline that integrates MR imaging and non-imaging features to attain strong TKR prediction performance, reporting accuracy of $78.5 \pm 0.134\%$, sensitivity of $81.8 \pm 0.643\%$, and specificity of $78.4 \pm 0.138\%$ (intervals calculated with standard error of measurement (s.e.m.), p < 0.05). Comparisons of AUCs showed the MRI pipeline to outperform the X-ray pipeline for patients without OA and with severe OA, thereby showing the MRI model to have a better combination of sensitivity and specificity in these OA classifications. That it did so particularly for patients without OA shows the utility of the MRI pipeline in screening for patients at risk of TKR despite higher costs. It was also interesting that, particularly among patients with no OA, the X-ray model improved drastically more than the MRI model when non-imaging information was added, judging by disparities in AUCs. This suggests the MRI-trained DenseNet-121 may have learned to predict some of the non-imaging features from the images themselves, indicating that MRI images may intrinsically contain information regarding pain, quality of life, and physical performance, among other non-imaging variables used in this study. The utility of MRI in predicting these variables through DL is certainly worth further investigation.
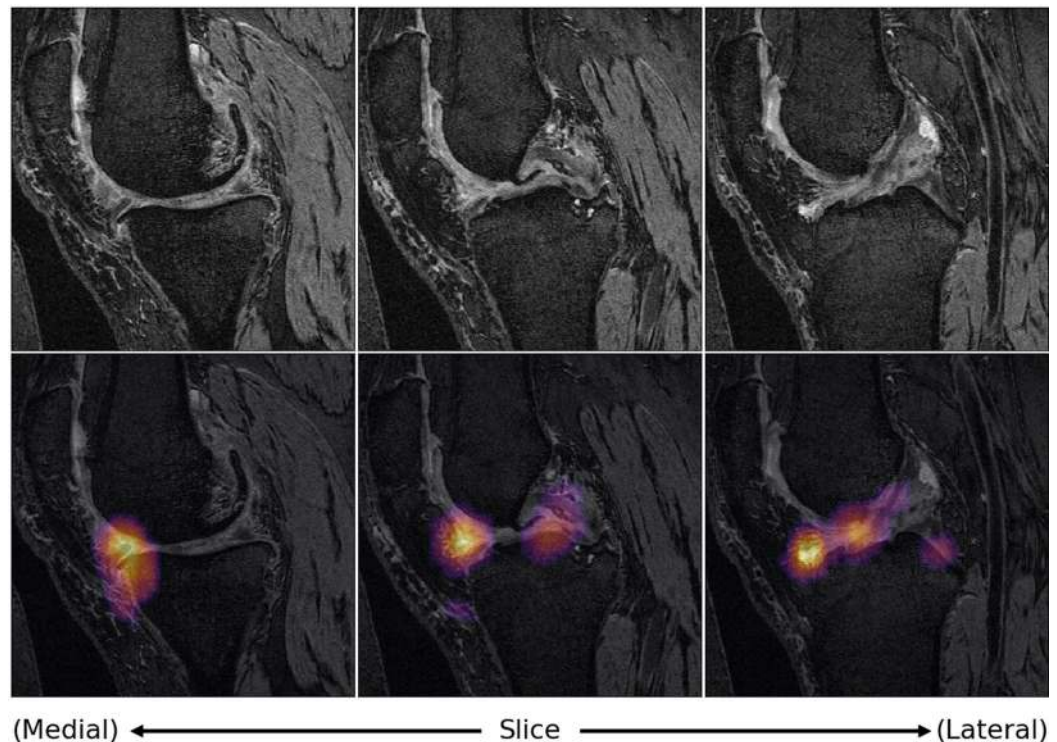
Figure 4. Slices of occlusion map of true positive detected by MRI pipeline, overlaid on corresponding slices of DESS MRI. Such maps were generated and analyzed for all 124 true positives and corresponding true negative controls of the integrated MRI pipeline.

A comparison of the MRI pipeline performance to past work is insightful. The closest analog to our work was conducted by Wang, T. *et al.*[24], who trained independent residual networks to predict TKR from both DESS and Turbo Spin Echo (TSE) MRI images, integrating both predictions with non-imaging variables in an LR model to yield a final TKR prediction. This yielded a model with AUC of 0.86 ± 0.01 ($p < 0.01$) when solely DESS or TSE images were used, and 0.88 ± 0.02 ($p < 0.01$) when both images and non-imaging features were integrated. Our MRI image-only model saw AUC of 0.886 ± 0.020 (image only, $p < 0.05$) and an integrated AUC of 0.834 ± 0.036 (combined, $p < 0.05$). Our image-only model thus yields performance superior to its image-only counterpart, with a 95% confidence interval lying entirely above the mean AUC of the image-only model by Wang, T. *et al.*[24]. Our integrated model, as discussed previously, was optimized to maximize Youden's index within each OA classification rather than overall AUC, explaining why our integrated model has a lower overall AUC than our image-only model. However, due to this decision, we obtained strong performance at early and moderate OA stages, with sensitivity and specificity of $92.2 \pm 1.68\%$ and $82.4 \pm 0.173\%$ at no OA, respectively, and $78.9 \pm 0.974\%$ and $74.7 \pm 0.228\%$ at moderate OA (intervals calculated using s.e.m., $p < 0.05$). In particular, the AUC of 0.943 ± 0.029 (interval calculated with s.d., $p < 0.05$) obtained by the MRI pipeline for patients without OA, the most difficult OA classification from which to predict TKR, by far surpasses that of past TKR predictive models that include patients across all stages of OA. This performance marks progress towards a model that identifies patients at risk for TKR such that nonsurgical treatment strategies can be implemented to delay TKR.

The biomarker analysis conducted also has implications, as it identified several tissues located within or near the tibiofemoral joint as reducing risk of TKR when identified as hotspots by the full MRI pipeline—namely, these were medially and laterally located cartilage and bone, all examined meniscal regions, and the ACL. These tissues or damage within them all have been associated with progression or onset of OA, and that our model shows TKR onset to be less reliant on these imaging features in cases compared to controls demonstrates TKR onset to be a more complicated problem than OA progression, despite the relationship between the two. On the other hand, the model identifies three tissues as increasing risk of TKR when identified as hotspots in the pipeline: the medial patellar retinaculum, gastrocnemius tendon, and plantaris muscle. The medial patellar retinaculum is crucial for lateral stabilization of the knee joint, and as such, damage to it results in a patella that more easily dislocates[51]. Past work has shown patellar dislocation increases risk for OA, and TKR can be an effective procedure to treat inveterate patellar dislocation, showing a previous link between this tissue's functionality and eventual OA and TKR[52,53]. The gastrocnemius tendon and plantaris muscle, on the other hand, are both posteriorly located tissues within the knee that play a key role in knee flexion[54]. While literature regarding the plantaris muscle is rather sparse, injuries to the muscle can be implicated in knee and calf pain felt by a patient[55]. Given their related functionality and location, the gastrocnemius tendon and plantaris muscle can jointly be implicated in conditions such as "tennis leg," which refers to mid-calf pain felt during extension of the leg, usually due to damage to one of these tissues or their associated muscles or tendons[56]. The significance of the plantaris muscle and gastrocnemius

Tissue type	Tissue	OR (95% CI, n = 124)	P value (n = 124)
Cartilage	TFJ medial*	0.05 (0.00–0.48)	3.36×10^{-3}
	TFJ lateral*	0.03 (0.00–0.25)	3.89×10^{-5}
	PFJ	1.03 (0.60–1.77)	1.00×10^{0}
Meniscus	Medial anterior*	0.33 (0.12–0.79)	1.04×10^{-2}
	Medial posterior*	0.40 (0.16–0.89)	2.37×10^{-2}
	Lateral anterior*	0.23 (0.05–0.67)	5.05×10^{-3}
	Lateral posterior*	0.26 (0.06–0.80)	1.49×10^{-2}
Bone	TFJ medial*	0.17 (0.00–0.91)	3.57×10^{-2}
	TFJ lateral*	0.02 (0.00–0.22)	8.48×10^{-6}
	PFJ	1.11 (0.64–1.92)	7.93×10^{-1}
Ligament	ACL*	0.49 (0.23–0.99)	4.72×10^{-2}
	PCL	1.58 (0.89–2.87)	1.27×10^{-1}
	Popliteal	1.62 (0.96–2.77)	7.51×10^{-2}
Tendon	Medial patellar retinaculum*	1.98 (1.02–3.99)	4.19×10^{-2}
	Lateral patellar retinaculum	1.08 (0.60–1.96)	8.88×10^{-1}
	Popliteal	1.49 (0.87–2.57)	1.56×10^{-1}
	Patellar	1.76 (0.92–3.48)	9.00×10^{-2}
	Gastrocnemius*	2.97 (1.12–10.0)	2.67×10^{-2}
	Semimembranosus	0.50 (0.23–1.03)	6.17×10^{-2}
	Quadriceps	3.18 (0.88–20.4)	8.38×10^{-2}
	Gracilis	4.52 (0.74–290)	1.20×10^{-1}
Fat pad	Hoffa	2.38 (0.92–7.38)	7.80×10^{-2}
Muscle	Popliteus	1.98 (1.00–4.14)	5.11×10^{-2}
	Vastus medialis	1.26 (0.54–3.00)	6.93×10^{-1}
	Gastrocnemius	1.35 (0.73–2.54)	3.76×10^{-1}
	Plantaris*	2.84 (1.47–5.82)	1.29×10^{-3}
	Biceps femoris	4.52 (0.74–290)	1.20×10^{-1}
	Tibialis anterior	2.37 (0.24–161)	6.22×10^{-1}
	Semimembranosus	0.35 (0.05–1.32)	1.36×10^{-1}
Synovium	General	1.17 (0.50–2.82)	8.41×10^{-1}

Table 5. Summary of occlusion map analysis comparing frequencies with which selected knee joint tissues were indicated as hotspots in analysis. Hotspots were defined as pixels that, when occluded, were among the top 5% of all pixels in change of pipeline TKR prediction output metric when occluded. Odds ratios, 95% confidence intervals calculated using Cornfield's method, and p values calculated using Fisher's exact test are displayed. Tissues that were significant at $\alpha = 0.05$ are designated with a *. N value for all tests was n = 124.

tendon to OA progression and TKR, however, have not been well characterized, and these results justify future studies to these ends.

This study had some limitations. The first is specific to the OAI dataset, which tends towards older, female patients, all from the United States: across 4,796 patients, the mean age is 61 years and 58% of patients are female. This is not emblematic of the general population, so the robustness of the pipeline could be strengthened by testing on a dataset such as the Multicenter Osteoarthritis Study (MOST). A further limitation of the dataset is that, despite the fairly large size, there are a very limited number of patients with the classification of most interest: those without radiographic OA that still undergo TKR within 5 years. Only 66 such cases existed in the entire OAI dataset, and 12 were in the test set. As such, the OAI dataset and the number of comparison experiments we ran within and across OA classifications limits the statistical power of our conclusions. Furthermore, in this study, pixels in MRI images were compressed to 14 possible values to optimize performance—a version of the pipeline was also constructed and evaluated without the compression, but its TKR prediction performance was not as strong. Ideally, a model that uses all available information would be used in occlusion map analysis to draw more precise conclusions regarding anatomic regions that associate with TKR, but this compromise was necessary to improve performance. A final limitation was computational intensity in occlusion map generation: the voxel size and stride used were $12 \times 32 \times 32$ and 12, respectively. These ideally would be smaller so maps could yield more precise insights but doing so was infeasible in a reasonable amount of time.

To conclude, this work presents a predictive model that delivers performance not previously seen in predicting TKR, especially for patients without OA. By delivering such performance, this pipeline can identify patients at risk of TKR with high sensitivity and specificity, and for patients with no or moderate OA, this can allow a non-invasive treatment to be implemented that prolongs good health of the knee and delays TKR. The biomarker analysis identifies the medial patellar retinaculum, gastrocnemius tendon, and plantaris muscle as increasing risk of TKR when identified as a hotspot by the model, while its assessment that several tissues within and near the tibiofemoral joint appear to reduce risk of TKR helps demonstrate the added complexity of predicting TKR onset

as opposed to OA progression. Beyond this, additional directions include investigating a more effective means of integrating non-image information with image predictions to improve TKR prediction performance, assessing the efficacy of alternate network architectures, and reducing computational time to make predictions.

Data availability

The datasets analyzed during the current study are available through the Osteoarthritis Initiative, which can be accessed at https://nda.nih.gov/oai/. In addition, model checkpoints, code, and label files used to produce presented results can be accessed at https://bitbucket.org/atolpadi7/dl-tkr-sci-rep/.

References

1. Kremers, H. M. *et al.* Prevalence of total hip and knee replacement in the United States. *J. Bone Joint Surg. Am.* **97**, 1386–1397 (2015).
2. Deshpande, B. R. *et al.* The number of persons with symptomatic knee osteoarthritis in the United States: impact of race/ethnicity, age, sex, and obesity. *Arthritis Care Res. (Hoboken)* **68**, 1743–1750 (2016).
3. Cross, M. *et al.* The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Ann. Rheum. Dis.* **73**, 1323–1330 (2014).
4. Murphy, L. B., Cisternas, M. G., Pasta, D. J., Helmick, C. G. & Yelin, E. H. Medical expenditures and earnings losses among US adults with arthritis in 2013. *Arthritis Care Res. (Hoboken)* **70**, 869–876 (2018).
5. Lawrence, R. C. *et al.* Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part II. *Arthritis Rheum.* **58**, 26–35 (2008).
6. Ostrander, R. V., Leddon, C. E., Hackel, J. G., O'Grady, C. P. & Roth, C. A. Efficacy of unloader bracing in reducing symptoms of knee osteoarthritis. *Am. J. Orthop. (Belle Mead N. J.)* **45**, 306–311 (2016).
7. Kellgren, J. H. & Lawrence, J. S. Radiological assessment of osteo-arthrosis. *Ann. Rheum. Dis.* **16**, 494–502 (1957).
8. Ringdahl, E. & Pandit, S. Treatment of knee osteoarthritis. *Am. Fam. Physician* **83**, 1287–1292 (2011).
9. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P. & Saarakkala, S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci. Rep.* **8**, 1727, https://doi.org/10.1038/s41598-018-20132-7 (2018).
10. Nguyen, L. C., Lehil, M. S. & Bozic, K. J. Trends in total knee arthroplasty implant utilization. *J. Arthroplasty* **30**, 739–742 (2015).
11. Inacio, M. C. S., Paxton, E. W., Graves, S. E., Namba, R. S. & Nemes, S. Projected increase in total knee arthroplasty in the United States – an alternate projection model. *Osteoarthritis Cartilage* **25**, 1797–1803 (2017).
12. Tanzer, M. & Makdhom, A. M. Preoperative planning in primary total knee arthroplasty. *J. Am. Acad. Orthop. Surg.* **24**, 220–230 (2016).
13. Sassoon, A., Nam, D., Nunley, R. & Barrack, R. Systematic review of patient-specific instrumentation in total knee arthroplasty: new but not improved. *Clin. Orthop. Relat. Res* **473**, 151–158 (2015).
14. Parvizi, J. *et al.* High level of residual symptoms in young patients after total knee arthroplasty. *Clin. Orthop. Relat. Res* **472**, 133–137 (2014).
15. Chang, M. J., Lim, H., Lee, N. R. & Moon, Y. Diagnosis, causes and treatments of instability following total knee arthroplasty. *Knee Surg. Relat. Res* **26**, 61–67 (2014).
16. Kim, K. T. *et al.* Causes of failure after total knee arthroplasty in osteoarthritis patients 55 years of age or younger. *Knee Surg. Relat. Res* **26**, 13–19 (2014).
17. Raynauld, J. P. *et al.* Correlation between bone lesion changes and cartilage volume loss in patients with osteoarthritis of the knee as assessed by quantitative magnetic resonance imaging over a 24-month period. *Ann. Rheum. Dis.* **67**, 683–688 (2008).
18. Raynauld, J. P. *et al.* Total knee replacement as a knee osteoarthritis outcome: predictors derived from a 4-year long-term observation following a randomized clinical trial using chondroitin sulfate. *Cartilage* **4**, 219–226 (2013).
19. Everhart, J. S., Abouljoud, M. M., Kirven, J. C. & Flanigan, D. C. Full-thickness cartilage defects are important independent predictive factors for progression to total knee arthroplasty in older adults with minimal to moderate osteoarthritis: data from the osteoarthritis initiative. *J. Bone Joint Surg. Am* **101**, 56–63 (2019).
20. Riddle, D. L., Kong, X. & Jiranek, W. A. 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* **16**, 494–500 (2009).
21. Hawker, G. A. *et al.* A prospective population-based study of the predictors of undergoing total joint arthroplasty. *Arthritis Rheum.* **54**, 3212–3220 (2006).
22. Lewis, J. R., Dhaliwal, S. S., Zhu, K. & Prince, R. L. A predictive model for knee joint replacement in older women. *PLoS One* **8**, e83665, https://doi.org/10.1371/journal.pone.0083665 (2013).
23. Yu, D. *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.* **78**, 91–99 (2019).
24. Wang, T., Leung, K., Cho, K., Chang, G. & Deniz, C.M. Total knee replacement prediction using structural MRIs and 3D convolutional neural networks. In International Conference on Medical Imaging with Deep Learning – Extended Abstract Track, **79** (2019).
25. Akobeng, A. K. Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta. Paediatr.* **96**, 644–647 (2007).
26. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778 (2016).
27. Huang, G., Lui, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2261-2269 (2017).
28. Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis* **115**, 211–252 (2015).
29. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat. Rev* **521**, 436–444 (2015).
30. Peterfy, C. G., Schneider, E. & Nevitt, M. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis Cartilage* **16**, 1433–1441 (2008).
31. Antony, J., McGuinness, K., O'Connor, N. E. & Moran, K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. *In 23rd International Conference on Pattern Recognition (ICPR)*, 1195-1200 (2016).
32. Norman, B., Pedoia, V., Noworolski, A., Link, T. M. & Majumdar, S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J. Digit. Imaging* **32**, 471–477 (2019).
33. Tiulpin, A. *et al.* Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Sci Rep.* **9**, 20038, https://doi.org/10.1038/s41598-019-56527-3 (2019).
34. Heidari, B. Knee osteoarthritis prevalence, risk factors, pathogenesis, and features. *Caspian J. Intern. Med.* **2**, 205–12 (2011).
35. Cooper, C. *et al.* Risk factors for the incidence and progression of radiographic knee osteoarthritis. *Arthritis Rheum.* **43**, 995–1000 (2000).
36. Pisters, M. F. *et al.* The course of limitations in activities over 5 years in patients with knee and hip osteoarthritis with moderate functional limitations: risk factors for future functional decline. *Osteoarthr. Cartilage* **20**, 503–10 (2012).

37. Sharma, L. *et al*. Physical functioning over three years in knee osteoarthritis. *Arthritis Rheum.* **48**, 3359–70 (2003).
38. Dietrich, S. *et al*. Random survival forest in practice: a method for modelling complex metabolomics data in time to event analysis. *Int. J. Epidemiol.* **45**, 1406–20 (2016).
39. Chen, C. & Breiman, L. Using random forest to learn imbalanced data. University of California, Berkeley (2004).
40. Hara, K., Kataoka, H. & Satoh, Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6546-6555 (2018).
41. Youden, W. J. Index rating for diagnostic tests. *Cancer* **3**, 32–35 (1950).
42. Salzberg, S. L. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min. Knowl. Disc.* **1**, 317–328 (1997).
43. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **10**, 1895–1923 (1998).
44. Lawson, R. Small sample confidence intervals for the odds ratio. *Commun. Stat-Simul. C* **33**, 1095–113 (2004).
45. Upton, G. J. G. Fisher's exact test. *J. R. Statist. Soc. A* **155**, 395–402 (1992).
46. Roos, E. M. & Arden, N. K. Strategies for the prevention of knee osteoarthritis. *Nat. Rev. Rheumatol.* **12**, 92–101 (2016).
47. Menashe, L. *et al*. The diagnostic performance of MRI in osteoarthritis: a systematic review and meta-analysis. *Osteoarthritis Cartilage* **20**, 13–21 (2012).
48. Collins, J. E. *et al*. Semi-quantitative imaging biomarkers of knee osteoarthritis progression: data from the FNIH OA biomarkers consortium. *Arthritis Rheumatol.* **68**, 2422–2431 (2016).
49. Khan, T. *et al*. ACL and meniscal injuries increase the risk of primary total knee replacement for osteoarthritis: a matched case-control study using the clinical practice research datalink (CPRD). *Br. J. Sports Med.* **53**, 965–968 (2019).
50. Simon, D. *et al*. The relationship between anterior cruciate ligament injury and osteoarthritis of the knee. *Adv. Orthop.* **2015**, 928301, https://doi.org/10.1155/2015/928301 (2015).
51. Diederichs, G., Issever, A. S. & Scheffler, S. MR imaging of patellar instability: injury patterns and assessment of risk factors. *RadioGraphics* **30**, 961–81 (2010).
52. Moatshe, G. *et al*. High prevalence of knee osteoarthritis at a minimum 10-year follow-up after knee dislocation surgery. *Knee Surg. Sports Traumatol. Arthrosc* **25**, 3914–22 (2017).
53. Figueroa, D., Garin, A. & Figuera, F. Total knee replacement in patients with osteoarthritis and concomitant inveterate patellar dislocation. *Arthroplast. Today* **5**, 68–72 (2019).
54. Souza, R. B. & Doan, R. Anatomy and physiology of the knee in Advances in MRI of the Knee for Osteoarthritis (ed. Majumdar, S.) 1-26 (World Scientific, 2010).
55. Spina, A. A. The plantaris muscle: anatomy, injury, imaging, and treatment. *J. Can. Chiropr. Assoc.* **51**, 158–65 (2007).
56. Zetaruk, M. & Hyman, J. Leg injuries in Clinical Sports Medicine (ed. Frontera, W. R., Herring, S. A., Micheli, L. J., Silver, J. K. & Young, T. P.) 441-57 (Saunders, 2007).

Acknowledgements

Author contributions

A.T. designed and optimized the pipeline, performed the experiments, and wrote the manuscript. J.L. modified the DenseNet architecture to allow for 3D volumetric input, guided development of coding infrastructure in initiation of the project, provided feedback regarding appropriate statistical tests, and edited the manuscript. V.P. and S.M. guided the study design and provided valuable feedback at all stages of the project. V.P. and S.M. also originated the idea for and supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-63395-9.

Correspondence and requests for materials should be addressed to S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.