Radiology

Prediction of Total Knee Replacement and Diagnosis of Osteoarthritis by Using Deep Learning on Knee Radiographs: Data from the Osteoarthritis Initiative

Kevin Leung, BS • Bofei Zhang, BS • Jimin Tan, BS • Yiqiu Shen, MS • Krzysztof J. Geras, PhD • James S. Babb, PhD • Kyunghyun Cho, PhD • Gregory Chang, MD, MBA • Cem M. Deniz, PhD

From the Courant Institute of Mathematical Sciences (K.L., K.C.) and Center for Data Science (B.Z., J.T., Y.S., K.J.G., K.C.), New York University, New York, NY; The Bernard and Irene Schwartz Center for Biomedical Imaging (K.J.G., J.S.B., C.M.D.) and Department of Radiology (K.J.G., J.S.B., G.C., C.M.D.), New York University Langone Health, 660 1st Ave, New York, NY 10016. Received September 17, 2019; revision requested October 30; revision received April 8, 2020; accepted April 30. **Address correspondence to** C.M.D. (e-mail: *cem.deniz@nyulangone.org*).

Study supported by National Institutes of Health (R01 AR074453).

Conflicts of interest are listed at the end of this article.

See also the editorial by Richardson in this issue.

Radiology 2020; 296:584–593 • https://doi.org/10.1148/radiol.2020192091 • Content codes: MK IN

Background: The methods for assessing knee osteoarthritis (OA) do not provide enough comprehensive information to make robust and accurate outcome predictions.

Purpose: To develop a deep learning (DL) prediction model for risk of OA progression by using knee radiographs in patients who underwent total knee replacement (TKR) and matched control patients who did not undergo TKR.

Materials and Methods: In this retrospective analysis that used data from the OA Initiative, a DL model on knee radiographs was developed to predict both the likelihood of a patient undergoing TKR within 9 years and Kellgren-Lawrence (KL) grade. Study participants included a case-control matched subcohort between 45 and 79 years. Patients were matched to control patients according to age, sex, ethnicity, and body mass index. The proposed model used a transfer learning approach based on the ResNet34 architecture with sevenfold nested cross-validation. Receiver operating characteristic curve analysis and conditional logistic regression assessed model performance for predicting probability and risk of TKR compared with clinical observations and two binary outcome prediction models on the basis of radiographic readings: KL grade and OA Research Society International (OARSI) grade.

Results: Evaluated were 728 participants including 324 patients (mean age, 64 years ± 8 [standard deviation]; 222 women) and 324 control patients (mean age, 64 years ± 8; 222 women). The prediction model based on DL achieved an area under the receiver operating characteristic curve (AUC) of 0.87 (95% confidence interval [CI]: 0.85, 0.90), outperforming a baseline prediction model by using KL grade with an AUC of 0.74 (95% CI: 0.71, 0.77; $P < .001$). The risk for TKR increased with probability that a person will undergo TKR from the DL model (odds ratio [OR], 7.7; 95% CI: 2.3, 25; $P < .001$), KL grade (OR, 1.92; 95% CI: 1.17, 3.13; $P = .009$), and OARSI grade (OR, 1.20; 95% CI: 0.41, 3.50; $P = .73$).

Conclusion: The proposed deep learning model better predicted risk of total knee replacement in osteoarthritis than did binary outcome models by using standard grading systems.

Online supplemental material is available for this article.

Osteoarthritis (OA) is the most common form of arthritis, diagnosed by clinical joint symptoms and radiographic findings (1,2). It is a major cause of physical disability in the elderly. In the United States, 14 million people aged 25 years and older have symptomatic knee OA (3), and more than half diagnosed will undergo primary total knee replacement (TKR) before death, with over 600 000 TKRs performed each year (3,4).

Clinical OA symptoms include joint pain, stiffness, and decreased range of motion. Radiographic OA is diagnosed by using a grading system such as the Kellgren-Lawrence (KL) grade (1) or OA Research Society International (OARSI) atlas (5) on the basis of the assessment of osteophytes and joint space narrowing. The presence of definite osteophytes with possible joint space narrowing (ie, KL grade ≥ 2) defines the radiographic knee OA in KL system. In the OARSI atlas, the radiographic knee OA is defined by any one of the following three separate criteria: joint space narrowing grade 2 or greater, sum of osteophyte grades 2 or greater, or joint space narrowing grade 1 and osteophyte grade 1. However, radiographic knee OA grading systems have multiple versions with no uniform agreement (6).

Automated methods for diagnosing knee OA from radiographs include a distance-based active shape model that calculates the geometric parameters between the tibia and femur (7). Transfer learning applied to a convolutional neural network pretrained on ImageNet (8) demonstrated state-of-the-art multiclass accuracy of approximately 67% in predicting KL grade from radiographs (9,10). In addition, for structural OA progression, Lazzarini et al (11) achieved a maximum area under the receiver operating characteristic curve (AUC) of 0.790 in predicting 30-month incidence of knee OA in a cohort

Abbreviations

AUC = area under the receiver operating characteristic curve, BMI = body mass index, CI = confidence interval, DL = deep learning, KL = Kellgren-Lawrence, OA = osteoarthritis, OAI = OA Initiative, OARSI = OA Research Society International, OR = odds ratio, TKR = total knee replacement, WOMAC = Western Ontario and McMaster Universities OA Index

Summary

A multitask deep learning model based on knee radiographs accurately classified patients with osteoarthritis at high risk of total knee replacement compared with binary outcome models that used standard grading systems.

Key Results

- A multitask deep learning model accurately predicted osteoarthritis progression in patients who would require a total knee replacement within 9 years (odd ratio, 7.7; 95% confidence interval: 2.3, 25; $P < .001$).
- The model also predicted Kellgren-Lawrence grade at levels acceptable compared with human graders (κ coefficient, 0.78).
- Our model outperformed an outcome model that used Kellgren-Lawrence grade (area under the receiver operating characteristic curve, 0.87 vs 0.74; $P < .001$).

of overweight middle-aged women by using a random forest algorithm on clinical variables and baseline KL grade.

Our study investigated the use of convolutional neural networks on baseline bilateral posteroanterior fixed-flexion knee radiographs to automatically predict structural OA progression and simultaneously diagnose radiographic OA. The likelihood of a patient undergoing a TKR within 9 years and KL grade was used for predicting the OA progression and diagnosing radiographic OA, respectively. The purpose of our study was to develop a deep learning (DL) prediction model for risk of OA progression by using knee radiographs from patients who underwent TKR and matched control patients who did not undergo TKR.

Materials and Methods

Data Collection

This retrospective study used data from the Osteoarthritis Initiative (OAI), which is a multicenter, longitudinal, prospective observational study of knee OA (12). OAI recruited 4796 participants for collecting clinical, imaging, and biospecimen data to study OA. The observational OAI study was performed between February 2004 and October 2015. Baseline Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (13) pain score and knee-related Quality of Life from Knee Injury and OA Outcome Score (14) provide the clinical observations used in our study. WOMAC pain score and Quality of Life from Knee Injury and Osteoarthritis Outcome Score are associated with TKR (15). WOMAC is a health status questionnaire that is used to assess the condition of patients (knee pain, stiffness, and knee-related physical function). The Knee Injury and Osteoarthritis Outcome Score is used to assess the opinion from patients about their knee and associated problems. In OAI, the Knee Injury and Osteoarthritis Outcome Score was administered separately to assess knee symptoms and function with different activity conditions (eg, during sports and recreation) than are evaluated by the WOMAC. Known clinical risk factors (16,17) included Heberden nodes (bony enlargement of 1+ distal interphalangeal joint in both hands), family history (a TKR for OA in a biologic parent or sibling), history of knee injury (difficulty walking for at least a week), and contralateral WOMAC pain score. The OAI data set excludes patients with MRI contraindications, inflammatory arthropathies, bilateral TKR, positive pregnancy test, and comorbid conditions that might interfere with the ability to participate in the study.

The OAI data set includes bilateral posteroanterior fixed-flexion knee radiographs from patients collected for 8 years and the knee replacement outcomes data collected during 9 years. Semiquantitative KL and OARSI grades (in ordinal scales) and quantitative minimum joint space width (in millimeters) in the medial compartment were assessed by auxiliary radiography studies. KL and OARSI grades were obtained from project 15 from file kxr_sq_bu00 version 0.8 (18) and minimum joint space width measurements were obtained from project 16 from file kxr_qjsw_duryea00 version 0.8 (19). The images were centrally graded by two expert readers who were blinded to each other's readings and to clinical data from the patient (see Appendix D of reference 15 for the flowchart of the reading process). The test-retest reliability of these readings was good, with κ coefficient values of 0.70–0.80 for KL grades, 0.75–0.88 for joint space narrowing variables, and 0.69–0.82 for osteophyte variables (20). The reliability of minimum joint space width measurements was high (intraclass correlation coefficient, 0.984) (21). Semiquantitative radiographic readings were available for the matched subcohort in the OAI data set. Participants were recruited at four clinical sites, and the Health Insurance Portability and Accountability Act–complaint study was approved by the institutional review board at each of the sites. All participants gave written informed consent.

Cohort Selection

A balanced case-control cohort was selected by matching patients and control patients by using the baseline confounding variables: age, body mass index (BMI), sex, and ethnicity. We defined patients as individuals who underwent a TKR in either knee after the baseline enrollment date and control patients as individuals who appeared at the 108-month follow-up visit and did not undergo a TKR in either knee. Each patient was matched to a control patient (without replacement) who was the same sex, ethnicity, and age, and with an additional constraint on the baseline BMI within a 10% tolerance. The data set from case-control pairs contained either the left or right knee radiographs from each patient and control patient. If a patient underwent TKR in both knees during OAI data collection, we included the knee that first underwent TKR. The inclusion criterion was to be enrolled in the OAI study. Exclusion criteria were the presence of a knee replacement at the baseline visit, missing 108-month follow-up visit and baseline demographics, partial knee replacement during the OAI study, and not to match a patient or a control.

DL Model Development and Evaluation

We used a multiple-task (hereafter, referred to as multitask) learning on a DL model for predicting simultaneously the TKR outcome of the patients and control patients, and KL grade of the radiographs. Multitask learning improves the generalization of a single-learning task (22). The output of the proposed multitask DL model trained with transfer learning provided both the prediction of TKR outcome and KL grade. We used a publicly available ResNet with 34 layers (ResNet34) model (23) with PyTorch (version 1.0.1; *pytorch.org*). We compared multitask DL model trained with transfer learning with single-task DL models trained either with transfer learning or by using random weight initialization. Details of the DL models, training with nested cross-validation, and regions of visualization are in Appendix E1 (online). The source code for this study is available at *https://github.com/denizlab/oai-xray-tkr-klg*.

We compared the performance of DL models with two binary outcome prediction models on the basis of radiographic readings: KL grade and OARSI grade. By using a KL grade threshold of 2, we defined patients more likely to undergo a TKR with probability 1 if the KL grade is 2 or greater and 0 otherwise (referred to as KL model). We developed a binary OARSI model by defining the patients more likely to receive a TKR with probability 1 if they met OARSI atlas radiographic knee OA definition criterion and 0 otherwise.

We computed five AUCs for the five classes in our learning task for KL grade prediction, each time treating one of the five KL grade classes as a positive finding and the remaining four as negative findings. In addition to individual AUCs, the macro average of the five AUCs measured KL grade prediction task. Cohen κ coefficient assessed the agreement between predictions of the multitask DL model trained with transfer learning and expert annotations provided in the OAI data set.

Statistical Analysis

The output of the multitask DL model trained with transfer learning was combined with clinical risk factors in each patient to develop an outcome prediction model for knee OA. Conditional logistic regression was used to assess the risk of TKR related to the several clinical observations and radiographic readings (ie, risk factors) by using clogit function (R package version 3.1–8; *https://CRAN.R-project.org/package=survival*) with "exact" method. Each risk factor was examined separately with univariable and multivariable analyses. Crude and adjusted odds ratios (ORs) of TKR were calculated for each risk factor. Crude OR assesses the effect of a given factor when it is used as the only predictor of outcome. Adjusted OR assesses the effect of the given factor adjusted for the effects of all other factors included in a multivariable model to predict the outcome. Paired *t* tests assessed differences in WOMAC pain score and Quality of Life from Knee Injury and Osteoarthritis Outcome Score between patients and control patients. The DeLong test (24) was used to assess significance (at 5% level) of the AUC difference between the baseline KL model and other models (pROC package version 1.15.3). Confidence intervals (CIs) were computed across 5000 bootstrap samples. All statistical calculations were performed by using software (R version 3.6.0.; *r-project.org*). *P* values less than .01 indicated statistical significance.

Results

Participant Characteristics

Our one-to-one case-control matching approach resulted in a study cohort of 728 participants. Figure 1 provides a flowchart of our method to select the case-control pairs. We excluded 63, 33, and 12 patients because they had a knee replacement at the baseline visit, underwent a partial knee replacement during the OAI study, and did not match with control patients, respectively. We excluded 1341 patients because they missed the 108-month follow-up visit, four patients were excluded because they did not have baseline demographics, and 2615 control patients were excluded because they did not match with patients. There were no differences between patient and control groups regarding age ($P > .99$), height ($P = .91$), weight ($P = .20$), and BMI ($P = .20$). Mean patient age was 64 years ± 8 (standard deviation). Patients were predominantly women (444 of 728; 61%) and were predominantly overweight (BMI, 29.8 kg/m ± 4.6). Study cohort characteristics are in Table 1.

KL grades for the study matched subcohort were as follows: grade 0, 155 knees; grade 1, 89 knees; grade 2, 186 knees; grade 3, 187 knees; and grade 4, 111 knees. From the baseline visit, the case-defining TKR date was within 12 months for 20 knees (20 of 364; 5.5%), 24 months for 35 knees (35 of 364; 9.6%), 36 months for 38 knees (38 of 364; 10.4%), 48 months for 50 knees (50 of 364; 13.7%), 60 months for 52 knees (52 of 364; 14.3%), 72 months for 47 knees (47 of 364; 12.9%), 84 months for 42 knees (42 of 364; 11.5%), 96 months for 41 knees (41 of 364; 11.3%), and 108 months for 39 knees (39 of 364; 10.7%).

Regarding the WOMAC pain score (range, 0–20), patients differed from control patients (4.8 ± 3.8 vs 2.0 ± 2.8, respectively; $P < .001$). Similarly, patients had lower Quality of Life from Knee Injury and Osteoarthritis Outcome Score (range, 0–100) than did control patients (52.4 ± 20.8 vs 71.5 ± 20.1, respectively; $P < .001$). The distribution of WOMAC pain and Quality of Life from Knee Injury and Osteoarthritis Outcome scores per KL grade is in Table 2.

Convolutional Neural Network Model

Receiver operating characteristic curve analysis of DL models is in Table 3 and Figure 2. The multitask DL model trained with transfer learning outperformed other image-based TKR outcome prediction models with an AUC of 0.87 (95% CI: 0.85, 0.90), which was a higher AUC compared with the KL model, which had an AUC of 0.74 (95% CI: 0.71, 0.77; $P < .001$). The KL model underperformed single-task DL models (AUCs, 0.84 [95% CI: 0.81, 0.86] and 0.86 [95% CI: 0.84, 0.89]; $P < .001$). Comparing DL-based models, the multitask DL model trained with transfer learning had a higher AUC than single-task DL model trained by using random weight initialization ($P < .001$) and single-task DL model trained with transfer learning ($P = .17$). The OARSI model showed an AUC of 0.75 (95% CI: 0.72, 0.78; $P = .35$). The KL model had the highest sensitivity (334 of 364; 91%) and

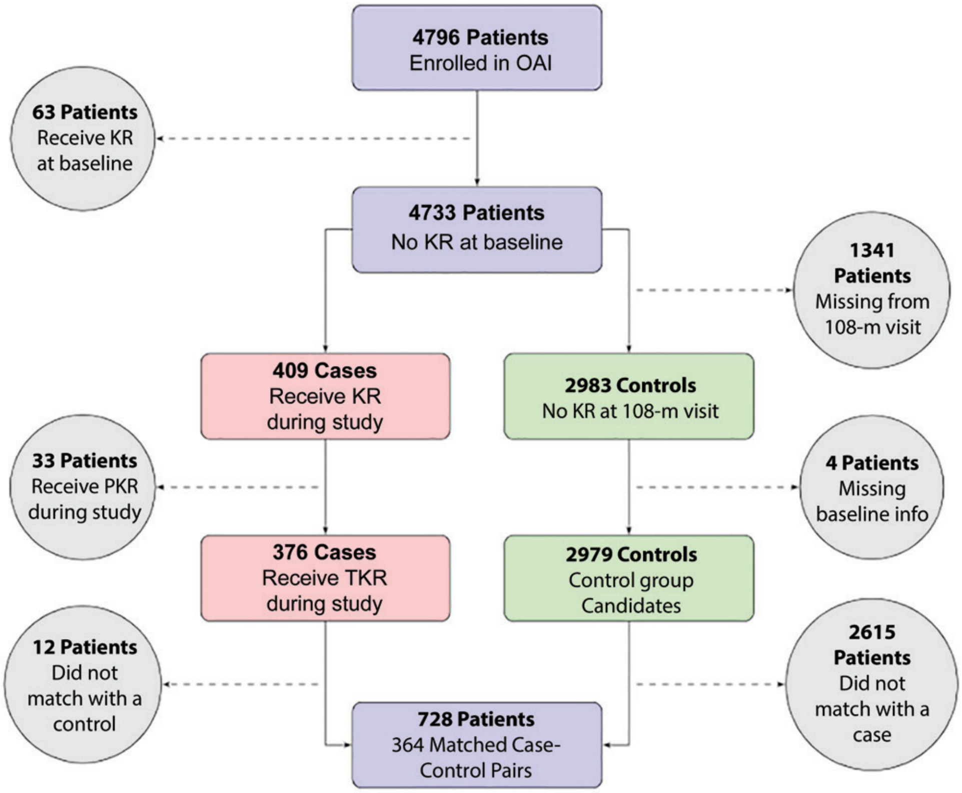

Figure 1: Flowchart for case-control matching method. Solid arrows represent patients included in our case-control cohort, whereas dashed arrows and circular nodes represent patients excluded from our case-control cohort. One-to-one case-control matching uses a combination of propensity score and exact matching on the baseline clinical variables of age, body mass index, sex, and ethnicity to create the final cohort. KR = knee replacement, OAI = Osteoarthritis Initiative, PKR = partial knee replacement, TKR = total knee replacement.

Table 1: Summary Statistics for Demographic Variables in Matched Case-Control Cohort

	Men			Women		
Parameter	Patients	Control Patients	*P* Value	Patients	Control Patients	*P* Value
No. of patients	142	142		222	222	
Mean age (y)	64 ± 8	64 ± 8	>.99	64 ± 8	64 ± 8	>.99
Mean height (m)	1.76 ± 0.06	1.77 ± 0.06	.73	1.62 ± 0.06	1.62 ± 0.06	.53
Mean weight (kg)	93.5 ± 14.0	92.1 ± 12.6	.40	79.0 ± 15.0	77.3 ± 13.8	.24
Mean BMI (kg/m^2)	29.9 ± 3.8	29.4 ± 3.4	.31	30.1 ± 5.4	29.7 ± 4.8	.37
Ethnicity*						
White	129	129		182	182	
African American	11	11		35	35	
Asian	0	0		2	2	
Other nonwhite	2	2		3	3	
Kellgren-Lawrence grade*						
0	4	67		8	76	
1	11	26		11	41	
2	24	28		55	79	
3	44	19		102	22	
4	59	2		46	4	

Note.—Mean data are ± standard deviation. *P* value compares the difference in means between the case and control groups for each confounding variable. BMI = body mass index.

* Number of patients are shown.

Table 2: Distribution of Western Ontario and McMaster Universities Osteoarthritis Index Pain and Quality of Life from Knee Injury and Osteoarthritis Outcome Score Scores per Kellgren-Lawrence Grade

	WOMAC Pain			KOOS QoL		
KL Grade	Patients	Control Patients	*P* Value	Patients	Control Patients	*P* Value
0	0.1 ± 0.3	1.4 ± 2.1	<.001	77.6 ± 20.2	75.4 ± 18.8	.72
1	3.6 ± 4.5	1.4 ± 2.3	.04	58.0 ± 22.2	73.2 ± 19.6	.007
2	4.6 ± 4.0	2.5 ± 3.4	<.001	52.9 ± 21.5	68.1 ± 20.0	<.001
3	4.9 ± 3.5	3.4 ± 3.3	.02	53.0 ± 20.5	64.9 ± 23.4	.005
4	5.8 ± 3.5	3.0 ± 3.8	.13	47.1 ± 17.9	67.7 ± 16.5	.03

Note.—Mean data are ± standard deviation. *P* values indicate difference in means between the case and control groups for Western Ontario and McMaster Universities Osteoarthritis Index pain and Knee-related Quality of Life from Knee injury and Osteoarthritis Outcome scores grouped by Kellgren-Lawrence grade. KL = Kellgren-Lawrence, KOOS QoL = Knee-related Quality of Life from Knee Injury and Osteoarthritis Outcome Score, WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index.

Table 3: Comparison of Image-based Total Knee Replacement Prediction Models

Model	AUC	*P* Value	Specificity (%)	Sensitivity (%)
Kellgren-Lawrence	0.74 (0.71, 0.77)	…	58 (52, 63)	91 (87, 93)
OARSI	0.75 (0.72, 0.78)	.35	59 (54, 64)	90 (87, 93)
DL	0.84 (0.81, 0.86)	<.001	85 (80, 88)	70 (65, 75)
DL-TL	0.86 (0.84, 0.89)	<.001	85 (81, 89)	77 (73, 82)
DL-TL-MT	0.87 (0.85, 0.90)	<.001	77 (72, 81)	83 (78, 86)

Note.—Data in parentheses are 95% confidence intervals. The *P* value of the difference between areas under the receiver operating characteristic curve from each model to the baseline Kellgren-Lawrence model was calculated by using the DeLong test (24), and it was adjusted by using the Holm correction (25) for multiple comparisons. Because we used one-to-one matching between patients and control patients, the sample sizes in patients and control groups do not reflect the real prevalence of the disease. Therefore, positive and negative predicted values cannot be estimated accurately, and they are not presented. AUC = area under the receiver operating characteristic curve, DL = deep learning, MT = multitask, OARSI = Osteoarthritis Research Society International, TL = transfer learning.

the lowest specificity (210 of 364; 58%). The single-task DL model trained with transfer learning had the highest specificity (311 of 364; 85%) and the single-task DL model trained by using random weight initialization had the lowest sensitivity (256 of 364; 70%). Unlike conventional image-based models (ie, the KL and OARSI models), DL models balance sensitivity and specificity. Figure 3 shows a zombie plot (26) to depict the performance of five outcome prediction models. All models are within the boomerang-shaped area formed by white and light-gray zones in the upper left corner of receiver operating characteristic plot, defining an acceptable efficacy for each prediction model. Specifically, DL models trained with transfer learning are mostly within the optimal zone (white zone in the upper left corner of the receiver operating characteristic plot), and they are credible predictors for the TKR outcome in patients and control patients.

Figure 4 shows regions with high impact on the decision of the multitask DL model trained with transfer learning. The examples provided in Figure 4 are the control patients with KL grade 0 [$p\ (y|x) < 0.001$], the patient with KL grade 2 [$p\ (y|x) = 0.991$], and the control with KL grade 2 [$p\ (y|x) = 0.004$]. The multitask DL model trained with transfer learning focused on regions near the knee joint space, which suggests that our network learned features related to the knee joint and bones to classify samples between the two groups.

Receiver operating characteristic curve analysis of KL grade predictions by using the multitask DL model trained with transfer learning is in Table 4. The model achieved the best identification of the radiographs with KL grade 4 (AUC, 0.99; 95% CI: 0.97, 1.00). The most challenging identification of KL grade was for KL grade 1 radiographs (AUC, 0.80; 95% CI: 0.76, 0.83). The macro average of the five AUCs for the KL grade prediction task was 0.91. The average multiclass accuracy was 72.7% (529 of 724). The weighted Cohen κ coefficient and mean square error for multiclass prediction were 0.78 (95% CI: 0.75, 0.81) and 0.47, respectively. The confusion matrix is in Figure E1 (online).

Receiver operating characteristic curve analysis of individual test sets is in Table 5 for multitask DL model trained with transfer learning and KL model. The multitask DL model trained with transfer learning achieved AUCs ranging from 0.85 to 0.90 and outperformed the KL model (AUC, 0.74–0.81). At univariable analysis, the multitask DL model trained with transfer learning yielded a stronger association with the risk of TKR (OR, 48–130; $P < .001$) compared with the KL model (OR, 2.9–8.9; $P < .001$).

Association of Risk Factors with TKR

Table 6 shows ORs of TKR for risk factors in the matched subcohort of 728 patients and control patients. The output of multitask DL model trained with transfer learning yielded

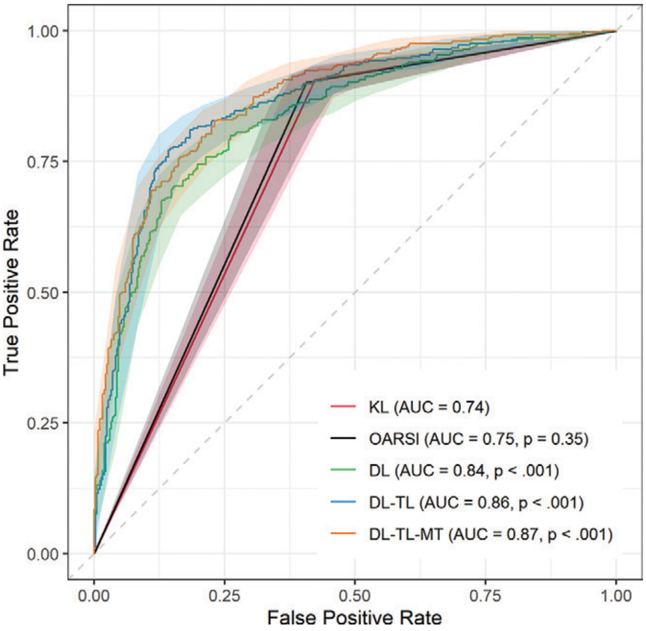

Figure 2: Receiver operating characteristic curves for image-based binary outcome prediction models. Outcome is defined as undergoing a total knee replacement (TKR) within 9 years or not. Area under the receiver operating characteristic curve (AUC) values are for the outcome of predicting TKR. Kellgren-Lawrence (KL) model provides a baseline model for assessing the significant differences between prediction models by using Delong test (24) with Holm correction (25) for multiple comparisons. DL = deep learning, MT = multitask, OARSI = Osteoarthritis Research Society International, TL = transfer learning.

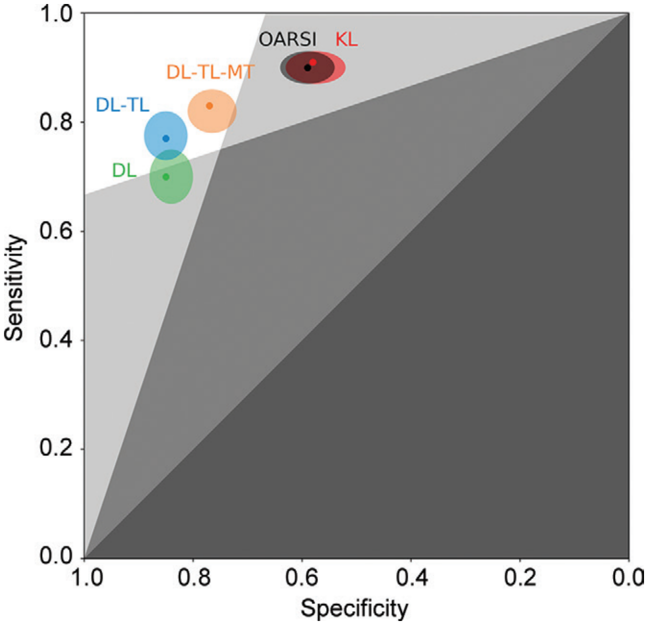

Figure 3: Combined zombie plot (ie, receiver operating characteristic plot divided into zones of mostly bad imaging efficacy) for image-based outcome prediction methods: Kellgren-Lawrence (KL), Osteoarthritis Research Society International (OARSI), deep learning (DL) model, DL-transfer learning (TL) model, and DL-TL-multitask (MT) model.

the highest OR in both univariable and multivariable analysis compared with other risk factors with an OR of 82 (95% CI: 34, 194) and 7.7 (95% CI: 2.3, 25), respectively. The addition of predicting KL grade task into the DL model by a multitask learning approach improved the quality of TKR predictions as depicted by the attenuated odds of TKR from single-task DL model trained with transfer learning (crude and adjusted OR, 55 and 6.0, respectively).

Multivariable and Univariable Analyses

At multivariable analysis, the odds of TKR increased with an increase on the output (ie, the probability of TKR within 9 years) of the multitask DL model trained with transfer learning (OR, 7.7; *P* < .001), KL grade (OR, 1.9; *P* = .009), BMI (OR, 1.4; *P* = .04), and a decrease on the Quality of Life from Knee Injury and Osteoarthritis Outcome Score (OR, 0.98; *P* = .02). The order of risk association with TKR from strongest to the weakest was multitask DL model trained with transfer learning, KL grade, BMI, and Quality of Life from Knee Injury and Osteoarthritis Outcome Score.

The univariable analysis of OARSI-grade yielded an association with the risk of TKR (*P* < .001) but not at multivariable analysis (*P* = .73). Similarly, at univariable analysis, WOMAC pain score was associated with the risk of TKR (*P* < .001) but not at multivariable analysis (*P* = .26). The risk factors derived from the contralateral knee were associated with TKR separately: KL grade, OARSI grade, and WOMAC pain score with ORs of 1.8 (*P* < .001), 3.6 (*P* < .001), and 1.12 (*P* < .001), respectively. At multivariable analysis, the TKR association with the risk factors from contralateral knee was attenuated: KL grade, OARSI grade, and WOMAC pain score with ORs of 1.1 (*P* = .79), 1.1 (*P* = .80), and 1.01 (*P* = .86), respectively. As shown in Table 7, patients with previous injury exhibited a higher risk for TKR compared with patients without a previous injury (*P* = .03). Similarly, at a univariable analysis, minimum joint space width yielded an association with the risk of TKR (*P* < .001) but not at multivariable analysis (*P* = .60). The family history and Heberden node yielded ORs of 1.90 (*P* = .15) and 1.15 (*P* = .27), respectively.

Discussion

In the United States, knee osteoarthritis (OA) affects nearly 27 million Americans, and there is a growing need for disease-modifying therapies that prevent or delay the need for total knee replacement (TKR). However, the methods for assessing knee OA do not provide enough comprehensive information to make robust and accurate outcome predictions (28). We developed a deep learning (DL) model on the basis of convolutional neural networks for the prediction of OA progression leading to TKR within 9 years by using baseline radiographs from a matched case-control subcohort of 728 patients. Learning for Kellgren-Lawrence (KL) grade prediction and TKR outcome prediction tasks jointly improved the performance of a DL model aimed at predicting only the TKR outcome. Our proposed model resulted in a higher area under the receiver operating characteristic curve (AUC, 0.87; 95% confidence interval [CI]: 0.85, 0.90) compared with a baseline binary outcome model on the basis of radiography readings of KL grade (AUC, 0.74; 95%

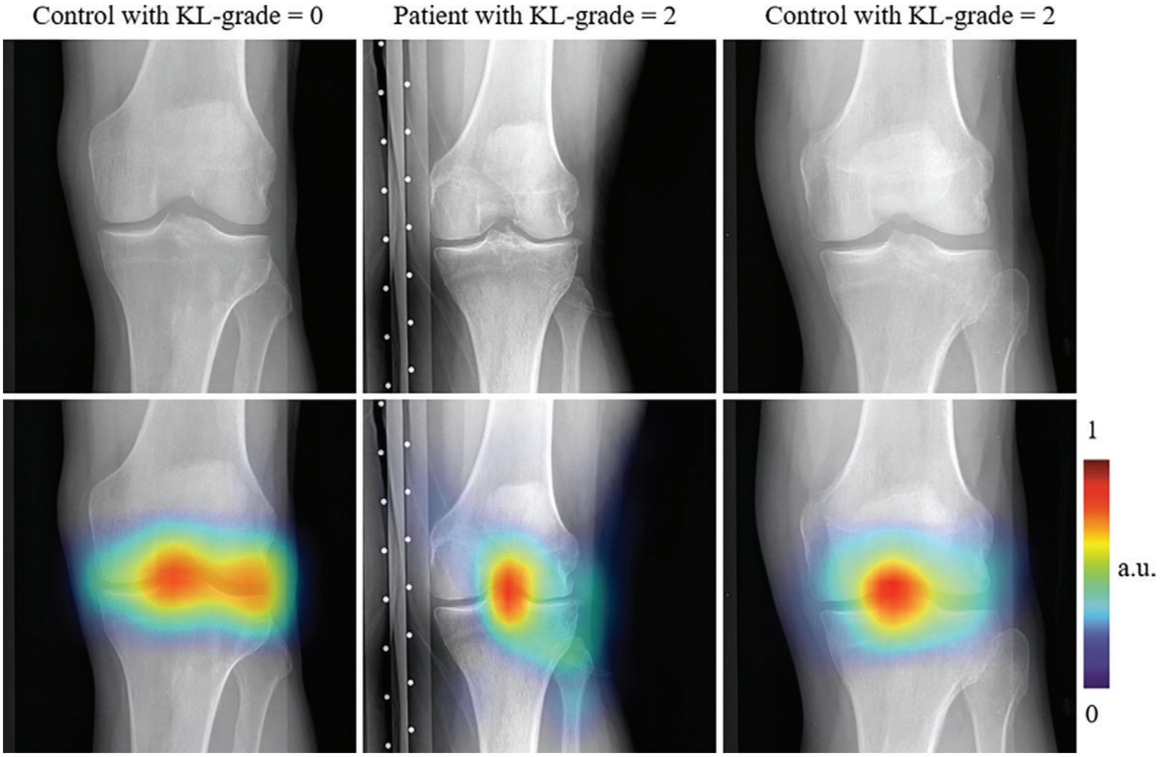
Control with KL-grade = 0
Patient with KL-grade = 2
Control with KL-grade = 2
1
a.u.
0

Figure 4: Knee radiographs overlaid with the heatmaps obtained by using the Grad-CAM method (27) (Appendix E1 [online]) show regions affecting the prediction of the multitask deep learning (DL) model trained with transfer learning. Colored regions show areas where multitask DL model trained with transfer learning focuses on decisions regarding the probability of total knee replacement in the patient within 9 years. Each column represents radiographs and heatmaps from a 75-year-old male control patient with Kellgren-Lawrence (KL) grade 0 (top and bottom left; body mass index [BMI], 26.0 kg/m^2), a 76-year-old female control patient with KL grade 2 (top and bottom middle; BMI, 28.4 kg/m^2), and a 70-year-old male patient with KL grade 2 (top and bottom middle; BMI, 27.6 kg/m^2). Patients underwent total knee replacement and control patients did not undergo total knee replacement within 9 years.

Table 4: AUC for the Kellgren-Lawrence Grade Prediction from Multitask Deep Learning Model Trained with Transfer Learning

KL Grade	AUC	*P* Value
0	0.93 (0.91, 0.95)	…
1	0.80 (0.76, 0.83)	<.001
2	0.88 (0.85, 0.91)	.007
3	0.96 (0.95, 0.97)	.03
4	0.99 (0.97, 1.00)	<.001

Note.—Data in parenthesis are 95% confidence intervals. The *P* value of the difference between receiver operating characteristic curves from each KL grade prediction to the Kl grade 0 prediction was calculated using the DeLong test (24), and it was adjusted using the Holm correction (25) for multiple comparisons. Each AUC treated the KL grade class as a positive and the remaining four as negative. AUC values are for the outcome of predicting the KL grade. AUC = area under the receiver operating characteristic curve, KL = Kellgren-Lawrence.

CI: 0.71, 0.77; $P < .001$). In our analysis of risk factors, we found that the DL model predicted TKR more strongly than other risk factors (odds radio [OR], 7.7; $P < .001$) versus KL grade (OR, 1.9; $P = .009$) and OA Research Society International grade (OR, 1.2; $P = .73$).

Previous approaches of DL for knee OA assessment focused on the prediction of KL grade directly from knee radiographs (9,10). However, we developed a model to predict OA progression outcomes directly from baseline radiographs with an additional KL grade prediction task. Our model, on the basis of multitask DL, provided the prediction of TKR and radiographic KL grade readings from each radiograph simultaneously. Our KL grade prediction results are comparable with previous approaches that focused only on diagnosing knee OA (9,10). Moreover, unlike previous methods of applying DL models to assess knee OA by predicting KL grade, our approach used image-based features to achieve higher outcome prediction accuracy of undergoing TKR.

Numerous clinical, laboratory, and imaging assessments aimed to identify prognostic risk factors for prediction of knee OA progression (29). Multiple definitions of OA progression exist, including a clinically important outcome of undergoing TKR. Many prediction models for OA progression risk use logistic regression (16,30). Three Nottingham knee OA risk prediction models developed by Zhang et al (16) used an internal retrospective cohort of 424 patients. These models were tested on a subset of the OAI data set containing 1489 people, of whom 162 were diagnosed as having a radiographic knee OA at 3rd-year follow-up. These models resulted in AUCs of 0.60, 0.60, and 0.52 in discriminating the risk of knee OA for incidence of radiographic OA (KL ≥ 2), incidence of

Table 5: Areas Under the Receiver Operating Characteristic Curve of Total Knee Replacement Prediction Models and Crude Odds Ratios

Model	Test Set No. 1	Test Set No. 2	Test Set No. 3	Test Set No. 4	Test Set No. 5	Test Set No. 6	Test Set No. 7
DL-TL-MT							
AUC	0.90 (0.84, 0.96)	0.88 (0.81, 0.94)	0.85 (0.77, 0.92)	0.86 (0.78, 0.93)	0.88 (0.81, 0.94)	0.84 (0.76, 0.91)	0.87 (0.79, 0.93)
Crude OR	67 (8, 579)	51 (8, 345)	48 (6, 366)	60 (7, 498)	130 (9, 1729)	87 (7, 1012)	114 (9, 1418)
KL							
AUC	0.80 (0.72, 0.88)	0.80 (0.73, 0.87)	0.74 (0.66, 0.82)	0.77 (0.68, 0.85)	0.77 (0.68, 0.85	0.78 (0.70, 0.86)	0.81 (0.73, 0.88)
Crude OR	8.9 (2.3, 34)	4.0 (1.9, 8.3)	3.5 (1.9, 6.6)	7.3 (2.4, 22)	2.9 (1.6, 5.1)	3.3 (1.7, 6.3)	3.8 (1.8, 7.7)

Note.—Data in parentheses are 95% confidence intervals. Analysis of the performance of outcome prediction models used seven disjoint groups defined for nested cross validation. In nested cross validation, stratified random sampling was used to partition the 728 patients and control patients in the matched subcohort into seven disjoint groups and each group consisted of 52 patients with total knee replacement and 52 control patients. Each of the seven groups served as a test set to assess the performance of a prediction model (outer loop) and test set number specifies which disjoint group was used to analyze the performance of outcome prediction models (Appendix E1 [online]). Each test set was not used for either training or validation of the DL models. Test sets did not contribute data to the derivation of the best-fit model, and the test data were independent of the data used to fit the model. Model fit was achieved by using the remaining 312 patients and 312 control patients. All *P* values for the crude odds ratios are less than .001. AUC = area under the receiver operating characteristic curve, DL = deep learning, KL = Kellgren-Lawrence, MT = multitask, OR = odds ratio, TL = transfer learning.

Table 6: Odds Ratios of Total Knee Replacement for Clinical Risk Factors and Radiographic Readings

Parameter	Crude Odds Ratio	Adjusted Odds Ratio	*P* Value[*]
DL-TL[†]	55 (26, 119)	6.0 (2.1, 17)	<.001
DL-TL-MT	82 (34, 194)	7.7 (2.3, 25)	<.001
Kellgren-Lawrence	4.0 (3.0, 5.3)	1.9 (1.2, 3.1)	.009
Body mass index	1.7 (1.4, 2.2)	1.4 (1.0, 1.9)	.04
OARSI	19 (10, 36)	1.2 (0.4, 3.5)	.73
KOOS QoL	0.96 (0.95, 0.97)	0.98 (0.96, 0.997)	.02
WOMAC	1.31 (1.23, 1.39)	1.06 (0.96, 1.18)	.26
Contralateral knee			
Kellgren-Lawrence	1.8 (1.5, 2.1)	1.1 (0.7, 1.7)	.79
OARSI	3.6 (2.6, 5.1)	1.1 (0.4, 3.2)	.80
WOMAC	1.12 (1.06, 1.18)	1.01 (0.92, 1.11)	.87

Note.—Data in parenthesis are 95% confidence intervals. Multivariable analysis was performed by using risk factors from 364 patients and 364 control patients. *P* values are for adjusted odds ratios. All *P* values for the crude odds ratios are <.001 for the risk factors in the table. DL = deep learning, KOOS QoL = Knee-related Quality of Life from Knee injury and Osteoarthritis Outcome Score, MT = multitask, OARSI = Osteoarthritis Research Society International, TL = transfer learning, WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index.

[*] Wald test was used to assess the significance levels of individual risk factors.

[†] DL-TL model is used to demonstrate the effect of addition of Kellgren-Lawrence-grade task into the total knee replacement prediction model only. Adjusted odds ratios presented after the first row are calculated from a multivariable analysis of risk factors and DL-TL-MT model.

symptomatic knee OA (KL ≥ 2 and current pain in the same knee), and progression of knee OA models (KL increased one grade or more), respectively. The study population of Nottingham data was slightly younger and lighter in BMI compared with our cohort. In another study, Joseph et al (30) developed models for OA risk predictions by using a subset of the OAI data set with 641 patients with KL 2 or less and WOMAC pain score of 1 or less. The study population was slightly younger and lighter in BMI compared with our cohort. A baseline risk prediction model resulted in an AUC of 0.60 by using age, sex, and BMI. Adding KL grade and previous injury to the baseline model improved the AUC to 0.67. The model with highest AUC included age, sex, BMI, KL grade, previous injury, and measurements from MRI (whole-organ MRI score and mean cartilage T2 in the medial tibia and medial femur) (AUC, 0.72). Kerkhof et al (31) used a data set from Rotterdam study I (32) to develop a prediction model for radiographic knee OA incidence (defined as KL < 2 at baseline and KL ≥ 2 at follow-up of 9.4 years ± 2.2). The prediction model was tested on Rotterdam study II data set of 69 patients with incident knee OA and 856 control patients by using a follow-up time of 4.1 years ± 0.6. The study population of Rotterdam study II was similar in age and slightly lighter in BMI compared with our cohort. The risk prediction model resulted in an AUC of 0.86 (95% CI: 0.82, 0.90) by using sex, age, BMI, knee pain (defined as pain

Table 7: Odds Ratios of Total Knee Replacement for Clinical Risk Factors and Radiographic Readings from a Subset of the Study Cohort

Risk Factor	No. of Patients	No. of Control Patients	Crude OR	Crude *P* Value	Adjusted OR*	Adjusted *P* Value
Family history	359	353	1.49 (0.99, 2.23)	.06	1.90 (0.79, 4.57)	.15
Previous injury	360	364	1.40 (1.04, 1.90)	.03	0.47 (0.24, 0.93)	.03
Heberden node	363	363	1.01 (0.90, 1.13)	.88	1.15 (0.90, 1.46)	.27
Minimum JSW	355	257	0.70 (0.62, 0.78)	<.001	1.06 (0.86 1.30)	.61

Note.—Data in parenthesis are 95% confidence intervals. Analysis of family history, previous injury, Heberden node and minimum joint space width was performed on a subset of the study cohort because of missing observations. Wald test was used to assess the significance levels of individual risk factors. JSW = joint space width, OR = odds ratio.

* Adjusted odds ratio from multivariable analysis uses clinical risk factors and radiographic readings from Table 5 from 345 patients and 257 control patients.

during the last month during most of the days) and baseline KL score of 0 or 1. The performance of our model cannot be directly compared with previous models because of changes in the outcome definition, cohort selection, and inclusion of different risk factors.

The use of convolutional neural networks to predict the KL grade automatically from radiographs was proposed by Antony et al (9). Performing transfer learning on the OAI data set resulted in a multiclass classification accuracy of 59.6%. Tiulpin et al (10) proposed a siamese convolutional neural network to predict KL grade and tested the developed model on the OAI data set. This approach yielded an average multiclass accuracy of 66.71%, weighted κ coefficient of 0.83, and mean square error of 0.48. Our multitask learning approach provided improved KL grade prediction performance regarding average multiclass accuracy and mean square error. It provided a slight reduced performance for weighted κ coefficient but performed at acceptable levels compared with human raters (33) (weighted κ, 0.56; 95% CI: 0.38, 0.73).

We used only baseline patient information to predict the risk of TKR. However, the OAI data set includes radiographs, clinical observations, and radiographic readings collected multiple times within 8 years. Changes in clinical observations and/or radiographic readings over subsequent years may affect a patient's decision to undergo a TKR. This type of information could be used to improve the predictive capability of DL models.

In the development of DL models, we used bilateral posteroanterior fixed-flexion knee radiographs used by radiologists to grade radiographs with KL grading scheme to diagnose radiographic OA. However, other types of radiographs could identify the progression of the knee OA. Our developed models could not be directly applied to other type of radiographs for predicting TKR outcome. However, they can provide a baseline model to benefit from transfer learning approach by facilitating the use of limited data set size and/or improving the predictive performance.

Radiographic data as part of the OAI are obtained by using standardized methods across sites and are regularly reviewed for quality by the OAI Quality Assurance Center. Variation in image quality still exists and would affect the training of the DL models. However, this variation in image quality would make it more challenging for training the DL model to perform accurately in the test data set. In addition, if the model is to be deployed in clinical practice, training on image data sets with varying quality is a characteristic of a real-world scenario.

In addition to knee radiographs, the OAI data set included knee MRI. We used radiographs and radiographic readings because they are used for clinical diagnosis of radiographic OA. However, the use of DL models to predict the progression of knee OA and to score or grade automatically is not limited to radiographs. Models with three-dimensional MRI can be developed by using an extension of two-dimensional convolutional neural network approaches to three-dimensional data or by using three-dimensional convolutional neural network approaches directly. In the future, MRI-based DL approaches may predict knee OA progression.

Our study had limitations. First, the data set size was limited for training a DL model from scratch. Our case-control cohort included 728 patients and control patients and baseline images from them. There were more than 3000 patients in the OAI data set who did not meet our cohort selection criteria. Because of limited data set size, the experiments were performed by using transfer learning and sevenfold nested cross-validation. Second, the outcome variable was defined as undergoing a TKR within 9 years. Even though the TKR outcome was preferable as a single clinical outcome measure, the decision to undergo a TKR can be affected by the constraints regarding comorbidities, insurance status, and other factors (34,35). Third, we defined the TKR outcome as a binary variable by neglecting the association of the time from baseline to time of TKR. This definition of outcome variable enabled us to train the DL model as a balanced binary classification problem on a matched case-control subcohort. Reformulating the TKR prediction training as a regression problem would enable predicting the time for TKR from baseline, and it could improve the TKR risk prediction. However, the small sample size of patients with TKR could impede the generalization accuracy of the DL models, resulting in suboptimal prediction of the time to TKR. Fourth, we viewed the regions where a DL model focuses to predict the TKR outcome of patients. However, we did not identify the parameters extracted from radiographs. Fifth, we analyzed the effect of clinical measurements and radiographic assessments available in the OAI data set. These clinical risk factors were selected on the basis of their association with the knee OA progression as identified by previous publications that analyzed the risk of knee OA progression (15,16,30). Finally, we applied the DL model to predict the risk of TKR. The output of the DL model is the probability that a person will undergo TKR and our goal was to identify predictive biomarkers and models for OA. In the future, it would also be of interest to apply survival or hazard analysis to the OAI data set to predict not only the risk of TKR, but the actual time to TKR, which would provide useful additional information to clinicians.

In summary, we developed a deep learning (DL) model to predict both the probability of total knee replacement (TKR) within 9 years and the Kellgren-Lawrence grade by using baseline radiographs. Our proposed DL model better predicted risk of TKR in osteoarthritis than did binary outcome models with standard grading systems.

Author contributions: Guarantors of integrity of entire study, B.Z., C.M.D.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, K.L., J.T., K.C., G.C., C.M.D.; experimental studies, Y.S., K.C., C.M.D.; statistical analysis, K.L., B.Z., J.T., K.J.G., J.S.B., C.M.D.; and manuscript editing, B.Z., J.T., Y.S., K.J.G., J.S.B., K.C., G.C., C.M.D.

Acknowledgements: The Osteoarthritis Initiative (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, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories, Novartis Pharmaceuticals Corporation, GlaxoSmithKline, and Pfizer. Private-sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared by using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the National Institutes of Health, or the private funding partners.

Disclosures of Conflicts of Interest: **K.L.** disclosed no relevant relationships. **B.Z.** disclosed no relevant relationships. **J.T.** disclosed no relevant relationships. **Y.S.** disclosed no relevant relationships. **K.J.G.** disclosed no relevant relationships. **J.S.B.** disclosed no relevant relationships. **K.C.** disclosed no relevant relationships. **G.C.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed money paid to author for consultancy from Regeneron. Other relationships: disclosed no relevant relationships. **C.M.D.** disclosed no relevant relationships.

References

1. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16(4):494–502.
2. Altman R, Asch E, Bloch D, 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. Arthritis Rheum 1986;29(8):1039–1049.
3. Deshpande BR, Katz JN, Solomon DH, et al. Number of Persons With Symptomatic Knee Osteoarthritis in the US: Impact of Race and Ethnicity, Age, Sex, and Obesity. Arthritis Care Res (Hoboken) 2016 Dec;68(12):1743–1750.
4. Weinstein AM, Rome BN, Reichmann WM, et al. Estimating the Burden of Total Knee Replacement in the United States. J Bone Joint Surg Am 2013;95(5):385–392.
5. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage 2007;15(Suppl A):A1–A56.
6. Culvenor AG, Engen CN, Øiestad BE, Engebretsen L, Risberg MA. Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Knee Surg Sports Traumatol Arthrosc 2015;23(12):3532–3539.
7. Lee HC, Lee JS, Lin MCJ, Wu CH, Sun YN. Automatic Assessment of Knee Osteoarthritis Parameters from Two-Dimensional X-ray Image. In: First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06), Beijing, China, August 30–September 1, 2006. Piscataway, NJ: IEEE, 2006; 673–676.
8. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009; 248-255.
9. Antony J, McGuinness K, O'Connor NE, Moran K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, December 4–8, 2016. Piscataway, NJ: IEEE, 2017; 1195–1200.
10. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. Sci Rep 2018;8(1):1727.
11. Lazzarini N, Runhaar J, Bay-Jensen AC, et al. A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. Osteoarthritis Cartilage 2017;25(12):2014–2021.
12. Nevitt M, Felson D, Lester G. The Osteoarthritis Initiative: Protocol for the cohort study. 2006; 1–74. https://oai.epi-ucsf.org/datarelease/docs/StudyDesignProtocol.pdf. Accessed April 5, 2015.
13. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15(12):1833–1840.
14. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. J Orthop Sports Phys Ther 1998;28(2):88–96.
15. Hochberg MC, Favors K, Sorkin JD. Quality of life and radiographic severity of knee osteoarthritis predict total knee arthroplasty: data from the osteoarthritis initiative. Osteoarthritis Cartilage 2013;21(Suppl):S11.
16. Zhang W, McWilliams DF, Ingham SL, et al. Nottingham knee osteoarthritis risk prediction models. Ann Rheum Dis 2011;70(9):1599–1604.
17. Dahaghin S, Bierma-Zeinstra SMA, Reijman M, Pols HAP, Hazes JMW, Koes BW. Does hand osteoarthritis predict future hip or knee osteoarthritis? Arthritis Rheum 2005;52(11):3520–3527.
18. Central Reading of Knee X-rays for Kellgren & Lawrence Grade and Individual Radiographic Features of Tibiofemoral Knee OA. 2016; 1–30. https://oai.epi-ucsf.org/datarelease/SASDocs/kXR_SQ_BU_descrip.pdf. Accessed December 12, 2019.
19. Central assessment of longitudinal knee x-rays for quantitative JSW. 2016; 1–8. https://oai.epi-ucsf.org/datarelease/SASDocs/kXR_QJSW_Duryea_descrip.pdf. Accessed December 10, 2019.
20. Project 15 Test-Retest Reliability of Semi-quantitative Readings from Knee Radiographs. 2016; 1–17. https://oai.epi-ucsf.org/datarelease/SASDocs/kXR_SQ_Rel_BU_descrip.pdf. Accessed December 10, 2019.
21. Project 20 Test-Retest Reliability of Joint Space Width Measurements from Knee Radiographs. 2012; 2–5. http://oai.epi-ucsf.org/datarelease/SASDocs/kXR_QJSW_Rel_Duryea_Descrip.pdf. Accessed May 21, 2020.
22. Caruana R. Multitask Learning. Mach Learn 1997;28(1):41–75.
23. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 27–30, 2016. Piscataway, NJ: IEEE, 2016; 770–778.
24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
25. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6(2):65–70. https://www.jstor.org/stable/4615733.
26. Richardson ML. The Zombie Plot: A Simple Graphic Method for Visualizing the Efficacy of a Diagnostic Test. AJR Am J Roentgenol 2016;207(4):W43–W52.
27. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22–29, 2017. Piscataway, NJ: IEEE, 2017; 618–626.
28. Jamshidi A, Pelletier JP, Martel-Pelletier J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. Nat Rev Rheumatol 2019;15(1):49–60.
29. Bastick AN, Belo JN, Runhaar J, Bierma-Zeinstra SMA. What Are the Prognostic Factors for Radiographic Progression of Knee Osteoarthritis? A Meta-analysis. Clin Orthop Relat Res 2015;473(9):2969–2989.
30. Joseph GB, McCulloch CE, Nevitt MC, et al. Tool for osteoarthritis risk prediction (TOARP) over 8 years using baseline clinical data, X-ray, and MRI: Data from the osteoarthritis initiative. J Magn Reson Imaging 2018;47(6):1517–1526.
31. Kerkhof HJM, Bierma-Zeinstra SMA, Arden NK, et al. Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. Ann Rheum Dis 2014;73(12):2116–2121.
32. Hofman A, Grobbee DE, de Jong PTVM, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol 1991;7(4):403–422.
33. Gossec L, Jordan JM, Mazzuca SA, et al. Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the OARSI-OMERACT task force. Osteoarthritis Cartilage 2008;16(7):742–748.
34. Hunter DJ, Zhang W, Conaghan PG, et al. Systematic review of the concurrent and predictive validity of MRI biomarkers in OA. Osteoarthritis Cartilage 2011;19(5):557–588.
35. Kwoh CK, Vina ER, Cloonan YK, Hannon MJ, Boudreau RM, Ibrahim SA. Determinants of patient preferences for total knee replacement: African-Americans and whites. Arthritis Res Ther 2015;17(1):348.