Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images

Jaynal Abedin[1,*]**, Joseph Antony**[2]**, Kevin McGuinness**[2]**, Kieran Moran**[2,3]**, Noel E O'Connor**[2]**, Dietrich Rebholz-Schuhmann**[1,5]**, and John Newell**[1,4]

[1]Insight Centre for Data Analytics, National University of Ireland Galway, Ireland
[2]Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland
[3]School of Health and Human Performance, Dublin City University, Dublin, Ireland
[4]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Ireland
[5]ZB MED – Information Centre for Life Sciences, University of Cologne, Germany
[*]joystatru@gmail.com, jaynal.abedin@insight-centre.org

ABSTRACT

Knee osteoarthritis (KOA) is a disease that impairs knee function and causes pain. A radiologist reviews knee X-ray images and grades the severity level of the impairments according to the Kellgren and Lawrence grading scheme; a five-point ordinal scale (0–4). In this study, we used Elastic Net (EN) and Random Forests (RF) to build predictive models using patient assessment data (i.e. signs and symptoms of both knees and medication use) and a convolution neural network (CNN) trained using X-ray images only. Linear mixed effect models (LMM) were used to model the within subject correlation between the two knees. The root mean squared error for the CNN, EN, and RF models was 0.77, 0.97 and 0.94 respectively. The LMM shows similar overall prediction accuracy as the EN regression but correctly accounted for the hierarchical structure of the data resulting in more reliable inference. Useful explanatory variables were identified that could be used for patient monitoring before X-ray imaging. Our analyses suggest that the models trained for predicting the KOA severity levels achieve comparable results when modeling X-ray images and patient data. The subjectivity in the KL grade is still a primary concern.

Introduction

Osteoarthritis (OA) is the result – and the observable status – of inflammatory processes in a joint leading to functional and anatomical impairments. The resulting status often shows irreversible damages to the joint cartilage and the surrounding bone structures[1,2]. The knees are the most commonly affected joints in the human body and knee osteoarthritis (KOA) is more prevalent in females aged 60 years or more compared to males of the same age (13% vs 10%)[3]. Severity of KOA amongst females with more than 55 years of age is higher compared to their male counterparts and the severity of KOA is higher compared to other types of OA[4,5]. Approximately one in every six patients consult with a general practitioner in their first year of an OA episode [4,5]. The incidence of KOA has a positive association with age and weight and the prevalence is more common in younger age groups, particularly those who have obesity problems[6].

Swelling, joint pain, and stiffness are the prominent symptoms among others, such as restrictions in movement including walking, stair climbing, and bending[7]. The symptoms worsen over time and elderly patients are affected more frequently than patients in other age groups. The presence of OA in the knee reduces activity in daily life and eventually leads to disability, which can incur high costs related to loss in productivity[8]. It is estimated that functional impairment of the knee and the hip are the eleventh highest disability factors[9] contributing to considerable socio-economic burden with an estimated cost per patient per year of approximately 19,000 Euro[10]. The estimated prevalence of disability due to arthritis is expected to reach 11.6 million individuals by the year 2020[11], which is greater than the estimated risk of disability attributable to cardiovascular diseases or any other medical condition[12]. Total joint replacement surgery is the most favorable option to treat advanced stage OA. However, diagnosing the status of KOA at an early stage and providing behavioral interventions could be beneficial for prolonging a healthy life for a patient[13].

In a review of possible risk factors of KOA, Heidari[7] concluded that age, obesity, gender (i.e., female), repetitive knee trauma and kneeling are the most common risk factors for KOA. The common symptoms include pain, functional impairment, swelling and stiffness. The severity of KOA and the pain status is measured based on the Kellgren and Lawrence (KL) scale of 0 to 4 by visual inspection of the knee X-ray images[14].

Considering the impact of KOA on disability and the subsequent unavoidable economic burden, there is a need to quantify the severity of KOA during the early stages of development. KOA severity level helps in determining appropriate treatment decisions and for the monitoring of disease progression[15]. The classical way of quantifying KOA severity is by inspection of X-ray images of the knee by a radiologist who then grades the images according to the KL scale (from 0 for "normal" up to 4 for "severe" stage)[14]. This approach suffers from high levels of subjectivity as there is no gold standard grading system: the semi-quantitative nature of the KL grading scale creates ambiguity, thus giving rise to disagreements between raters (for details please refer to[14, 16, 17]).

To reduce the influence of subjectivity in quantifying KOA severity from X-ray images, computer-aided diagnosis has been very helpful[18]. To date the sample size of available images has been the main limiting factor to train an efficient model[19–22]. The Osteoarthritis Initiative (OAI)[23] and the Multi-centre Osteoarthritis Study (MOST)[24] mitigated this small sample size limitation by making thousands of patients data and X-ray images available. Recently, several researchers have used these resources to develop an automatic approach for quantifying KOA severity by analyzing X-ray images[21, 25–28]. Although there have been multiple attempts to quantify KOA severity based on an automated analysis of X-ray images, so far there has been no attempt to build a predictive model on a patient's assessment data such as signs, symptoms, medication and other characteristics about a patient (later on referred as patient's questionnaire data) and to compare this approach against the X-ray based prediction. Developing predictive models using patient data other than X-ray images offers additional advantages such as identifying those variables that contribute strongest towards predicting the severity of KOA. A good predictive model based on patient's questionnaire data could reduce treatment costs and could also contribute to a prolonged healthy life of a patient due to early behavioral intervention.

The Osteoarthritis Initiative (OAI) is a multi-center longitudinal study for men and women sponsored by the National Institute of Health (NIH) to better understand KOA. Data collected through the OAI can provide useful information about the marginal distributions of relevant patient characteristics, their demographics, signs & symptoms and medication history. To date, there are more than 200 scientific publications that have used data collected through the OAI including several attempts to automate the KL grade quantification using X-ray images. But to date no study (or publication) has tried to predict KOA severity based on patient questionnaire data. Our primary goal is to compare the prediction accuracy of a statistical model based on patient questionnaire data to the prediction accuracy based on X-ray image based modeling to predict KOA severity score. In this paper we present several statistical approaches to predict the severity of KOA using patient questionnaire data. Furthermore, we use a convolution neural network (CNN) model to predict the same outcome using corresponding X-ray images for the same patients. The performance of both the approaches has been compared using the calculated root mean squared error on a validation set. As a secondary goal, we identified key variables with the strongest predictive ability, which may be useful to monitor a patients over time and design early interventions for prolonging healthy life in patients of concern.

Results

Exploratory analysis

The OAI dataset contains data for $4,796$ individuals. After initial pre-processing, $2,951$ patients with sufficient data on potential candidate explanatory variables were selected, representing 62% of the original patients. The remaining 38% of individuals did not have enough data for the potential explanatory variables and were not included in the analysis. The list of candidate variables, their labels and type (binary, numeric and categorical) has been provided in a supplementary table (Supp table 1).

To train and validate the predictive models we used a training and validation data split as shown in Table 1 (roughly a 70% - 30% split). To make valid comparisons, we used the same validation set in the models developed on patient's questionnaire data and the model developed using X-ray images[26, 27]. The validation set contained data for both knees for 846 patients, i.e. $1,692$ data points for a knee in total. The training set for the predictive models included data from $2,105$ patients, i.e. $4,210$ knees.

Severity level	Training: Freq (%)	Validation: Freq (%)	Total: Freq (%)
Level 0	1818 (43.2)	685 (40.5)	2503 (42.4)
Level 1	728 (17.3)	312 (18.4)	1040 (17.6)
Level 2	1045 (24.8)	416 (24.6)	1461 (24.8)
Level 3	503 (12.5)	237 (14.0)	740 (12.5)
Level 4	115 (2.7)	42 (2.5)	157 (2.7)

Table 1. Distribution of KOA severity between training and validation data

The validation data that contains 30% of the original patients data are the same patients information that has been used in X-ray image based modeling. To make our results comparable with X-ray based prediction we used same validation patients

information, although we performed cross validation to check sensitivity of our entire analysis. The cross validation result is consistent with our original 70-30 split.

Relevant summary statistics for patient characteristics are given in Table 2 for the entire dataset and for the training and the validation subsets. Good balance is evident when comparing the mean and variability for each patient characteristic across the training and validation data and it is plausible that they can be considered as representative samples taken from the same overall population. Maintaining a similar distribution of patient characteristics between the training and validation data is paramount for making reliable inference. We dropped occupation from the patient characteristics table and from subsequent analysis as this variable has more than 30% missing data.

Characteristics	Training: Mean (SD)	Validation: Mean (SD)	Total: Mean (SD)
Age	60.3 (9.2)	61.1 (8.9)	60.5 (9.1)
Female (Freq. %)	1177 (56.0)	454 (53.7)	1631 (55.3)
Height (mm)	1685.2 (93.2)	1687.3 (92.6)	1685.8 (93.0)
Weight (kg)	80.7 (16.3)	80.5 (15.7)	80.6 (16.1)
BMI (kg/m^2)	28.3 (4.8)	28.2 (4.6)	28.3 (4.7)
Systolic	123.3 (15.9)	123.7 (16.7)	123.3 (16.1)
Diastolic	75.5 (9.8)	75.4 (9.6)	75.5 (9.8)

Table 2. Summary statistics of patient characteristics between complete, training and validation data

The box-plot (Figure 1) displays the distribution of several patient characteristics. We have introduced minor displacement quantity (jitters) along the horizontal axis and alpha-blending to make the display of the distribution of the points clearer. The level of KOA severity appears to be higher among elderly people. Height, weight and BMI show similar patterns but in contrast the distribution of blood pressure measurements does not indicate any strong obvious pattern with severity score. (Figure 1).
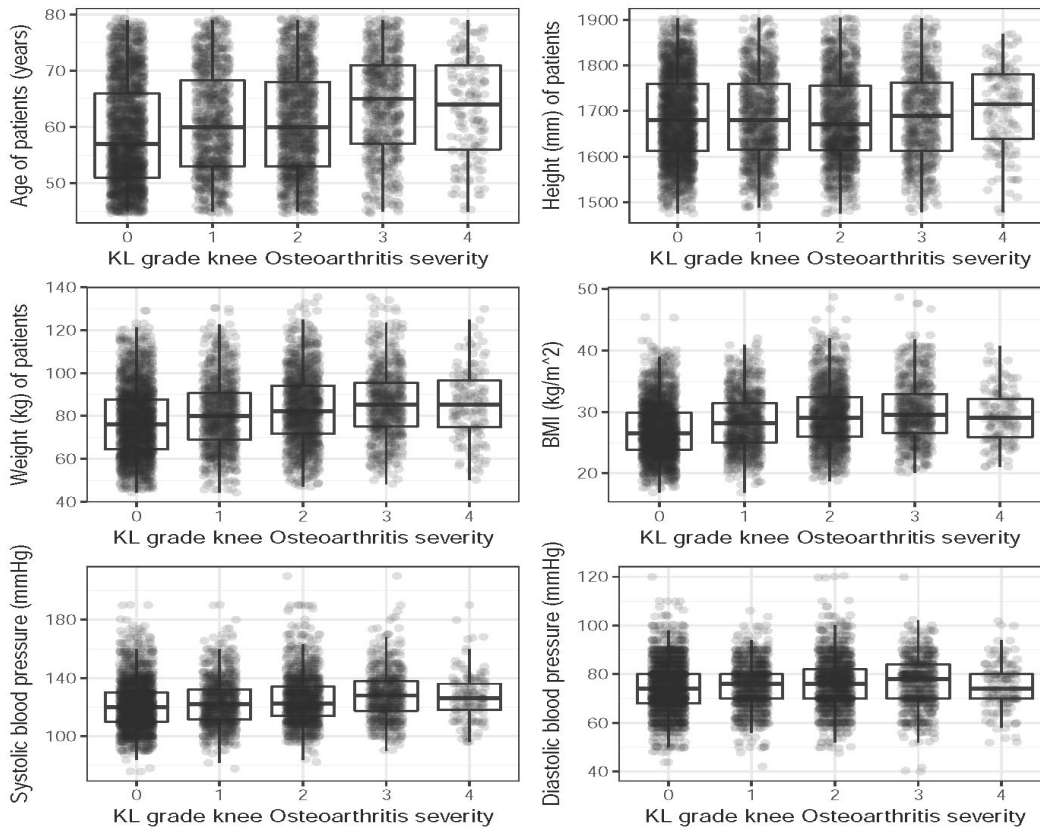

Figure 1. Boxplots of patient characteristics by Knee Osteoarthritis Severity

The data contains a mixture of continuous, binary, and categorical predictors. To observe the relationship among such a collection of predictor variables we calculated Pearson correlation between continuous predictors, polyserial correlations

between continuous and categorical predictors and polychoric correlation between categorical predictors[29]. The correlation matrix (Figure 2) shows the relationship among the predictor variables of interest where a higher color intensity indicates a stronger correlation between variables. The blue color indicates a positive correlation and red color indicates a negative correlation. We can observe that the predictor variables are positively correlated with each others to a moderate degree. Patients sex, height and weight shows weak negative correlation with other variables but only sex and height show a strong negative correlation. The upper block represents correlation among signs and symptoms in the left knee whereas the lower right block represents correlation among signs and symptoms of the right knee. The lower left block is the correlation between signs and symptoms of left to right knee. Other than the three blocks of correlation there are some variables that represent neither of the knees; rather, those variables represents medication history and other characteristics (Figure 2). What is clear from Figure 2 is the large number of candidate predictors and the presence of multicollinearity amongst predictors which will have to be accounted for accordingly in any subsequent model.
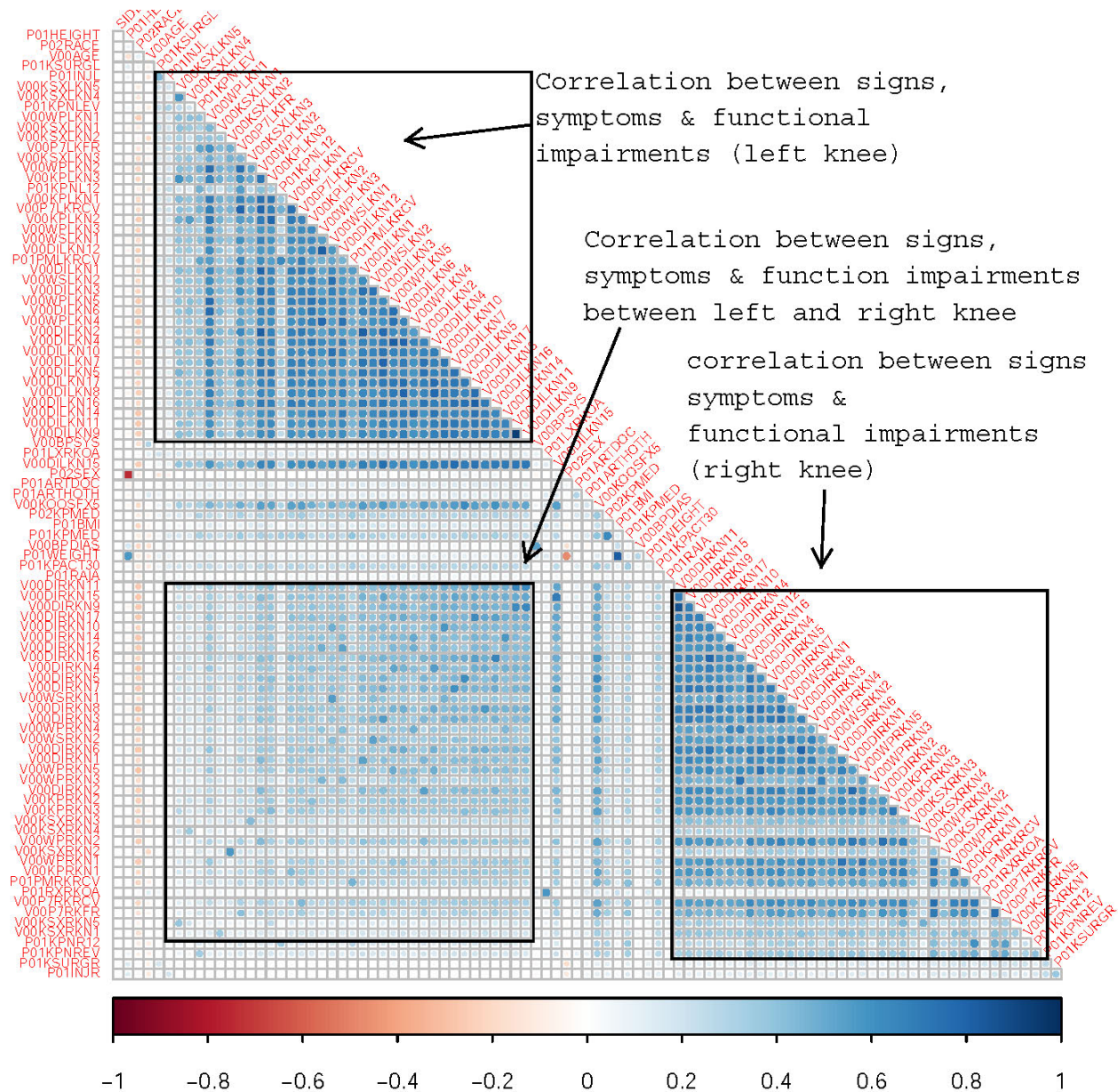

Figure 2. Correlation among predictors (dark color indicates stronger correlation)

Among the five levels of KOA severity there were very few patients from severity level 4 (KL grade) in comparison to other

categories. Overall 42% of patients were from severity level 0 indicating the normal knee followed by mild severity level 2 with 24% and doubtful severity level 1 with 17%. The distribution of severity level frequencies across the training and the validation data is well balanced indicating that it is plausible that the training and validation data came from same underlying population (Table 1.)

Model building, evaluation, and comparison

Initially an Elastic Net regression[30], a weighted combination of LASSO and Ridge regression, was fitted. An Elastic Net regression model can be used to select variables with high predictive power. The weighting is controlled by the mixing parameter α that controls the amount of mixing between LASSO and Ridge penalties, whereas the parameter λ controls the amount of shrinking in the regression coefficients. To estimate a suitable value for the shrinkage parameter λ we performed repeated cross validation using a fixed $\alpha = 0.5$, which corresponds to the minimum cross-validation RMSE. Using this value of α the value of λ that also minimizes the RMSE (Figure 3) was selected. The contribution (i.e. direction and magnitude) of each predictor variable has been extracted from the corresponding estimated regression coefficients (Figure 4).
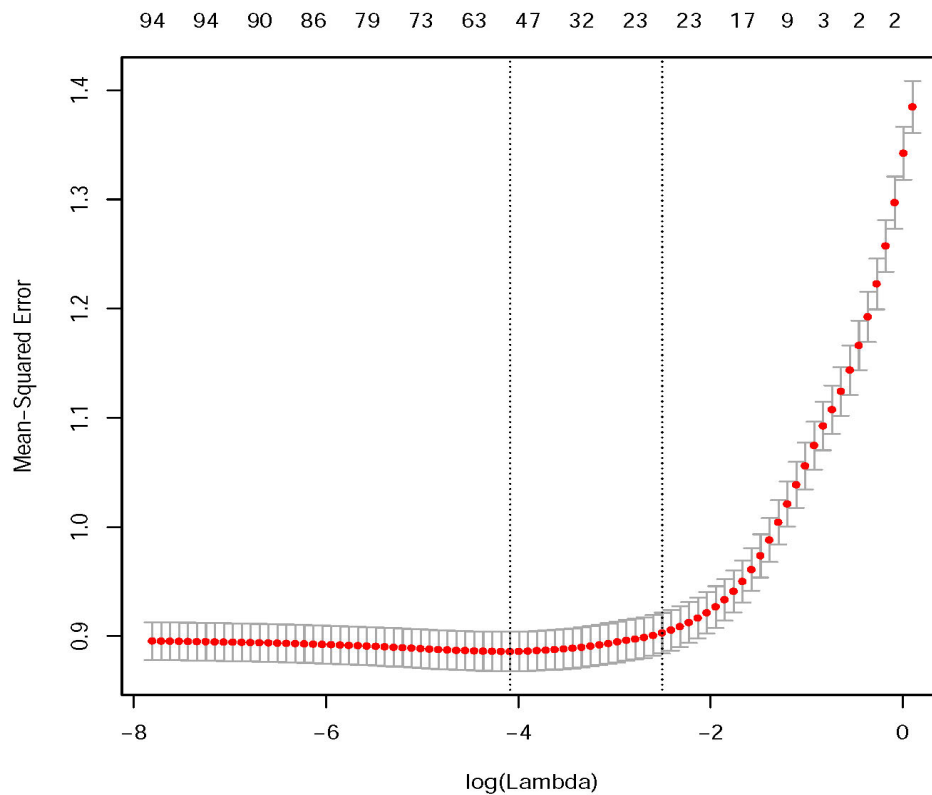

Figure 3. Estimation of hyper-parameter λ. The Y-axis represents RMSE for different values of λ whereas the upper horizontal axis represents the number of predictors. The RMSE increases as the number of predictors decrease but stabilizes after a certain number of predictors are added. The red points represents RMSE and the gray line segments represents a 95% confidence interval corresponding to each RMSE. The optimal value of λ is the minimum value corresponding to a steady-state RMSE.

A Random Forest[31] regression model was then fitted using differing numbers of trees where the RMSE was calculated for each scenario. Based on these evaluations, we found that using 100 trees produced the lowest RMSE in the validation set. We also identified those predictors with highest variable importance in terms of improved predictive ability of the final forest.

The overall RMSE for the Elastic Net regression model is 0.97 and the RMSE for the random forest model is 0.94. Both models give higher accuracy for the prediction of the severity levels 1 and 2 in contrast to the other categorical levels. The RMSE from the X-ray image based CNN model is 0.77, which is slightly lower than the RMSE from the Elastic Net regression and the Random Forest model. The advantage of using Elastic Net regression over a Random Forest regression model and

X-ray image based CNN model is that we can easily identify the variables that have high predictive power and also the direction of the contribution of each variables by looking at the magnitude and sign of their regression coefficients (Figure 4).
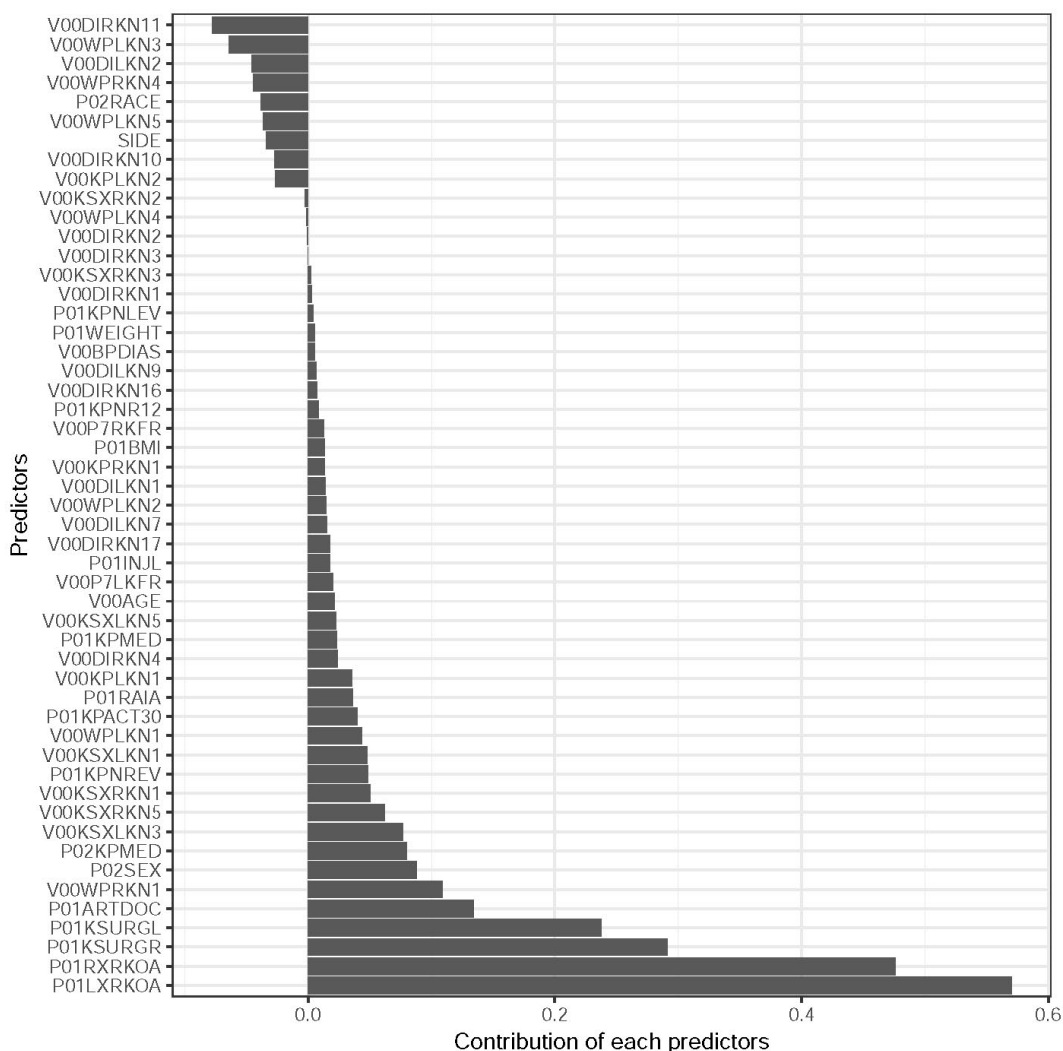
Figure 4. Contribution of each variable on KOA severity score prediction by Elastic Net Regression

The Elastic Net regression model produced higher prediction accuracy for severity level 1 and 2, in comparison to other levels. A similar result is noted in the predictions by the Random Forest model. The overall RMSE of Elastic Net regression and Random Forest regression models are 0.974 and 0.943. The overall accuracy of the CNN model is higher than Elastic Net and Random Forest regression. The performance of each of the three models show their lowest outcome for the KOA severity level 4 as there is less data available in that category. Relatively higher accuracy in predicting KOA severity using an X-ray based CNN model has been observed however, the margin of difference between the RMSE of the predictions from the X-ray image based CNN model in comparison to the predictions from the patient's questionnaire data models is considerably small. Table 3 shows the RMSE for the models trained with patient data (Elastic Net and Random Forest)and the model trained with X-ray images (CNN regression).

Both the Elastic Net and Random Forest models allow us to variable importance of the individual predictors on overall predictive ability. There are some variables commonly identified by both these models with higher contribution towards the final predictions. However, the variables identified by the Elastic Net have more interpretable properties than the variables selected by the Random Forest model. The sign of the regression coefficients in an Elastic Net regression allows us to understand the direction of the contribution; whether it increases the severity score or reduces it. A negative sign indicates a reduction in the overall severity score for increasing values of the predictor, whereas a positive sign indicates an increase in the severity score for increasing values of the predictor. The direction of the contribution by predictors selected by the random forest model

Severity Level	Elastic Net Regression	Linear Mixed Model (LMM)	Random Forest Regression	CNN Regression
Level 0	0.917	0.920	0.909	0.816
Level 1	0.563	0.591	0.511	0.485
Level 2	0.881	0.895	0.853	0.840
Level 3	1.320	1.320	1.270	0.795
Level 4	2.140	2.10	2.02	0.846
Overall	0.973	0.978	0.943	0.770

Table 3. Estimated RMSE from different models for each level of KOA severity level

is unclear as it gives similar importance to both directions. Figure 4 shows the sign and magnitude of the contribution for each of the selected variables. The positive sign indicates an increase in the severity score whereas a negative sign indicates a decrease in the severity score. The identified variables could be a proxy indicators of patient knee's anatomical structure which ultimately indicates the level of severity.

Since the data have a hierarchical structure (i.e. knee nested within patient) the number of replicates at the individual level is more appropriately modeled using a mixed effects model with a random effect to capture the correlation between knees within an individual. To explore the random effect of patient level information, a linear mixed effect model was fitted using the predictor variables initially selected from the Elastic Net regression. There is clear evidence for the need of random effects due to the study design. In addition a small p-value (less than 0.001) was evident for the test for the need of the random effect term due to subject level within knee correlation in the model. The intra-class correlation coefficient is 0.265 which indicates that the proportion of the variance explained by the random effect component (patient level information) in the population 26.5%. The overall RMSE for the linear mixed effects model is 0.978, which is almost the same as the RMSE from the Elastic Net regression. However, the predicted severity levels, and more importantly the corresponding uncertainty, is correctly adjusted to account for the within patient correlation.

Discussion

Judging the impairment for patients with KOA requires a thorough understanding of the disease condition. Expert radiologists or clinicians assess the functional knee impairments and the KOA severity level from the X-ray images. Ideally, the image analysis should give an objective measure of the impairments; however, in reality not all functional impairments show up in anatomical transformations of the knee, and the patho-physiological evaluation relies on the subjective perception of the patient and the physician jointly.

Our primary goal was to explore whether the prediction accuracy of a statistical model based on patient's questionnaire data is comparable to the prediction accuracy based on X-ray image based modeling to predict KOA severity. We have demonstrated that statistical models, using patients' questionnaire data, could predict KOA severity level with a good level of accuracy (RMSE: 0.974 & 0.943). The prediction performance of the statistical models presented in this paper are comparable to models using X-ray image data based on model performance as assessed by RMSE measures[26–28]. In particular we have demonstrated that functional impairment at severity levels 1 and 2 can be predicted by our statistical models (Elastic Net & Random Forest and LMM) trained from the patients' assessment data to a level of accuracy similar to the accuracy achieved on the basis of CNN model trained on X-ray images. There are very subtle structural variations in the knee joints (minimal joint space narrowing (JSN) and osteophytes formation) belonging to grade 0 and grade 1, and these are not fully reflected in the KL grades. Also, there are relatively large overlaps in the JSN measurements for KL grades 0 and 1 compared to the other grades[32]. These factors make them challenging to distinguish by inspecting the X-ray images. Also, patients share almost similar distribution on their characteristics, signs, symptoms and functional impairments. Due to very subtle differences of the predictors between KOA levels the prediction accuracy gets affected.

We were able to identify the key variables that contributed most to the predictive ability in our models. These identified variables can be monitored over time to assess the progression of KOA severity. The strong indicator variables are reporting on knee baseline radiographic OA status for the right or left knee (P01LXRKOA, P01RXRKOA) and on treatments such as surgery on the right or left knee (P01KSURGR, P01KSURGL) as well as other reasons to see the doctor (P01ARTDOC). Patient's sex also plays important role in predicting KOA severity. The next indicator variables cover medication (P02KPMED) and functional impairments, pain or other symptoms to the right or left knee (V00WPRKN1, V00KSXLKN3, V00KSXRKN5, V00KSXRKN1, V00KSXLKN1, V00WPLKN1, P01KPNREV, P01KPACT30). A final parameter notes whether a doctor "ever said you have rheumatoid arthritis or other inflammatory arthritis" (P01RAIA). The predictors variable that we found as important predictor of KOA severity were also reported important risk factor in previous studies[7,33].

Importantly, an early behavioral intervention could be developed based on the identified variables to prolong the healthy life

of a patient. By observing the identified variables that have higher predictive ability to predict KOA severity, we can identify the subjects who are currently taking medication for pain relief and facing functional difficulty in their daily life. Variables representing limited knee functions in particular are the potential indicators for quantifying KOA severity that could lead to developing targeted interventions for further treatment and medications.

When making predictions the LMM is favored as it is the only approach that correctly adjusts for the hierarchical structure present in the data. It is interesting that the severity levels 1 and 2 can be predicted with good accuracy in all the four models (EN, RF, LMM, and CNN), while the other levels of severity are more challenging to predict. For higher severity levels, i.e. levels 3 and 4, this could be due to the lack of patient data, i.e. the sample sizes at these levels are smaller than for levels 1 and 2 (Figure 1 & table 2).

As a conclusion based on the results in this paper, we can say that the patients' questionnaire data can predict KOA severity level with good accuracy and it is comparable with the prediction based on X-ray images. Patient's assessment data also enables us to identify some of the key variables that can be used to design early interventions and monitor the patients over the treatment period. The accuracy of the model developed using patient's assessment data is almost comparable to the CNN model. Moreover, the statistical models have an edge over the CNN model by identifying key variables that helps the physicians to design interventions and helps the patients for further treatment.

There is at least one potential limitation in developing statistical models to predict KOA severity, that is the KL grade score itself is not a gold standard and suffers from subjectivity. The KL grade is dependent on the perception of the radiologist who is inspecting the X-ray images. In the model building process, we are effectively using a quasi-gold standard outcome. Considering this potential limitation, one way to improve the prediction accuracy could be to build a model of the X-ray image data in combination with the patients' assessment data. The prediction of KOA severity based on patients data shows comparable accuracy, it would be interesting to see the performance of prediction based on a statistical model combining both patient's questionnaire data and with X-ray images.

Methods

Data

The data used in this study were obtained from the Osteoarthritis Initiative (OAI), which is available for public access at http://www.oai.ucsf.edu/. The specific dataset used is labeled 0.2.2. This is the data from the multi-center longitudinal and prospective observational study of KOA. We have used the baseline dataset for this work. The description of each variable used in our analysis has been given in the supplementary table 1.

Data pre-processing and descriptive statistics

The baseline dataset contains a large number of variables related to patients' characteristics, their vital signs, symptoms of KOA, medication history, and functional impairment. At the early stage of the analysis, we manually inspected each of the variables and selected a subset of candidate variables that were clinically relevant and previously reported risk factor for KOA[7,33]. We inspected the completeness of the data in terms of missing values. We calculated the amount of missing values (in percent) for each variable. The variables that had at least 85% non-missing values were kept for further analysis. If it can be assumed that missing data are missing at random (i.e. missingness is explained by the covariates available) a multiple imputation step is unnecessary if a linear mixed model is used as the likelihood is correctly specified under this assumption. Moreover, we excluded categorical variables with very low discriminatory power, for example, the variables with very low frequency in one of the categories compared to the rest within the same variable. The reduced set of candidate variables was used for further processing and analysis. The dataset was split into two parts, training and validation sets, by taking a random sample of 70% of the data for training and the remaining 30% for validation. To make valid comparisons, we used the same validation set in the models developed on patient's questionnaire data and the model developed using X-ray images[26,27]. The data pre-processing steps are summarized in (Figure 5).

To summarize and explore the explanatory variables, we calculated descriptive statistics: mean and standard deviation for numeric variables, frequency and percentage for categorical variables. The relationship among predictors was also explored by calculating Pearson correlation between numeric variables, polyserial correlations between numeric and categorical variables and polychoric correlation between categorical variables[29].

The KL grade score was recorded on an ordinal scale from 0: normal to 4: severe. To model an ordinal outcome, ordinal logistic regression[34,35] is the typical approach used. We fitted ordinal logistic regression models, but the prediction performance was poor. In this paper we have treated the severity score as a continuous response to investigate if this would improve predictive ability. Moreover, the data are hierarchical in structure; for each patient we have data for both knees. To capture this structure appropriately we have used a linear mixed effect model incorporating a random effect at the subject level[36].
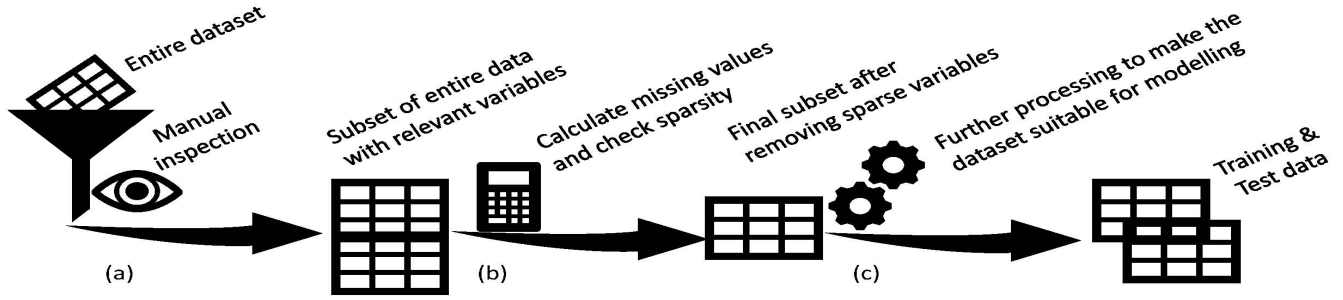
Figure 5. Data pre-processing work flow: (a) Inspecting entire dataset manually to get subset of relevant candidate variables, (b) calculate percentage of missing values for each variables and also inspect sparsity of the categorical variables. Drop a variable that has more than 15% missing values or very low e.g. less than 5% into one category in a binary variable, (c) creating dummy variables from multi-category variables and then split the dataset into training and test data for predictive model building.

Elastic Net regression

Elastic Net regression is a combination of ridge regression and LASSO, and this model is appropriate in the presence of correlated predictors[30]. We denote the outcome variable: KL grade score by Y (considered as a continuous variable) and all predictors by $X_1, X_2, ..., X_p$. The Elastic Net regression linearly combines L_1 and L_2 penalties as follows:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda[(1-\alpha)\sum_{j=1}^{p}\theta_j^2 + \alpha\sum_{j=1}^{p}|\theta_j|] \tag{1}$$

The L_1 penalty is defined as the sum of absolute value of the regression coefficients and the $L_1 = \sum_{j=1}^{p}|\theta_j|$ and L_2 penalty is defined as the sum of squared values of regression coefficients: $L_2 = \sum_{j=1}^{p}\theta_j^2$. The amount of mixing between two penalty term is controlled by a mixing parameter α. If the value of $\alpha = 0$ then it leads to a ridge regression whereas a value of $\alpha = 1$ leads to LASSO regression. The hyper-parameter λ controls the amount of shrinkage of regression coefficients for various values of α. A higher value of λ leads to shrink the regression coefficients towards zero and a very small value of λ has little effect on the regression coefficients. Using both L_1 and L_2 penalty enables us to select appropriate variables that have higher predictive power by shrinking some of the regression coefficient to zero using an appropriate value of hyper-parameter λ. To estimate the most suitable value for the shrinkage parameter λ we performed repeated cross validations with fixed values of the mixing parameter $\alpha = 0.5$ and choose the value of λ that minimizes the root mean squared error (RMSE). Figure 3, shows the cross-validation results while selecting λ.

Random Forest

Random Forests (RF) are an ensemble method that combines the predictive ability of multiple tree based models. The RF model is an extension of the original work of Tin Kam Ho[37] who developed the algorithm for random decision forests. Leo Breiman[31] used the idea of bagging (bootstrap aggregating) and random variable selection. The principle of random forest is to combine multiple tree based models to form a single model that can achieve better accuracy compared to its individual counterparts. This method takes a random sample with replacement from the original data, then builds a decision tree model based on a random selection of variables at each branch in the tree. This process is repeated for multiple trees and stores the prediction from each tree. The predicted value is then the mode (for a categorical response) for the mean (for a continuous response) across the forest. The random forest model is popular because it can reduce the variance of single tree models and also overcomes the problem of correlated predictors as it takes only a subset of candidate predictor variables in each of the individual trees.

Linear Mixed Effect Model

There is a clear hierarchical structure in the dataset as we have patient level data along with knee level data. A linear mixed effect model (LMM)[36] is an extension of a linear model that accounts for the hierarchical structure in data. The primary benefit of using a LMM in this paper is that the uncertainty in knee level prediction is now correctly adjusted for through the introduction of a suitable random effect. This approach will account for measurements on both knees collected for each subject correctly. The Intra-class correlation has been reported that indicates how much variation in the dependent variable is due to random effect component in the LMM model. A random effect model can be formulated as:

$$y_{ij} = x_{ij}^t\beta + u_{ij}^t\gamma_i + \varepsilon_{ij}; i = 1, 2, \cdots, m; j = 1, 2, \cdots, n_i \tag{2}$$

Here y_{ij} is the KL grade of i-th knee of j-th patient, x_{ij} the covariate of vector of j-th member of cluster *i* for fixed effects; u_{ij} covariate vector of j-th member of cluster *i* for random effects; γ_i is the random effect parameter, *m* is the the number of cluster (in our case $m = 2$ representing left and right knee), β is the regression coefficient of the fixed effect covariates.

Convolution Neural Network

In the machine learning based approach to automatically assess the KOA severity, the first step is to localize the region of interest (ROI), that is to detect and extract the knee joint regions from the X-ray images, and the next step is to classify the localized knee joints based on KL grades. In our previous study[27], we introduced a fully convolutional neural network (FCN) to automatically detect and extract the knee joints, and trained CNNs from scratch to predict the KOA in both discrete and continuous scales using classification and regression respectively[26,27]. We used the baseline X-ray images from the OAI dataset to train the CNN model. After testing different configurations, the network in Table 4 was found to be the best for classifying knee images. The network contains five layers of learned weights: four convolutional layers and a fully connected layer. Each convolutional layer in the network is followed by batch normalization and a ReLU activation layer. After each convolutional stage there is a max pooling layer. The final pooling layer (maxPool4) is followed by a fully connected layer (fc5) with output shape of 1024 and a softmax dense (fc6) layer with output shape of 5 representing five level of KOA severity. To avoid overfitting, a drop out layer with a drop out ratio of 0.25 is included after the last convolutional (conv4) layer and a drop out layer with a drop out ratio of 0.5 after the fully connected layer (fc5). Also, a L2-norm weight regularization penalty of 0.01 is applied in the last two convolutional layers (conv3 and conv4) and the fully connected layer (fc5). Applying a regularization penalty to other layers increases the training time whilst not introducing significant variation in the learning curves. The network is trained to minimize categorical cross-entropy loss using the Adam optimizer with default parameters: initial learning rate $(\alpha) = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e^{-8}$. The inputs to the network are knee images of size 200×300. This size is selected to approximately preserve the aspect ratio based on the mean aspect ratio (1.6) of all the extracted knee joints.

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	$32 \times 100 \times 150$
maxPool1	–	3×3	2	$32 \times 49 \times 74$
conv2	64	5×5	1	$64 \times 49 \times 74$
maxPool2	–	3×3	2	$64 \times 24 \times 36$
conv3	96	3×3	1	$96 \times 24 \times 36$
maxPool3	–	3×3	2	$96 \times 11 \times 17$
conv4	128	3×3	1	$128 \times 11 \times 17$
maxPool4	–	3×3	2	$128 \times 5 \times 8$

Table 4. Best performing CNN for classifying the knee images

After training, this network achieves an overall root mean-squared error 0.771 on the test data. Figure 6 shows the learning curves whilst training this network. The learning curves show proper convergence of the training and validation losses with consistent increase in the training and validation accuracy until they reach constant values.
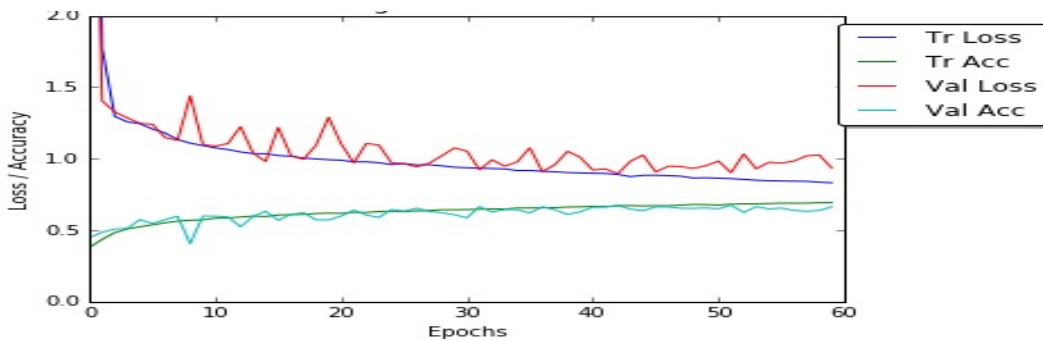
Figure 6. Learning curves: training and validation losses, and accuracy of the fully trained CNN.

References

1. Arden, N. *et al.* *Atlas of osteoarthritis* (Springer, 2014).

2. Eyre, D. R. Collagens and cartilage matrix homeostasis. *Clin. orthopaedics related research* **427**, S118–S122 (2004).

3. Murphy, S. L., Lyden, A. K., Phillips, K., Clauw, D. J. & Williams, D. A. Subgroups of older adults with osteoarthritis based upon differing comorbid symptom presentations and potential underlying pain mechanisms. *Arthritis research & therapy* **13**, R135 (2011).

4. Zhang, Y. & Jordan, J. M. Epidemiology of osteoarthritis. *Clin. geriatric medicine* **26**, 355–369 (2010).

5. Peat, G., McCarney, R. & Croft, P. Knee pain and osteoarthritis in older adults: a review of community burden and current use of primary health care. *Annals rheumatic diseases* **60**, 91–97 (2001).

6. Bliddal, H. & Christensen, R. The treatment and prevention of knee osteoarthritis: a tool for clinical decision-making. *Expert. opinion on pharmacotherapy* **10**, 1793–1804 (2009).

7. Heidari, B. Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part i. *Casp. journal internal medicine* **2**, 205 (2011).

8. Altman, R. D. Early management of osteoarthritis. *The Am. journal managed care* **16**, S41–7 (2010).

9. Cross, M. *et al.* The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Annals rheumatic diseases* annrheumdis–2013 (2014).

10. Puig-Junoy, J. & Zamora, A. R. Socio-economic costs of osteoarthritis: a systematic review of cost-of-illness studies. In *Seminars in arthritis and rheumatism*, vol. 44-5, 531–541 (Elsevier, 2015).

11. for Disease Control, C., (CDC, P. *et al.* Arthritis prevalence and activity limitations–united states, 1990. *MMWR. Morb. mortality weekly report* **43**, 433 (1994).

12. Guccione, A. A. *et al.* The effects of specific medical conditions on the functional limitations of elders in the framingham study. *Am. journal public health* **84**, 351–358 (1994).

13. Karsdal, M. *et al.* Disease-modifying treatments for osteoarthritis (dmoads) of the knee and hip: lessons learned from failures and opportunities for the future. *Osteoarthr. cartilage* **24**, 2013–2021 (2016).

14. Kellegren, J. & Lawrence, J. Radiological assessment of osteoarthritis. *Ann Rheum Dis* **16**, 494–501 (1957).

15. Braun, H. J. & Gold, G. E. Diagnosis of osteoarthritis: imaging. *Bone* **51**, 278–288 (2012).

16. Gossec, L. *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. *Osteoarthr. cartilage* **16**, 742–748 (2008).

17. Sheehy, L. *et al.* Validity and sensitivity to change of three scales for the radiographic assessment of knee osteoarthritis using images from the multicenter osteoarthritis study (most). *Osteoarthr. cartilage* **23**, 1491–1498 (2015).

18. Dacree, J. & Huskisson, E. The automatic assessment of knee radiographs in osteoarthritis using digital image analysis. *Rheumatol.* **28**, 506–510 (1989).

19. Shamir, L., Felson, D. T., Ferrucci, L. & Goldberg, I. G. Assessment of osteoarthritis initiative–kellgren and lawrence scoring projects quality using computer analysis. *J. Musculoskelet. Res.* **13**, 197–201 (2010).

20. Woloszynski, T., Podsiadlo, P., Stachowiak, G. & Kurzynski, M. A dissimilarity-based multiple classifier system for trabecular bone texture in detection and prediction of progression of knee osteoarthritis. *Proc. Inst. Mech. Eng. Part H: J. Eng. Medicine* **226**, 887–894 (2012).

21. Shamir, L. *et al.* Early detection of radiographic knee osteoarthritis using computer-aided analysis. *Osteoarthr. Cartil.* **17**, 1307–1312 (2009).

22. Thomson, J., O'Neill, T., Felson, D. & Cootes, T. Automated shape and texture analysis for detection of osteoarthritis from radiographs of the knee. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 127–134 (Springer, 2015).

23. Eckstein, F., Mosher, T. & Hunter, D. Imaging of knee osteoarthritis: data beyond the beauty. *Curr. opinion rheumatology* **19**, 435–443 (2007).

24. Segal, N. A. *et al.* The multicenter osteoarthritis study: opportunities for rehabilitation research. *PM&R* **5**, 647–654 (2013).

25. Oka, H. *et al.* Fully automatic quantification of knee osteoarthritis severity on plain radiographs. *Osteoarthr. Cartil.* **16**, 1300–1306 (2008).

26. Antony, J., McGuinness, K., O'Connor, N. E. & Moran, K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, 1195–1200 (IEEE, 2016).

27. Antony, J., McGuinness, K., Moran, K. & O'Connor, N. E. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, 376–390 (Springer, 2017).

28. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P. & Saarakkala, S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Sci. reports* **8**, 1727 (2018).

29. Kolenikov, S. & Ángeles, G. The use of discrete data in principal component analysis with applications to socio-economic indices. cpc. Tech. Rep., MEASURE Working Paper No. WP-04-85 (2004).

30. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **67**, 301–320 (2005).

31. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

32. Hart, D. & Spector, T. Kellgren & lawrence grade 1 osteophytes in the knee—doubtful or definite? *Osteoarthr. cartilage* **11**, 149–150 (2003).

33. Hunter, D. J., McDougall, J. J. & Keefe, F. J. The symptoms of osteoarthritis and the genesis of pain. *Rheum. Dis. Clin. North Am.* **34**, 623–643 (2008).

34. McCullagh, P. Regression models for ordinal data. *J. Royal Stat. Soc. Ser. B (Methodological)* **42**, 109–127 (1980).

35. Anderson, J. A. Regression and ordered categorical variables. *J. Royal Stat. Soc. Ser. B (Methodological)* **46**, 1–22 (1984).

36. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biom.* 963–974 (1982).

37. Ho, T. K. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, vol. 1, 278–282 (IEEE, 1995).

Acknowledgements

Author contributions statement

Jaynal Abedin [JA] has done the overall data management, statistical analysis & manuscript writing and coordinated with all the co-authors. Joseph Antony [JA] acquired the OAI dataset, trained the CNN model with X-ray images and contributed in writing the manuscript. Kevin McGuinness [KMG], Kieran Moran [KM] & Noel E O'Connor [NO] critically reviewed and provided feedback to revise the manuscript, and supervised this work. Dietrich Rebholz-Schuhmann [DRS] reviewed initial candidate variables for clinical relevance and reviewed the manuscript to provide critical inputs to improve the quality. John Newell [JN] oversaw the statistical analysis and critically reviewed and suggested necessary improvements from the statistical point of view, and he also supervised the first author throughout this work.

Additional information

Competing Interests: The authors declare that they have no competing interests either financially and non-financially.