Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data

Widera, Paweł; Welsing, Paco M.J.; Ladel, Christoph; Loughlin, John; Lafeber, Floris P.F.J.; Petit Dop, Florence; Larkin, Jonathan; Weinans, Harrie; Mobasheri, Ali; Bacardit, Jaume

DOI
10.1038/s41598-020-64643-8

Publication date
2020

Document Version
Final published version

Published in
Scientific Reports

SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data

Paweł Widera[1], Paco M. J. Welsing[2], Christoph Ladel[3], John Loughlin[4], Floris P. F. J. Lafeber[2], Florence Petit Dop[5], Jonathan Larkin[6], Harrie Weinans[7,8], Ali Mobasheri[9,10,11] & Jaume Bacardit[1]✉

Conventional inclusion criteria used in osteoarthritis clinical trials are not very effective in selecting patients who would benefit from a therapy being tested. Typically majority of selected patients show no or limited disease progression during a trial period. As a consequence, the effect of the tested treatment cannot be observed, and the efforts and resources invested in running the trial are not rewarded. This could be avoided, if selection criteria were more predictive of the future disease progression. In this article, we formulated the patient selection problem as a multi-class classification task, with classes based on clinically relevant measures of progression (over a time scale typical for clinical trials). Using data from two long-term knee osteoarthritis studies OAI and CHECK, we tested multiple algorithms and learning process configurations (including multi-classifier approaches, cost-sensitive learning, and feature selection), to identify the best performing machine learning models. We examined the behaviour of the best models, with respect to prediction errors and the impact of used features, to confirm their clinical relevance. We found that the model-based selection outperforms the conventional inclusion criteria, reducing by 20–25% the number of patients who show no progression. This result might lead to more efficient clinical trials.

Knee osteoarthritis (OA) is a chronic degenerative joint disease characterised by cartilage loss and changes in bones underneath it, causing pain and functional disability. The main clinical symptoms of knee OA are pain and stiffness, particularly after activity[1], leading to reduced mobility and quality of life, and eventually resulting in knee replacement surgery. OA is one of the leading causes of global disability in people aged 65 and older, and its burden is likely to increase in the future with the ageing of the population and rise in obesity worldwide[2].

OA is a heterogeneous disease where progression spreads over several years with periods of fast changes and periods of stability[3]. A major challenge in OA drug development is effective selection of patients to the clinical trials. In an ideal case, all selected patients would show disease progression within the trial period, and their response to the drug in trial would be properly assessed. However, identification of patients in need of treatment, that is those with a high probability of progression, is an open problem.

[1]School of Computing Science, Newcastle University, 1 Science Square, Newcastle, NE4 5TG, UK. [2]Department of Rheumatology & Clinical Immunology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, Netherlands. [3]Merck, Frankfurter Str. 250, 64293, Darmstadt, Germany. [4]Biosciences Institute, Newcastle University, International Centre for Life, Newcastle, NE1 3BZ, UK. [5]Immuno-inflammation Center of Therapeutic Innovation, Institut de Recherches Internationales Servier, Suresnes, France. [6]Novel Human Genetics Research Unit, GlaxoSmithKline, Collegeville, PA, 19426, USA. [7]Department of Orthopedics, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, Netherlands. [8]Department of Biomechanical Engineering, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, Netherlands. [9]Department of Regenerative Medicine, State Research Institute Centre for Innovative Medicine, Santariskiu 5, 08661, Vilnius, Lithuania. [10]Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Aapistie 5A, FIN-90230, Oulu, Finland. [11]Centre for Sport, Exercise and Osteoarthritis Research Versus Arthritis, Queen's Medical Centre, Nottingham, NG7 2UH, UK. ✉e-mail: jaume.bacardit@newcastle.ac.uk

	patients	periods	attributes	used timepoints	missing values
CHECK	1 002	3 001	513	0,2,5,8	34%
OAI	3 465	16 800	1 536	0,1,2,3,4,6,8	59%

Table 1. Summary of the main characteristics of the datasets used in this work.

To help analyse knee OA progression, the APPROACH consortium (a partnership of over 20 European clinical centres, research institutes, small enterprises and pharmaceutical companies) is running a 2-year observational study in 5 clinical centres from 4 European countries. One of the study objectives is to discover new markers of disease progression. The consortium recruits patients from centres with existing OA cohorts, and similarly to clinical trials, is interested in selecting only patients who will progress during the observation period.

The traditional approach to patient selection relies on expert knowledge and typically follows a set of consensus criteria defined by the American College of Rheumatology (ACR), mixed with a presence of limited joint damage (so further progression is possible) and significant pain complaints. When these criteria are satisfied, the patient's disease is expected to progress over time. However, the speed with which this will happen is unknown. This is a problem for clinical trials and short-term studies, like APPROACH, in which the observation time is typically limited to about 2 years.

The main hypothesis of this article is that machine learning can be more effective at identifying progressive patients than the traditional approach. We hypothesise that prediction models trained on historical data will be able to differentiate between patients for whom a fast progression happen during the observation period, and patients who show no progression or progress slowly and should not be selected to trials. Throughout the course of this article we examine different algorithms and learning process configurations, to finally develop predictive models for patient selection that outperform the conventional inclusion criteria used in clinical trials.

To train the models, and verify our hypothesis, we use longitudinal data from two large studies running in parallel in Europe and North America: the Cohort Hip and Cohort Knee (CHECK) study[4], and the Osteoarthritis Initiative (OAI) study[5]. We outline a data preprocessing strategy to handle missing values and different attribute types, and we define four classes of patients using clinically relevant measures of OA progression. We set up the experiments that allow us to estimate the typical performance of a model on out-of-sample instances and find the best approach to handle the class imbalance present in the data. We choose the best performing algorithm and test several of its multi-model/multi-label variants to further improve performance. We select the most effective configuration of parameters and train the final models on all data and estimate their performance. Then we interpret the behaviour of these models, by looking at the individual features contribution to the model output, and assess their clinical relevance. Next, we simulate two patient selection scenarios and compare the best model results against the selection with conventional clinical classification criteria. Finally, we include a discussion on limitations, the experiment design choices, related literature, and future work.

Materials and methods

Datasets. The CHECK cohort data used in this article were contributed by the CHECK steering committee (available upon request at http://check-onderzoek.nl/). Specifically, we used the clinical and X-ray image assessment (radiographic scoring and KIDA features[6]) data.

The OAI cohort data used in the preparation of this article were obtained from the Osteoarthritis Initiative (available at http://www.oai.ucsf.edu/). Specifically, we used the clinical, X-ray image assessment (semi-quantitative readings and joint space width measurements) and outcomes (knee replacements) data.

Both of these cohorts have been studied for over 10 years, and collected longitudinal data with typically yearly updates. For both cohorts we used the time points between the study baseline and the 8 year follow-up, for which the joint space width measurements (used in class definition) were available (see summary in Table 1). To maximise the size of the training set, instead of using only the baseline and 2 year follow-up time points, for every patient we used all available periods that were at least 2-year long (some periods were longer than two years, e.g. between CHECK time points 2–5 or 5–8). As a consequence, each instance in the training set represented a period, not a patient. We excluded all periods after a knee replacement, to avoid problems with a change in meaning of some attributes (e.g. pain would no longer be related to the knee but to issues with the prosthesis).

As we show in tab:datasets, both datasets contain a relatively large number of attributes and a small number of patients, together with a large proportion of missing values. This introduces a challenge to the machine learning algorithms and we tried to improve this balance with additional preprocessing steps (see below).

Preprocessing. We dropped all attributes with more than 50% missing values and all periods with over 40% missing attributes. These thresholds are quite conservative, as we tried to retain as much data as possible. We also dropped all attributes that did not vary across instances (i.e. had just a single non-null value), and thus were not useful in distinguishing between the classes. Finally, we removed attributes that could be exploited by the model, such as dates, visit numbers, barcodes and patient and staff IDs.

As the CHECK cohort is one of the recruitment sources for the APPROACH consortium, we spent extra time analysing the reasons behind the missing values and fixing them were possible. We filled forward values from the most recent time point, for attributes which values cannot change in the future (e.g. past diseases), and used a default value in place of a missing one where this was a reporting convention (e.g. for presence of rare disorders).

	N	P	S	P+S
CHECK	63% (1891)	12% (358)	20% (592)	5% (160)
OAI	74% (12502)	6% (953)	16% (2719)	4% (626)

Table 2. Balance between the classes for each dataset. Exact number of periods per class is given in brackets.

For both datasets, we assumed that all attributes with at most 10 different values are categorical. For CHECK, we additionally went through the cohort variable guide and manually identified ordinal and continuous attributes. This step was not practical for OAI, as its variable guide has almost 4000 pages.

We performed additional preprocessing during the model training. We imputed missing values, using only the values found in the training set (to avoid information leaks from the test set). We performed the imputation with the mode/mean value (for categorical/continuous attributes). We briefly tried other methods (cluster centroids, a vote of nearest neighbours), but as they did not produce better results, we settled for the simplest method.

The final step after imputation was the one-hot encoding of nominal attributes. That is, their replacement with dummy attributes, of which only one is "hot" at a time (set to 1, while others are zero). We encoded all categorical attributes with more than 2 distinct values, unless they were known to be ordinal.

Class definition. The APPROACH consortium decided to use similar patient categorisation to the OAI-based FNIH biomarker study[7], but defined more broadly and bounded in the observation time to 2 years. Patients were split into one non-progressive category (N), and three progressive categories related to pain (P), structure (S), and combined pain and structure (P + S).

To define the categories, the consortium relied on the measures of pain symptoms and structural damage at the beginning and at the end of a period. Pain was measured using the pain subscale from the WOMAC self-report questionnaire[8], which includes perceived level of pain during 5 different activities: walking, using stairs, in bed, sitting or lying, and standing upright. Structural progression was measured using radiographic readings of minimum joint space width (JSW) across both lateral and medial femorotibial compartments of the knee.

The exact definitions of the categories are given below:

- S period — a minimum total JSW must decrease by at least 0.3 mm per year,
- P period — patient must experience progressive or intense sustained pain (Eq. 1):
  - pain increase of at least 5 WOMAC points per year ($\Delta p \geq 5$) on 0–100 scale,
  - pain at the end of a period (p_e) must be substantial ($p_e \geq 40$),
  - for a rapid pain increase ($\Delta p \geq 10$), end pain can be lower ($p_e \geq 35$),
  - sustained pain must be substantial at both the start (p_s) and the end (p_e) of a period ($p_s \geq 40 \cap p_e \geq 40$).

$$((\Delta p \geq 5 \bigcap p_e \geq 40) \bigcup (\Delta p \geq 10 \bigcap p_e \geq 35)) \bigcup (p_s \geq 40 \bigcap p_e \geq 40) \quad (1)$$

For each period, the most affected knee (with greater JSW narrowing) and maximum pain (if reported for both knees) were used in the calculation of progression. When we could not measure the progression due to missing values, we excluded the period. This way, the class definition was never based on imputed numbers.

We assigned a period to the **P + S** category when criteria for both **P** and **S** were satisfied, and to the remaining **N** category if none were satisfied. We obtained imbalanced class distributions strongly skewed towards the non-progressive periods (see Table 2).

Experimental setup. All experiments were performed using the `scikit-learn` library[9] and its implementation of the machine learning algorithms. In data preprocessing, analysis and generation of statistics, we used `pandas`[10], `NumPy`[11] and `SciPy`[12]. For data visualisation, we used `seaborn`[13] and `Matplotlib`[14].

Measure of performance. To measure the classification performance we used the F_1 score[15]. It is commonly used in information retrieval, where relevant documents have to be identified amongst a large number of unrelated ones, and therefore is well suited for patient selection. F_1 score is defined as a harmonic mean of precision and recall, where *precision* is the probability that a (randomly selected) retrieved document is relevant, and *recall* is the probability that a (randomly selected) relevant document has been retrieved. In medical literature precision is known as positive predictive value and recall is equivalent to sensitivity.

Although F_1 score has been originally designed for binary classification, it can be extended to a multi-class case by averaging the F_1 scores across classes. Throughout this article we use weighted average of per class F_1 scores, with weights depending on the class instance frequency (to take into account the class imbalance).

Cross-validation. In all experiments we used out-of-sample estimation of the algorithm performance. That is, we kept some of the instances hidden from the algorithm during training, and used them later as an independent test set. Specifically, we followed the standard 10-fold stratified cross-validation (CV) protocol, in which the instances are split into 10 approximately equal-sized parts (folds) and the split preserves the overall class distribution within each fold. Each fold is then used in turn as a test set, and the remaining 9 folds are used as a training set. To score the method performance, rather than averaging the scores across all 10 folds, we pool the out-of-sample predictions together and use it to calculate a single score.

The cross-validation is repeated 10 times with different partitions into folds. As some of the machine learning algorithms are not deterministic, we also repeat the model training (25 times) with different random seeds (the seeds remain constant across folds and cross-validation repeats). We report typical performance of a configuration (algorithm + parameters), as a median score amongst the cross-validation repeats, where the score for each repeat is the median across all trained models.

Initial experiments. To test how well different machine learning algorithms can learn from the data, we initially simplified the problem to a case of balanced classification through down-sampling. We fixed the size of the classes to 150 for CHECK and 600 for OAI, and drew 11 different random samples of 600/2400 instances. For each sample we performed repeated cross-validation (as described in the previous section) using for each fold a fixed-size test set, and a subset of the training set of increasing size (10%, 20%, 100%), to obtain a learning curve.

We tested six machine learning algorithms with the default parameters:

- **logistic regression**[16] (using one-vs-rest scheme),
- **multinomial logistic regression** using cross-entropy loss with L-BFGS solver,
- **k nearest neighbours** classifier (kNN[17]) using KD tree (default k=5),
- **support vector classifier** (SVC[18]) using one-vs-rest scheme with **linear** kernel,
- **support vector classifier** using one-vs-rest scheme with the **Radial Basis Function** (RBF) kernel (default $C = 1.0, gamma = \frac{1}{num\_features}$),
- **random forest**[19] (with 100 trees (default in scikit-learn 0.22)).

For scale-sensitive algorithms (SVC and kNN) all attribute values in the training set were scaled to the [0, 1] range.

In these initial experiments, random forest (see *Results* was the best performing algorithm (in line with literature[20,21]), and we focused our further experiments on it.

Cost-sensitive learning. Random forest can be made cost-sensitive by incorporation of class weights to penalise the misclassification of the minority classes (as the weights influence the node split criteria). The cost-sensitive learning is an alternative to up/down sampling techniques that does not introduce artificial instances (as with up-sampling of the minority classes) and does not lose information (as with down-sampling of the majority classes). And specifically for random forest, the algorithm creators have demonstrated that the weighted variant performs better on imbalanced data, than on up/down-sampled ones[22].

To test the difference in performance between the cost-sensitive and the balanced learning, we first performed a repeated cross-validation (as before) using a full imbalanced dataset while incrementally increasing the training set size. Then we kept the imbalanced test sets unchanged, and down-sampled each of the imbalanced folds used to form the training set, to obtain a balanced training set that does not overlap with the imbalanced test set. We repeated this procedure 11 times with different sampling seeds. In the cost sensitive variant, we used weights inversely proportional to the class distribution in the full dataset.

The rationale behind this process is that regardless of the different training sets, the test sets have to remain the same in all cross-validation rounds, so that the performance scores obtained by the two strategies are truly comparable. With experiments set up this way, we are able to examine whether a larger training set is more important to performance than the class balance.

Multi-model methods. As we are trying to solve a multi-class problem, where the class labels are a combination of two clinical criteria (see *Class definition*), we have tested multi-model and multi-label strategies to further improve the performance of random forest. In particular, we first tested (1) a *one-vs-rest scheme*, in which a combination of 4 independent models is used, each trained to discriminate one class from the rest, and (2) a *multi-label classification*[23], in which a single model is trained to assign P and S labels independently (rather than to predict the class) that are later mapped to 4 classes. Finally, we combined the two strategies to create (3) a *duo classifier* that uses two independent models, each trained to predict a single label (P or S). We implemented this classifier as a wrapper class on top of the random forest algorithm that predicts one of the 4 class labels, but at the same time, provides independent P and S probabilities for each instance.

Parameter tuning. To tune the configuration of the duo classifier we exhaustively searched the space of 84 combinations of three key random forest parameters in the following range:

- **number of trees** ∈[100, 200, 400, 600, 800, 1000],
- **maximum tree depth** ∈[4, 5, 6, 7, 8, 9, 10],
- **split quality criterion** ∈[*gini*, *entropy*] — (standing for Gini impurity and information gain).

Because we tried multiple models, cross-validated performance of the best configuration is an optimistically biased estimate of the performance of the final model trained on all data. This "multiple induction" problem is conceptually equivalent to multiple hypothesis testing in statistics. To estimate the unbiased performance of the final model, we used a recently proposed bootstrap-based BBC-CV protocol[24]. It is a computationally efficient alternative to the popular nested cross-validation procedure and provides good bias estimation for datasets with 100+ instances.

BBC-CV uses the out-of-sample predictions to (1) select a configuration with best performance on a bootstrapped sample of instances, and (2) score the performance of the selected configuration on the out-of-bootstrap

instances only. The returned performance estimate is the average out-of-bootstrap score over all bootstrap iterations.

As we repeat each cross-validation 10 times, we used the most robust variant of the protocol — BBC-CV with repeats. It includes in the estimate the results from all CV-repeats, which reduces the variance introduced by the random partitioning into folds. The number of bootstraps in the protocol was set to 1000.

Recursive feature elimination. To test if a reduced set of features can lead to better performance, we added an inner 3-fold cross-validation loop that selects the best subset of features to use in model training. The inner loop operates on the training folds only. It starts from a full set of features and eliminates the worst, one by one, until only one feature is left. Then a subset of features that maximises inner cross-validation score is selected and used to train the model on the full training fold.

Model interpretation. As each tree in the random forest votes for a class label, it is possible to count how many times each of the features have contributed to the final decision and estimate the feature importance. The problem with the feature importance determined in this way, is that it treats all splits in a tree equally, while the early, close to the root splits, tend to have the most impact.

Therefore, we decided not to use the feature importance provided by the random forest, but to examine each tree using the TreeExplainer class from the SHAP module[25,26]. It provides consistent and locally accurate (per prediction) estimates of feature influence on the model output. It combines ideas from game theory (Shapley sampling values)[27] and local explanations (LIME method)[28] and goes beyond the impact magnitude, providing information on the direction of the influence (probability boost/reduce) in relation to the feature low/high values.

Comparison to the conventional inclusion criteria. To simulate conventional inclusion decisions, we used a logical conjunction of the following three criteria: (1) a combination of the ACR clinical classification criteria for knee OA[29], (2) the Kellgren & Lawrence grade of OA severity[30,31] between 1 and 3 (inclusive), and (3) pain complaints resulting in at least 40 points score on the WOMAC questionnaire. We applied the variant of ACR criteria that uses history, physical examination and radiographic findings. It requires presence of (1) pain in the knee and (2) one of: age over 50, less than 30 minutes of morning stiffness, crepitus (crackling noises) on active motion and osteophytes. We assumed the criteria are satisfied if one of the knees satisfy them.

To simulate selection with machine learning models we used two scenarios: ML-L (based on class labels) and ML-P (based on class probabilities). Both scenarios were based on predictions made by the best configuration of the duo classifier, specifically the median score model from the median cross-validation repeat.

In the ML-L scenario, we selected all instances classified as progressive (predicted to belong to the P, S, or P + S class). This scenario simplifies the task to a binary classification, and makes it comparable to the binary decision made using the conventional inclusion criteria.

In the ML-P scenario, for a more direct comparison, we selected the same number of instances as obtained with the conventional criteria. We used the progression probabilities $p(S)$ and $p(P)$ returned by the model to three-way sort the instances (in a descending order) by $p(P) + p(S)$, $p(S)$, and $p(P)$. Then we selected 1/3 of instances from each sorted group (to obtain balanced representation), in that exact order, disregarding the duplicates.

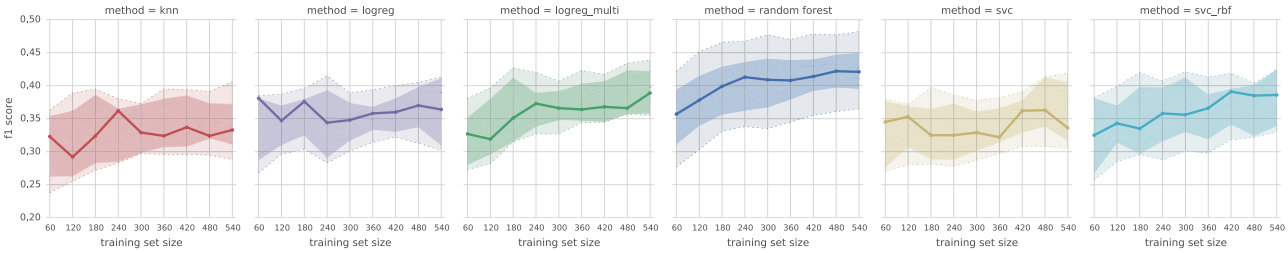
Results

Comparison of algorithms on balanced subsets. In the initial experiments on balanced subsets, the best performing algorithm was the random forest. For the CHECK dataset, the other algorithms were competitive only at small training set sizes, and otherwise were trailing 10% and more behind (see Fig. 1a). For the OAI dataset, logistic regression and SVC with the RBF kernel were closer, but on the other hand, the performance gap between random forest and the linear SVC or multi-modal regression was as large as 20% (see Fig. 1b).

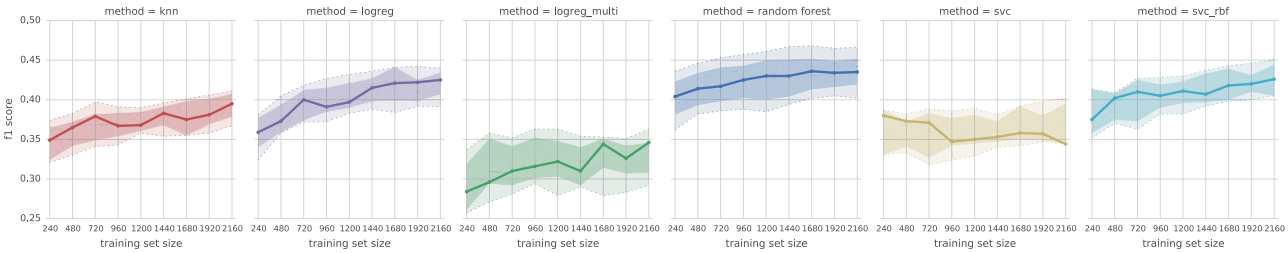
Performance on balanced and imbalanced training set. Figure 2 compares the performance of the cost-sensitive and balanced learning. Two observations arise from assessing the trade-off between balanced training set and potentially easier model training, and imbalanced training set with a larger number of instances to train on. Firstly, the bigger training set largely reduced the variance in model performance. Secondly, the typical (median) learning curve on the full set had a higher performance at every training set size compared. The difference was especially large in case of the OAI dataset (about 20% in relative numbers). Therefore, in all subsequent experiments we used the full imbalanced training set and the cost-sensitive learning.

Performance of multi-model methods. Figure 3 compares the performance of multi-label and multi-model strategies, to a single model 4-class random forest (indicated as "single"). Although all the strategies to some degree improved over the single model, the overall performance gain was minor, especially in case of the multi-label and one-vs-rest strategies. The *duo classifier* emerged as the best option, achieving a median F_1 score improvement of about 2% for CHECK and 1% for OAI. As a result, in subsequent experiments we used the *duo classifier*.

Random forest parameter tuning. Figures 4 and 5 show the typical performance of the *duo classifier* for different algorithm configurations. Each figure reports the F_1 score of the median run from the median CV-repeat. The best performing configurations for CHECK were located in a sweet spot around 800 trees of maximum depth of 9 (for information gain criterion) and depth 8 (for Gini impurity criterion). For OAI, we did not find a clear peak spot within the tested range of parameters. The best performing configuration was the one with largest

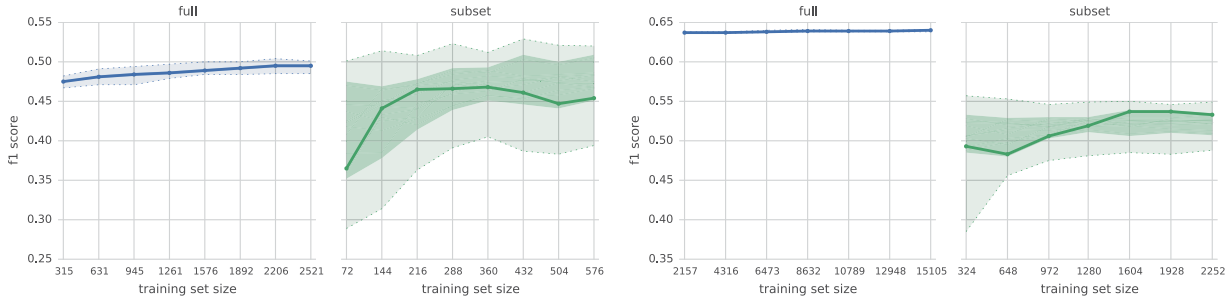


(a) CHECK dataset



(b) OAI dataset

Figure 1. Learning curves with F_1 score for models trained with different algorithms on balanced subsets of the dataset. The dotted lines show the total max/min score for each training set size across all subsets and CV-repeats. The solid lines (one per subset) represent elementwise median of curves for all CV-repeats. The thick line is the elementwise median of the 11 median curves shown. The shaded inner area contains all curves plus/minus their median average deviation (across all CV-repeats), and marks a range of the typical performance. For exact numbers and confidence intervals see Suppl. Tables S1 and S2.



(a) CHECK dataset

(b) OAI dataset

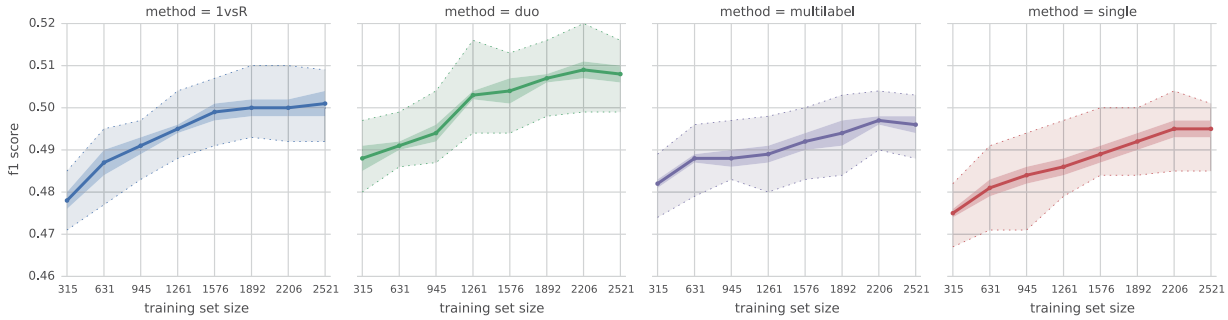
Figure 2. Learning curves with F_1 score for models trained on the full imbalanced training set (blue) or its balanced subsets (green), using the same test set. The dotted lines show the total max/min score for each training set size. The solid lines (one per subset) represent elementwise median of curves for all CV-repeats. The thick line shows the median score (or elementwise median curve across subsets). The shaded inner area represents the median average deviation (across all CV-repeats) around the median curve(s), and marks a range of the typical performance.

maximum depth of 10 and $\geq$400 trees. Perhaps configurations allowing for deeper trees could further improve the results.

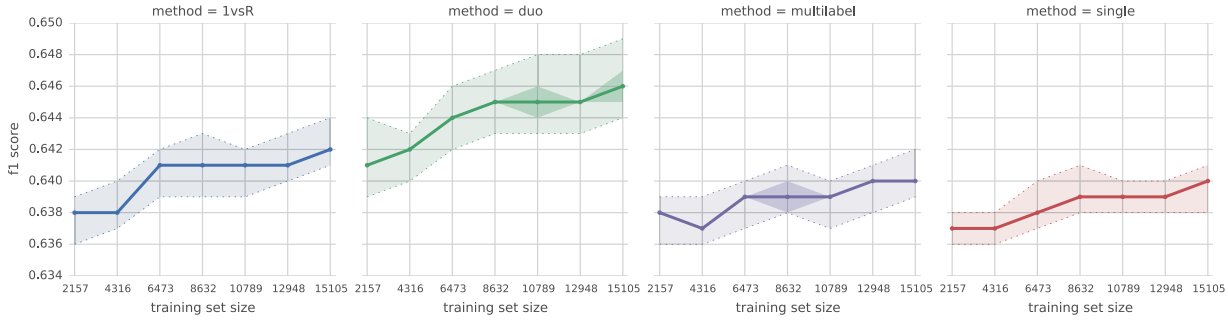
A general conclusion is that above 400 trees the improvement in performance is very small, and a difference in the maximum tree depth has the largest impact on the score. However, random forest is not over-training easily with more trees, and more trees can be useful (even if they do not improve performance), as they improve the reliability of the feature importance estimates. On the other hand, with increased depth and larger trees, their interpretability decreases and there is more potential for overfitting.

In subsequent experiments we used the best performing configuration with lowest median absolute deviation, preferring lower depth and less trees in case of ties, in particular: {800 trees, depth 9, *entropy* criterion} for CHECK, and {1000 trees, depth 10, *gini* criterion} for OAI.

The expected performance (F_1 score) of the **final models** trained on all data, estimated with the Bootstrap Bias Corrected Cross-Validation protocol (BBC-CV), was 0.584 — 95% CI (0.560, 0.609) for CHECK, and 0.689 — 95% CI (0.680, 0.698) for OAI. For both datasets, the estimate is the same (with respect to rounding) as the score of a typical run of the best configuration (median of median runs for each CV-repeat).

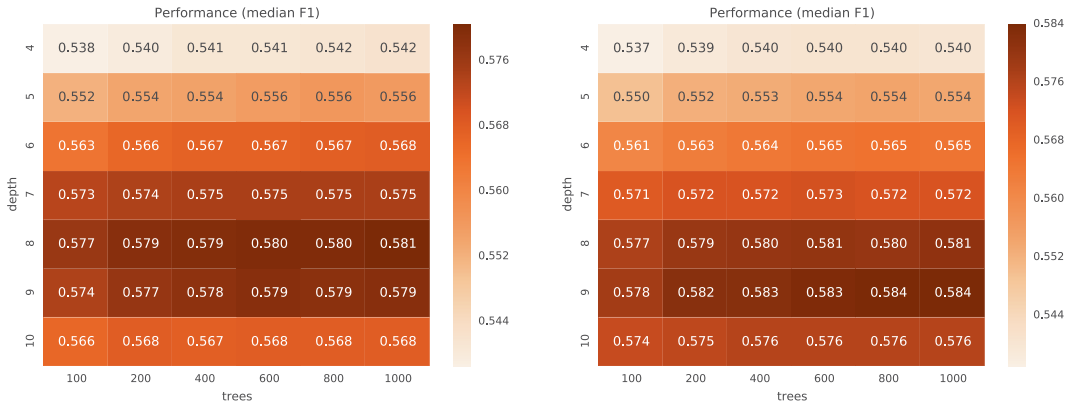


(a) CHECK dataset



(b) OAI dataset

Figure 3. Learning curves with F_1 score for multi-model/multi-label methods trained on imbalanced dataset. The dotted lines show the total max/min score across all CV-repeats for each training set size. The thick solid line shows the median score. The shaded area marks the median average deviation (across all CV-repeats) and contains ≥50% of scores. For exact numbers and confidence intervals see Suppl. Tables S3 and S4.



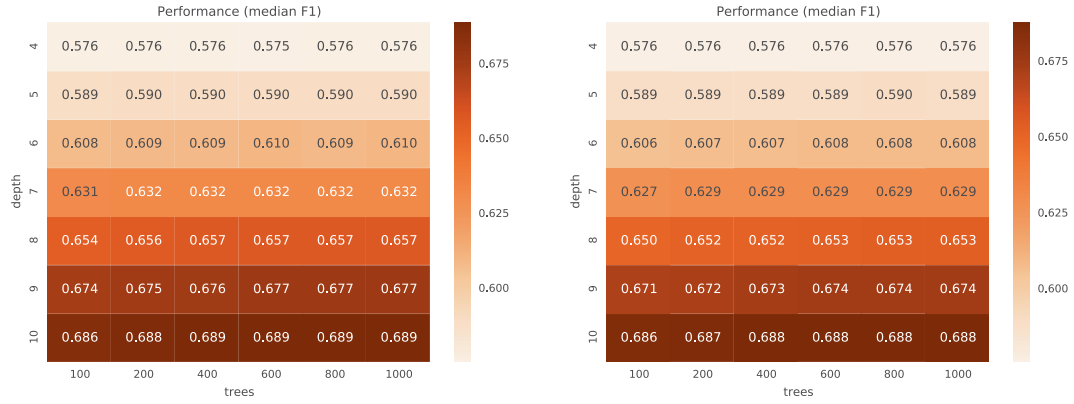
(a) with **Gini impurity** criterion

(b) with **information gain** criterion

Figure 4. Performance of different configurations of the duo classifier on the **CHECK** dataset.

Feature selection experiments. Table 3 summarises the results of experiments with the recursive feature elimination (RFE) procedure. As the table shows, the use of reduced set of features did not improve the model performance. Its median score was about 2% lower compared to configurations using all features. We counted the frequency with which each feature was selected (out of 100 selection rounds = 10 repeats * 10 folds). For CHECK only minimum JSW (left/right knee), WOMAC pain, WOMAC function, WOMAC total and height of the medial eminence (left/right) were selected 100% of the time (see Suppl. Figure S1). For OAI this subset was much larger, 181 features were selected every time, and overlapped with CHECK features (except eminence which was not measured in OAI), therefore not much can be learned there.

The main advantage of a smaller model (using a subset of features) is an easier interpretation, particularly with a substantial reduction to a median of just 12 features for CHECK (see Suppl. Figure S2). It is also an advantage from the clinical perspective, as data collection is costly and sometimes less measurements could be preferred over slightly better performance. However, it would not help much in case of the OAI models, where the median number of selected features was almost 20 times higher (see Suppl. Figure S3).



(a) with **Gini impurity** criterion

(b) with **information gain** criterion

Figure 5. Performance of different configurations of the duo classifier on the **OAI** dataset.

dataset	features		F_1 score	
			median	95% CI
CHECK	379	(all)	0.584	(0.583, 0.586)
	10–15	(subset)	0.573	(0.570, 0.575)
OAI	1299	(all)	0.69	(0.689, 0.690)
	209–364	(subset)	0.676	(0.675, 0.676)

Table 3. Performance of the best model using all features vs. a subset of features found with the RFE procedure. We report the median model score for a median CV-repeat and the 95% confidence interval around it (from binomial distribution). For the size of the selected subset of features, we report a range across all CV-repeats.

Model interpretation. Although the best learning strategy was to use all features, it does not mean that they all had the same impact. Figures 6 and 7 show features impact on the output of the **final model** trained on the entire CHECK dataset (see Suppl. Table S5a for feature description). For the **P** sub-predictor, the four most impactful features are the WOMAC scores (3 sub-scores and the total score). They all reduce the probability of assigning the **P** label if their value is low and boost that probability if their value is high (see left panel of Fig. 6). An example of an opposite direction of influence can be seen for the *rfys* feature (physical functioning from the SF-36 health survey), where higher values indicate a better health status.

For the **S** sub-predictor, the most impactful features are all related to structural degradation of the knee cartilage: the minimum JSW for both knees, with size of the osteophytes in medial tibia region and the varus angle (degree of outward bowing at the knee) further down. Low values of minimum JSW reduce the probability of assigning the **S** label. High values of minimum JSW, presence of large osteophytes and deviation in varus angle in range $[-2.5, 0.5]$ boost the probability.

Figures 8 and 9 show the impact of features on the output of the **final model** trained on OAI dataset (see Suppl. Table S5b for feature description). For the **P** sub-predictor, the most impactful features are the KOOS and WOMAC pain scores for the left and right knee.

For the **S** sub-predictor, some of the most impactful features are pain related: *DIRKN6* — pain level while walking in the last 7 days (part of the WOMAC questionnaire), and *P7RKACV* — knee pain severity in the last 7 days. But there are several impactful radiographic features as well, such as: *JSW175* — medial JSW at x=0.175*mm*, *MCAJSW* — average medial JSW, or *MCMJSW* — the minimum medial JSW. In the top 3, we can also find *GLCFQCV* — glucosamine frequency of use in past 6 months (glucosamine is a popular supplement used by OA patients).

A few features in the top make much less sense: *KIKBALL* — leg used to kick a ball, or *DFUCOLL* — difference in minutes between baseline and follow-up urine collection times, or *IMPIXSZ* — radiograph pixel size used in conversion to millimetres. This might be a sign of attribute exploitation, as with large number of attributes in OAI and not so many instances, the model might be finding dataset specific patterns, rather than discovering general rules, and perhaps these attributes should be removed from the dataset. Nevertheless, even if taken alone the contribution of a feature is difficult to explain, it might be useful in interaction with other features, e.g. *KIKBALL_3.0* indicates a person is ambipedal (has no dominant leg), which might trigger the use of radiographic features from both knees.

Simulated patient selection. We performed a selection from both datasets using the conventional clinical criteria, and compared that to two selection scenarios based on predictions of the best machine learning models: ML-L using the class labels, and ML-P using the class probabilities. In the simpler ML-L scenario, we selected all

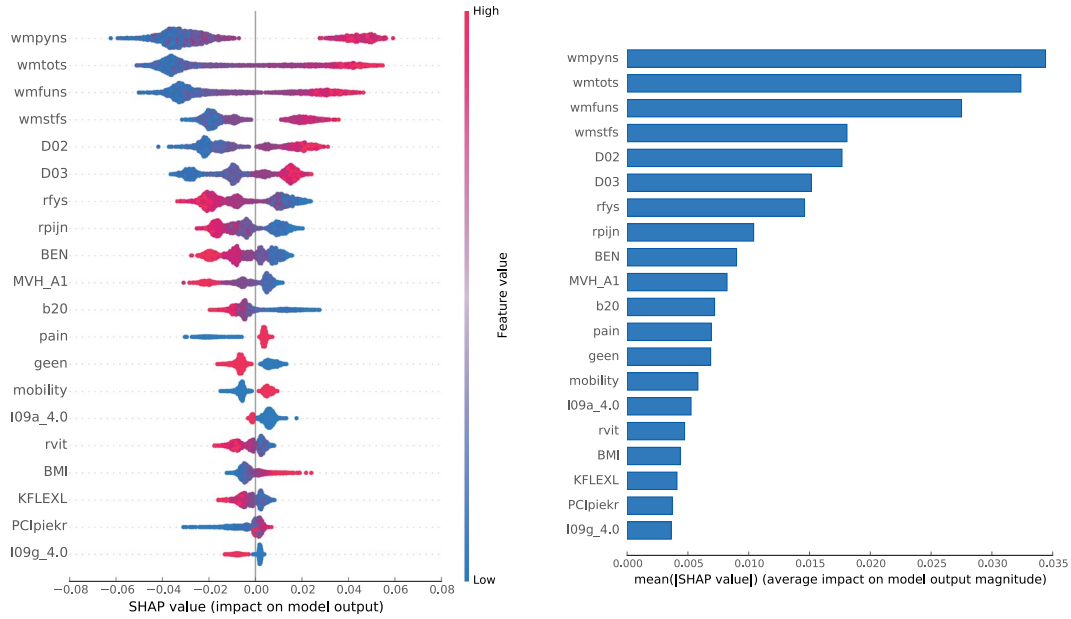


Figure 6. Features impact on **P** sub-predictor output for **CHECK** dataset. In the left panel, we show the distribution of the impact of a feature value on the model output across all instances. A positive SHAP value indicates a positive impact (probability boost). The colour represents the feature value (blue if low, red if high). In the right panel, we show the average impact magnitude for all instances. Features in both panels are ordered by their total impact.

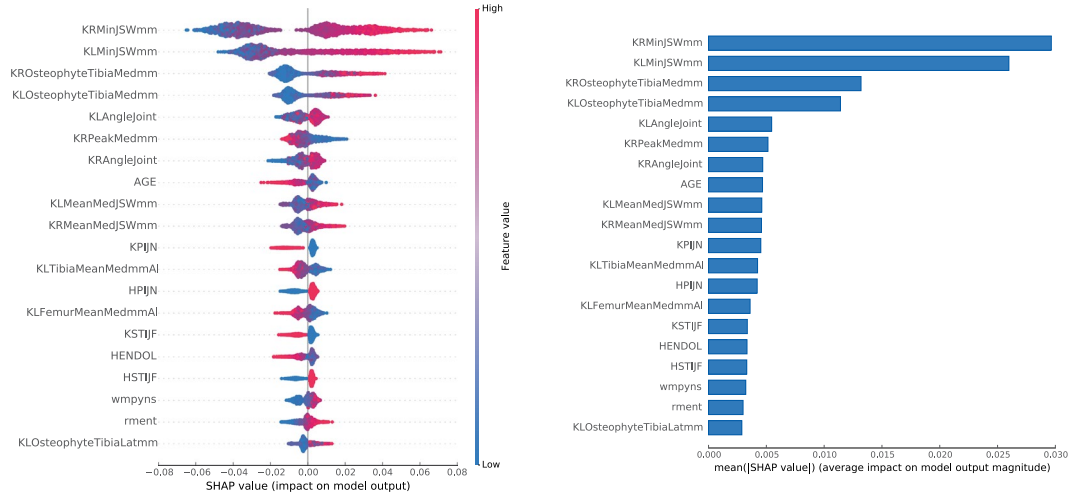


Figure 7. Features impact on **S** sub-predictor output for **CHECK** dataset. In the left panel, we show the distribution of the impact of a feature value on the model output across all instances. A positive SHAP value indicates a positive impact (probability boost). The colour represents the feature value (blue if low, red if high). In the right panel, we show the average impact magnitude for all instances. Features in both panels are ordered by their total impact.

instances predicted not to be in the non-progressive class (N). In the more refined ML-P scenario, we selected equal number of instances most likely to be in the P + S, S or P class.

Tables 4 and 5 summarise results of the selection with the conventional criteria and the ML-L selection scenario. The comparison between the two revealed several issues with the conventional criteria. Firstly, the retrieval of progressive periods was low (18% in total) for both CHECK and OAI, especially in the **S** category (only 7%). Secondly, the selection focused primarily on the **P** category, resulting in approximately half of the progressive periods from there. On the other hand, as desired, the percentage of retrieved non-progressive periods was low (5% for CHECK and 7% for OAI).

The ML-L selection scenario retrieved over 2 times more progressive periods (≈45% in total). In the **S** category the retrieval was 5 times higher than the conventional criteria result. The balance between the categories has

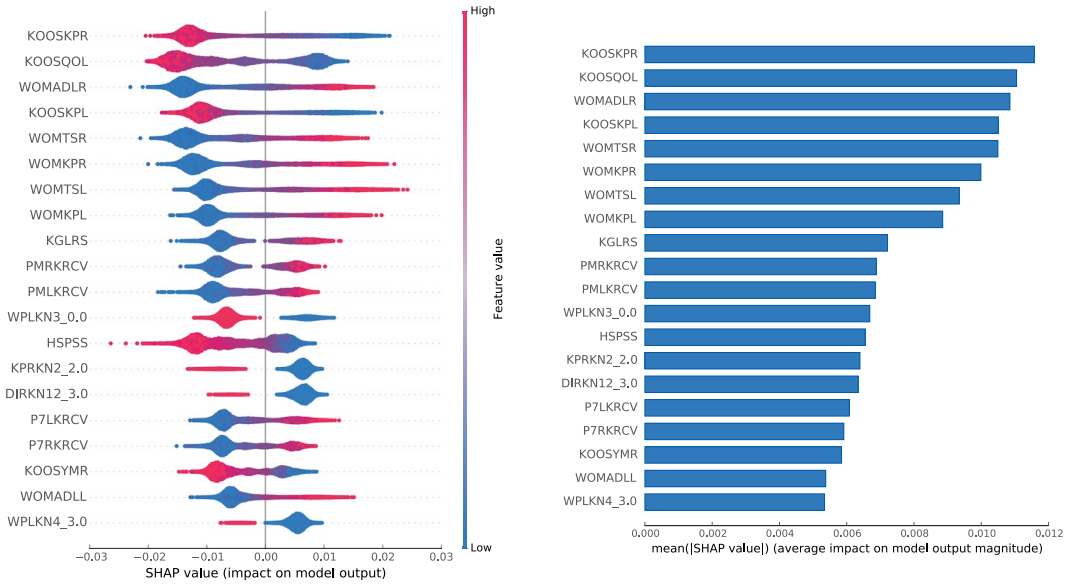


Figure 8. Features impact on **P** model output for **OAI** dataset. In the left panel, we show the distribution of the impact of a feature value on the model output across all instances. A positive SHAP value indicates a positive impact (probability boost). The colour represents the feature value (blue if low, red if high). In the right panel, we show the average impact magnitude for all instances. Features in both panels are ordered by their total impact.

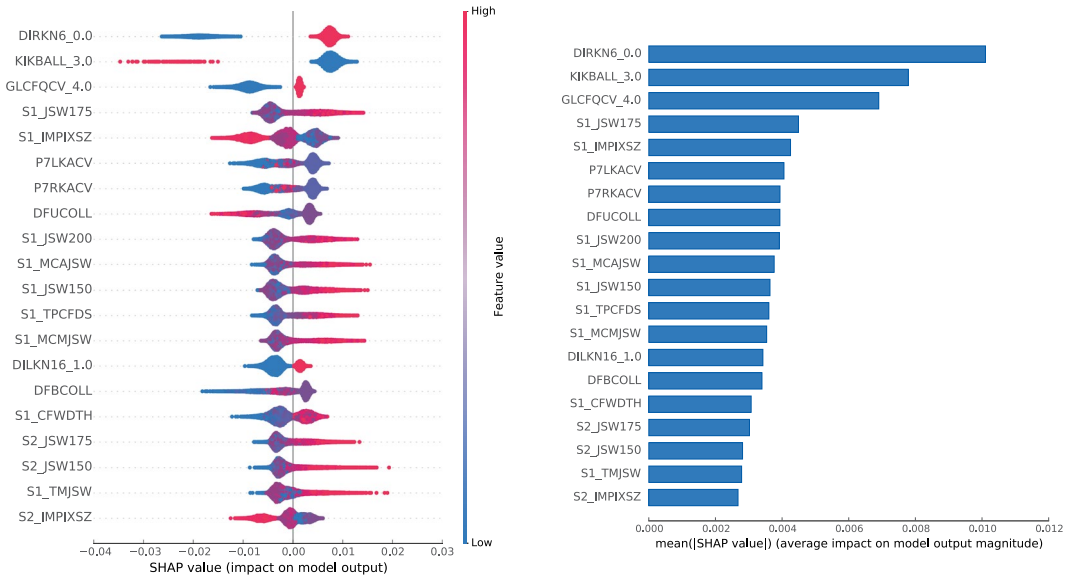


Figure 9. Features impact on **S** model output for **OAI** dataset. In the left panel, we show the distribution of the impact of a feature value on the model output across all instances. A positive SHAP value indicates a positive impact (probability boost). The colour represents the feature value (blue if low, red if high). In the right panel, we show the average impact magnitude for all instances. Features in both panels are ordered by their total impact.

improved for CHECK where **P** and **S** categories only differed by 2 p.p., but not for OAI, where the **S** category became dominant. Overall, we see that our machine learning models were less conservative (i.e. have made more non-N predictions) than the conventional criteria, which resulted in retrieving more progressive instances, at the cost of incorporating higher relative percentage of non-progressive ones.

Although in the ML-L scenario, the machine learning had some advantages in recall levels over the conventional criteria, it selected a larger number of non-progressive instances. It also selected 2.5–3 times more instances overall. To make a more direct comparison, in the ML-P scenario we selected the same total number of instances as obtained with the conventional criteria. The selection prioritised the instances more likely to progress and directly used the probabilities provided by the classifier.

	N(1704)			P(358)			S(579)			P+S(160)			not N
selection	abs	rel	recall	abs	rel	recall	abs	rel	recall	abs	rel	recall	recall
conventional	88	31%	5%	103	37%	29%	40	14%	7%	49	18%	31%	18%
ML–L	296	38%	17%	183	24%	51%	203	26%	35%	96	12%	60%	44%

Table 4. Subset of **CHECK** periods selected by the conventional clinical criteria and the ML-L scenario. The number of total instances of each category is reported next to the class name. For each category we report an absolute and relative number of included instances, and a recall percentage (how many instances of that category have been retrieved). The "not N" column shows the summarised recall percentage for all progressive instances.

	N(12489)			P(951)			S(2718)			P+S(626)			not N
selection	abs	rel	recall	abs	rel	recall	abs	rel	recall	abs	rel	recall	recall
conventional	858	52%	7%	366	22%	38%	187	11%	7%	229	14%	37%	18%
ML–L	2254	53%	18%	521	12%	55%	1059	25%	39%	385	9%	62%	46%

Table 5. Subset of **OAI** periods selected by the conventional clinical criteria and the ML-L scenario. The number of total instances of each category is reported next to the class name. For each category we report an absolute and relative number of included instances, and a recall percentage (how many instances of that category have been retrieved). The "not N" column shows the summarised recall percentage for all progressive instances.

dataset	selection	N	P	S	P + S
CHECK	conventional	31.4%	36.8%	14.3%	17.5%
	ML-P	25.4%	28.2%	22.5%	23.6%
OAI	conventional	52.3%	22.3%	11.3%	14.0%
	ML-P	38.5%	21.6%	22.3%	17.5%

Table 6. Comparison between selection with conventional clinical criteria and the ML-P scenario.

Table 6 shows the results of the ML-P selection scenario. Not only did it reduce the number of non-progressive instances compared to the conventional criteria (by ≈20% for CHECK and ≈25% for OAI), but it also increased the balance between the progressive categories (boosting selection from S and P + S, while reducing the bias towards P).

Discussion

We hypothesised that machine learning models predicting OA progression could be used to select fast progressing patients more effectively than the conventional inclusion criteria. In a search for the most performant learning process configuration, we used a careful evaluation focused on the median performance. For statistical stability of the results, we used repeated cross-validation and trained multiple models for each fold using different random seeds. We found random forest to stand out as the best learning algorithm. The cost-sensitive learning with random forest outperformed the balanced learning on down-sampled training set, and reduced the variance in model scores. The multi-model approach with the *duo classifier* further improved the results. Contrary to our expectations, we did not obtain better models with recursive feature elimination.

When predictions of the best models were used to simulate patient selection, we observed a substantial reduction in the number of undesired non-progressive cases. This findings could impact the future clinical trials design, and potentially improve their efficiency. A machine learning model similar to ours, could be applied to the screening data during the inclusion phase of a trial, and suggest which patients should be enrolled in the study. The screening visits could be continued until the trial is sufficiently enriched with patients who are likely to show disease progression within the trial period, and allow for more effective treatment evaluation.

Limitations and future work. A clear limitation of the experiment design, was the weak preprocessing strategy for the OAI dataset. We did not identify the ordinal attributes and therefore we applied one-hot encoding to every categorical attribute regardless of its semantics. A similar problem repeated for the continuous attributes with low number of unique values, which were treated as categorical and unnecessarily encoded. This led to a construction of less general decision trees, with splits relying on specific attribute values (rather than value ranges), and made the model less trustworthy from a clinical point of view.

A related issue is the clinical relevance of the features the models relied on. It is inevitable that some of the features will be exploited to make shortcut decisions, despite not representing any real knowledge. For that reason, it is important to look "inside" the models and iteratively refine the data representation in the training set, to gradually eliminate the potential for misuse. But this process is not trivial, as models can use hard to explain features (indirectly associated with progression) as a proxy for what is not directly observed. Although we eliminated

some of the feature misuse already (e.g. our first OAI models were misusing the image barcodes), still more work needs to be done in this regard, involving further dialogue with the domain experts.

In terms of further improvement of the model performance, it might be possible to achieve better results if the configuration of parameters used to train the *duo classifier* is not shared between its sub-classifiers. That is, each of the sub-classifiers could have been tuned separately, including a dedicated feature elimination procedure (perhaps even with more inner cross-validation folds), to maximise its individual performance. Whether that would lead to a better overall performance is a matter of experiment, as it might as well increase the risk of over-training. For certain, it would require a substantial additional computational effort — the longest RFE experiment we performed so far, already took over 200 CPU days on our HPC cluster (using Intel Xeon E5-2690 processor). Moreover, due to the sequential nature of the RFE procedure (features were eliminated one by one), it cannot be easily sped up through parallelisation.

Another question is, how easy would it be to implement our approach in clinical practice. The main obstacle would be the process of patients' data collection. It is usually performed on a rolling basis (over the course of several months), due to logistics reasons (e.g. limited access to equipment or personnel), which makes a single selection step, as we performed in this work, impractical. Therefore, further work is needed on extending this approach towards a multi-step selection, in which decisions are made on small batches of patients as their data become available, without sacrificing the overall selection quality.

Choice of performance measure. In this work, inspired by the similarity of the patient selection problem to the task of document retrieval, we decided to measure the classification performance with F_1 score. Below, we briefly discuss the advantages and drawbacks of several alternative measures.

Area under the ROC curve (AUC) is commonly used in medical binary classification tasks such as cases vs. controls analysis. Although a generalisation to multi-class problems, *M-score*, has been proposed by Hand and Till[32], the use of AUC for model comparison has been strongly criticised by Hand himself. He not only pointed out problems with comparison of the crossing ROC curves (where difference in AUC creates false impression that one curve dominates the other), but also demonstrated the measure incoherence[33] (AUC evaluates different classifiers with a different metric, as it depends on the score distributions, which depend on the classifier). Hand proposed *H-measure* as a replacement for the AUC, but it has been only defined for binary classification.

Matthew's Correlation Coefficient (MCC) is another measure of binary classification performance[34] that has been extended to handle multi-class problems[35]. Its main merit is in taking into account true negatives (accuracy or F_1 do not), which makes MCC especially useful when negative examples are the minority. Unfortunately, this is not the case in the patient selection task.

Measures based on the error matrix (like F_1 score or MCC), do not take into account the distance in the class probability space (they treat every mistake the same, regardless of its scale). There are several measures that do, but they lack in other aspects. For example, area under the precision-recall curve (AUPRC) does not generalise to a multi-class case. Log-loss or the Brier distance can handle multi-class problems, but they do not address the class imbalance directly. Perhaps the patient selection task would benefit from a dedicated measure of performance designed to align with the specific recruitment requirements.

Related work. Although several long-term OA clinical studies have been completed and their outcomes analysed in detail, very little research has been done on improving the patient selection process. To our best knowledge, this work is a first attempt at building machine learning models that can compete with the established clinical practice.

Our approach differs from most of the analyses found in the literature in two important ways. Firstly, it does not focus on determining the risk factors, but on the prediction of the disease progression. Secondly, it defines the progression within a strict time window and targets the change in fine-grained radiographic measurements (JSW), rather than just a categorical difference in the KL/JSN grade.

Most of the previous works do not focus on disease progression, but analyse OA incidence instead, where a patient can either be diagnosed with OA (typically when KL grade ≥ 2) or be "OA free" (when KL grade ≤ 1). The incidence of disease is then defined, as a change in diagnosis of the same knee between the baseline and the follow-up visit, and is analysed with statistical methods to determine the risk factors (usually odds ratios with univariate analysis of variance, or multivariate logistic regression). Some authors go a bit further and test the logistic regression models on a binary classification task (cases vs. controls)[36–38] hand-picking the input variables. However, as Jamshidi *et al.* point out in their recent perspective article[39], very few authors reach beyond statistical analysis and build machine learning models.

Yoo *et al.*[40] trained an artificial neural network with 7 inputs and 3 hidden layers to directly predict the KL grade, obtaining AUC >0.8. However, they only focused on discriminating between KL grade levels at baseline, rather than trying to predict future disease progression. Similar results were obtained with random forest by Minciullo *et al.*[41] who were able to discriminate between cases and controls with AUC >0.85, but in the prediction task (same cohort, OA incidence after 84 months) achieved a much lower score (≈ 0.6). Better OA incidence prediction (AUC >0.8) was reported by Lazzarini *et al.*[42] who used random forest with an iterative feature elimination heuristic (RGIFE).

These results are not directly comparable, as the models were trained on data from different cohorts. As a consequence, the models operated on a different input, and used inconsistent definition of the outcome (the OA incidence was defined over a period of varying length: 10[38], 7[41], or 2.5[42] years). Moreover, due to the AUC measure incoherence discussed earlier, any comparison between these models would be, at most, approximate.

When it comes to the definition of the progression used in this article, in many aspects it is similar to the definition used by the FNIH OA Biomarkers Consortium (e.g.[7,43]). They likewise defined four categories of patients (N, P, S, and P + S) based on the change in WOMAC/JSW over time, but flexibly allowed the progression to happen at 2, 3 or 4 year follow-up. In contrast to our fixed 2 year time period, this does not select for a fast progression. Furthermore, the analysis performed in these works, is again focused on the risk factors only. In the best case, a test of discriminatory power is performed (without correcting for overfitting) but no independent prediction is attempted. Notable exception is the work by Hafezi-Nejad *et al.*[44]. who used a small artificial neural network with 10 inputs and 1 hidden layer to predict the joint space loss, and with a single training/test set random split and 100 runs, obtained an average AUC of 0.669.

Conclusions

The aim of this work has been to test if the machine learning models can be more predictive of the future knee OA progression than the conventional clinical selection criteria. We focused on a short progression time window typical for clinical trials. Using data from two long-term knee OA studies (CHECK and OAI), we experimented with different learning strategies to build the final models, and obtained the best results with a custom-made *duo classifer*. The model-based selection, compared to the conventional criteria, resulted in 20–25% less non-progressive cases and more balanced retrieval of progressive cases.

These results put into question the effectiveness of the conventional selection criteria, which although straightforward to apply in practice, were found to be less predictive of the future disease progression. At the same time, these results reveal a potential to develop more precise screening tools, leading to better designed clinical trials, and in consequence, to more successful evaluation of therapies, which is important for patients, scientific community, pharmaceutical industry and the ageing society in general.

Further work is needed before this potential is fully understood. Our approach needs to be implemented into the clinical practice, and tested in a real study. That involves a number of challenges, from methodology of the model evaluation to logistics of the selection process. We hope to solve some of them in the APPROACH study recruitment process, and based on its future results, assess the practical impact of the model-based selection.

Disclaimer

This communication reflects the views of the authors and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. Roles of all contributors (whether formally listed as authors or named in acknowledgements) are described using the CRediT taxonomy[45].

Data availability

The data from machine learning experiments performed during this study are available under a CC0 licence at https://doi.org/10.25405/data.ncl.10043060.

The CHECK and OAI cohorts are controlled access datasets available from their owners at https://doi.org/10.17026/dans-252-qw2n and https://oai.epi-ucsf.org/.

Received: 25 October 2019; Accepted: 20 April 2020;
Published online: 21 May 2020

References

1. Felson, D. T. Developments in the clinical understanding of osteoarthritis. *Arthritis Research and Therapy* **11**, 203, https://doi.org/10.1186/ar2531 (2009).
2. Cross, M. *et al*. The global burden of hip and knee osteoarthritis: estimates from the Global Burden of Disease 2010 study. *Annals of the Rheumatic Diseases* **73**, 1323–1330, https://doi.org/10.1136/annrheumdis-2013-204763 (2014).
3. Felson, D. *et al*. Progression of osteoarthritis as a state of inertia. *Annals of the Rheumatic Diseases* **72**, 924–929, https://doi.org/10.1136/annrheumdis-2012-201575 (2012).
4. Wesseling, J. *et al*. Cohort Profile: Cohort Hip and Cohort Knee (CHECK) study. *International Journal of Epidemiology* **45**, 36–44, https://doi.org/10.1093/ije/dyu177 (2016).
5. Eckstein, F., Kwoh, C. K. & Link, T. M. Imaging research results from the Osteoarthritis Initiative (OAI): a review and lessons learned 10 years after start of enrolment. *Annals of the Rheumatic Diseases* **73**, 1289–1300, https://doi.org/10.1136/annrheumdis-2014-205310 (2014).
6. Marijnissen, A. *et al*. Knee Images Digital Analysis (KIDA): a novel method to quantify individual radiographic features of knee osteoarthritis in detail. *Osteoarthritis and Cartilage* **16**, 234–243, https://doi.org/10.1016/j.joca.2007.06.009 (2008).
7. Eckstein, F. *et al*. Brief Report: Cartilage thickness change as an imaging biomarker of knee osteoarthritis progression: data from the Foundation for the National Institutes of Health Osteoarthritis Biomarkers Consortium. *Arthritis & Rheumatology* **67**, 3184–3189, https://doi.org/10.1002/art.39324 (2015).
8. Bellamy, N. WOMAC: a 20-year experiential review of a patient-centered self-reported health status questionnaire. *The Journal of Rheumatology* **29**, 2473–2476, http://www.jrheum.org/content/29/12/2473 (2002).
9. Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830, http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (2011).
10. McKinney, W. pandas: a foundational Python library for data analysis and statistics. In *Workshop on Python for High-Performance and Scientific Computing (PyHPC 2011)* (Seattle, USA, 2011), https://www.dlr.de/sc/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf.
11. Oliphant, T. E. Python for Scientific Computing. *Computing in Science and Engineering* **9**, 10–20, https://doi.org/10.1109/MCSE.2007.58 (2007).
12. Jones, E. T. P. *et al*. SciPy: Open source scientific tools for Python (2001–), https://www.scipy.org/scipylib/.
13. Waskom, M. seaborn: statistical data visualization (2013–), http://seaborn.pydata.org/.
14. Hunter, J. D. Matplotlib: a 2D graphics environment. *Computing in Science and Engineering* **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55 (2007).
15. Sasaki, Y. The truth of the F-measure. Tech. Rep., School of Computer Science, University of Manchester (2007), https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

16. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* **9**, 1871–1874, http://www.csie.ntu.edu.tw/cjlin/papers/liblinear.pdf (2008).
17. Wu, X. *et al*. Top 10 algorithms in data mining. *Knowledge and Information Systems* **14**, 1–37, https://doi.org/10.1007/s10115-007-0114-2 (2008).
18. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**, 27:1–27:27, https://doi.org/10.1145/1961189.1961199 (2011).
19. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32, https://doi.org/10.1023/a:1010933404324 (2001).
20. Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* **15**, 3133–3181, http://jmlr.org/papers/v15/delgado14a.html (2014).
21. Zhang, C., Liu, C., Zhang, X. & Almpanidis, G. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications* **82**, 128–150, https://doi.org/10.1016/j.eswa.2017.04.003 (2017).
22. Chen, C., Liaw, A. & Breiman, L. Using random forest to learn imbalanced data. Tech. Rep., University of California, Berkeley (2004).
23. Tsoumakas, G. & Katakis, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* **3**, 1–13, https://doi.org/10.4018/jdwm.2007070101 (2007).
24. Tsamardinos, I., Greasidou, E. & Borboudakis, G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine Learning* **107**, 1895–1922, https://doi.org/10.1007/s10994-018-5714-4 (2018).
25. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *Computing Research Repository* arXiv:1802.03888v2 https://arxiv.org/abs/1802.03888 (2018).
26. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In I., Guyon *et al*. (eds.) *Advances in Neural Information Processing Systems (NIPS 2017)*, 4765–4774 (Long Beach, CA, USA, 2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
27. Å trumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* **41**, 647–665, https://doi.org/10.1007/s10115-013-0679-x (2014).
28. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?": explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144 (San Francisco, USA, 2016), https://doi.org/10.1145/2939672.2939778.
29. Altman, R. *et al*. Development of criteria for the classification and reporting of osteoarthritis: Classification of osteoarthritis of the knee. *Arthritis & Rheumatism* **29**, 1039–1049, https://doi.org/10.1002/art.1780290816 (1986).
30. Kohn, M. D., Sassoon, A. A. & Fernando, N. D. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research* **474**, 1886–1893, https://doi.org/10.1007/s11999-016-4732-4 (2016).
31. Kellgren, J. & Lawrence, J. Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases* **16**, 494–502, https://doi.org/10.1136/ard.16.4.494 (1957).
32. Hand, D. J. & Till, R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* **45**, 171–186, https://doi.org/10.1023/A:1010920819831 (2001).
33. Hand, D. J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* **77**, 103–123, https://doi.org/10.1007/s10994-009-5119-5 (2009).
34. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442–451, https://doi.org/10.1016/0005-2795(75)90109-9 (1975).
35. Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry* **28**, 367–374, https://doi.org/10.1016/j.compbiolchem.2004.09.006 (2004).
36. Zhang, W. *et al*. Nottingham knee osteoarthritis risk prediction models. *Annals of the Rheumatic Diseases* **70**, 1599–1604, https://doi.org/10.1136/ard.2011.149807 (2011).
37. Kinds, M. *et al*. Evaluation of separate quantitative radiographic features adds to the prediction of incident radiographic osteoarthritis in individuals with recent onset of knee pain: 5-year follow-up in the CHECK cohort. *Osteoarthritis and Cartilage* **20**, 548–556, https://doi.org/10.1016/j.joca.2012.02.009 (2012).
38. Kerkhof, H. *et al*. Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. *Annals of the Rheumatic Diseases* **73**, 2116–2121, https://doi.org/10.1136/annrheumdis-2013-203620 (2014).
39. Jamshidi, A., Pelletier, J.-P. & Martel-Pelletier, J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nature Reviews Rheumatology* **15**, 49–60, https://doi.org/10.1038/s41584-018-0130-5 (2019).
40. Yoo, T. K., Kim, D. W., Choi, S. B., Oh, E. & Park, J. S. Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: A cross-sectional study. *PLOS ONE* **11**, 1–17, https://doi.org/10.1371/journal.pone.0148724 (2016).
41. Minciullo, L., Bromiley, P. A., Felson, D. T. & Cootes, T. F. Indecisive trees for classification and prediction of knee osteoarthritis. In Q., Wang, Y., Shi, H.-I., Suk & K., Suzuki (eds.) *International Workshop on Machine Learning in Medical Imaging (MLMI 2017)*, 283–290 (Quebec City, Canada, 2017), https://doi.org/10.1007/978-3-319-67389-9_33.
42. Lazzarini, N. *et al*. A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. *Osteoarthritis and Cartilage* **25**, 2014–2021, https://doi.org/10.1016/j.joca.2017.09.001 (2017).
43. Kraus, V. B. *et al*. Predictive validity of biochemical biomarkers in knee osteoarthritis: data from the FNIH OA Biomarkers Consortium. *Annals of the Rheumatic Diseases* **76**, 186–195, https://doi.org/10.1136/annrheumdis-2016-209252 (2017).
44. Hafezi-Nejad, N. *et al*. Prediction of medial tibiofemoral compartment joint space loss progression using volumetric cartilage measurements: Data from the FNIH OA biomarkers consortium. *European Radiology* **27**, 464–473, https://doi.org/10.1007/s00330-016-4393-4 (2017).
45. Brand, A., Allen, L., Altman, M., Hlava, M. & Scott, J. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing* **28**, 151–155, https://doi.org/10.1087/20150211 (2015).

Acknowledgements

We thank Janet Wesseling for providing explanations of the CHECK codebook, Anne-Christine Bay-Jensen, Francis Berenbaum, Ida Haugen, Marieke Loef, Anne Marijnissen, Margreet Kloppenburg, Sjaak Peelen, Jérémie Sellam and Erwin van Spil for comments and suggestions on the draft of this manuscript, and Janneke Boere and Leonie Hussaarts for coordination of the research activity.

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under Grant Agreement no. 115770, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution. See http://www.imi.europa.eu/ and http://www.approachproject.eu/.

This research used the High Performance Computing cluster at the School of Computing at Newcastle University.

Author contributions

Jaume Bacardit: Conceptualisation, Methodology, Resources, Writing – Original Draft, Supervision, Project Administration, Funding Acquisition. **Anne-Christine Bay-Jensen**: Writing - Review & Editing, Funding Acquisition. **Francis Berenbaum**: Writing - Review & Editing. **Janneke Boere**: Project Administration. **Ida Haugen**: Writing - Review & Editing. **Leonie Hussaarts**: Project Administration. **Margreet Kloppenburg**: Writing - Review & Editing. **Christoph Ladel**: Conceptualisation, Supervision, Project Administration, Funding Acquisition. **Floris Lafeber**: Conceptualization, Writing - Review & Editing, Funding Acquisition. **Jonathan Larkin**: Conceptualization, Writing - Review & Editing, Project Administration, Funding Acquisition. **Marieke Loef**: Writing - Review & Editing. **John Loughlin**: Conceptualisation, Resources, Supervision, Project Administration, Funding acquisition. **Anne Marijnissen**: Writing - Review & Editing. **Ali Mobasheri**: Conceptualization, Funding Acquisition. **Sjaak Peelen**: Writing - Review & Editing. **Florence Petit Dop**: Conceptualization, Writing - Review & Editing, Funding Acquisition. **Jérémie Sellam**: Writing - Review & Editing. **Erwin van Spil**: Writing - Review & Editing. **Harrie Weinans**: Conceptualization, Funding Acquisition, Project Administration. **Paco Welsing**: Conceptualisation, Methodology, Writing – Review & Editing. **Paweł Widera**: Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Writing — Original Draft, Visualisation.

Competing interests

Florence Petit Dop is employee of Servier. Jonathan Larkin is employee and stockholder at GlaxoSmithKline. Christoph Ladel is employee of Merck. All other authors declare no potential conflict of interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-64643-8.

Correspondence and requests for materials should be addressed to J.B.

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

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