Universiteit Leiden
The Netherlands

A machine learning approach reveals features related to clinicians' diagnosis of clinically relevant knee osteoarthritis

Wang, Q.K.; Runhaar, J.; Kloppenburg, M.; Boers, M.; Bijlsma, J.W.J.; Bacardit, J.; ... ; CREDO Experts Grp

Citation
Wang, Q. K., Runhaar, J., Kloppenburg, M., Boers, M., Bijlsma, J. W. J., Bacardit, J., & Bierma-Zeinstra, S. M. A. (2022). A machine learning approach reveals features related to clinicians' diagnosis of clinically relevant knee osteoarthritis. *Rheumatology*. doi:10.1093/rheumatology/keac707

Version:	Publisher's Version
License:	Licensed under Article 25fa Copyright Act/Law (Amendment Taverne)
Downloaded from:	https://hdl.handle.net/1887/3562171

Note: To cite this publication please use the final published version (if applicable).

Rheumatology, 2022, **00**, 1–8
https://doi.org/10.1093/rheumatology/keac707
Advance access publication 19 December 2022
Original article

British Society for Rheumatology RHEUMATOLOGY OXFORD

Downloaded from https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keac707/6936417 by Universiteit Leiden - LUMC user on 27 February 2023

Clinical science

A machine learning approach reveals features related to clinicians' diagnosis of clinically relevant knee osteoarthritis

Qiuke Wang [1]*, Jos Runhaar [1], Margreet Kloppenburg[2], Maarten Boers[3], Johannes W. J. Bijlsma[4], Jaume Bacardit [5], Sita M. A. Bierma-Zeinstra[1,6], The CREDO Experts Group[†]

[1]Department of General Practice, Erasmus MC University Center Rotterdam, Rotterdam, The Netherlands
[2]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands
[3]Department of Epidemiology and Biostatistics, Amsterdam UMC, Amsterdam, The Netherlands
[4]Department of Rheumatology and Clinical Immunology, University Medical Centre Utrecht, Utrecht, The Netherlands
[5]School of Computing, Newcastle University, Newcastle, UK
[6]Department of Orthopaedics and Sport Medicine, Erasmus MC University Center Rotterdam, Rotterdam, The Netherlands

*Correspondence to: Qiuke Wang, Department of General Practice, Erasmus MC University Medical Centre Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands. E-mail: q.wang@erasmusmc.nl

[†]See Acknowledgements section for a list of the CREDO Experts Group.

Abstract

Objectives: To identify highly ranked features related to clinicians' diagnosis of clinically relevant knee OA.

Methods: General practitioners (GPs) and secondary care physicians (SPs) were recruited to evaluate 5–10 years follow-up clinical and radiographic data of knees from the CHECK cohort for the presence of clinically relevant OA. GPs and SPs were gathered in pairs; each pair consisted of one GP and one SP, and the paired clinicians independently evaluated the same subset of knees. A diagnosis was made for each knee by the GP and SP before and after viewing radiographic data. Nested 5-fold cross-validation enhanced random forest models were built to identify the top 10 features related to the diagnosis.

Results: Seventeen clinician pairs evaluated 1106 knees with 139 clinical and 36 radiographic features. GPs diagnosed clinically relevant OA in 42% and 43% knees, before and after viewing radiographic data, respectively. SPs diagnosed in 43% and 51% knees, respectively. Models containing top 10 features had good performance for explaining clinicians' diagnosis with area under the curve ranging from 0.76–0.83. Before viewing radiographic data, quantitative symptomatic features (i.e. WOMAC scores) were the most important ones related to the diagnosis of both GPs and SPs; after viewing radiographic data, radiographic features appeared in the top lists for both, but seemed to be more important for SPs than GPs.

Conclusions: Random forest models presented good performance in explaining clinicians' diagnosis, which helped to reveal typical features of patients recognized as clinically relevant knee OA by clinicians from two different care settings.

Keywords: knee OA, clinician's diagnosis, machine learning, CHECK cohort

Rheumatology key messages

- Which clinical features drive clinicians to make the osteoarthritis diagnosis remains unclear.
- Herewith, diagnoses by general practitioners and secondary care physicians were analysed using machine learning approaches.
- Results illustrated typical vignettes of clinically relevant knee osteoarthritis from two different care settings.

Introduction

As no gold-standard definition has been established for OA [1], diagnosing knee OA is never trivial. Clinical classification/diagnostic criteria, such as the ACR [2], the EULAR [3] and the National Institute for Health and Care Excellence (NICE) criteria [4], have been proposed to potentially help identify knee OA patients in research/clinical settings. However, the clinical relevance of these criteria is unclear or insufficiently validated [3, 5, 6].

In addition to diagnoses based on predefined criteria, clinicians' diagnoses are often used as a reference standard, because it usually reflects the treatment decision-making process in daily clinical practice. For instance, observational studies using registry data could identify OA patients according to recorded clinicians' diagnosis [7–9] and clinical trials, to facilitate participant recruitment, could use recorded diagnosis for narrowing population screening spectrum [10, 11]. Furthermore, registered clinicians' diagnosis was sometimes

Received: 14 July 2022. **Accepted:** 9 December 2022
© The Author(s) 2022. Published by Oxford University Press on behalf of the British Society for Rheumatology. All rights reserved.
For permissions, please email: journals.permissions@oup.com

employed for deducing regional OA prevalence and incidence rate, which would impact future public health planning [12–14]. Despite the widespread use, a paucity of studies exists on which clinical features drive clinicians to make the OA diagnosis. Moreover, the focus in diagnosis-making could be different between primary and secondary care (given the different specialty knowledge) and between the circumstances with and without radiographs [15]; the features are preferably to be identified for each situation.

One of the major challenges in distinguishing important features from numerous patient characteristics is establishing a proper analysis framework. Integrating a large number of clinical features into statistical models will result in the dimensionality problem, where the number of features exceeds the model capacity [16]. Machine learning approaches coupled with feature selection methods have been shown to perform well in tackling such issues and are capable of identifying important features from high-dimensional data [17].

Here, we performed a *post hoc* analysis on the data from a previous task [15, 18], in which general practitioners (GPs) and secondary care physicians (SPs) were recruited to evaluate patients' longitudinal medical data to diagnose whether clinically relevant knee OA was present. With the help of machine learning algorithms, the primary aim of this study was to identify the highly-ranked clinical features related to the diagnosis by GPs and SPs, in the situation with or without access to radiographs, respectively.

Methods

Patient data

We obtained patient data from the CHECK cohort (a longitudinal cohort study of patients with knee/hip complaints suspected of early stage OA and followed for 10 years). The inclusion and exclusion criteria of CHECK were explained in a previous study [19]. For the present study, patients with knee complaints at baseline and data available from 5-year to 10-year (T5 to T10) follow-up were included. This study complies with the Declaration of Helsinki; Medical Ethics Committee of the University Medical Center Utrecht has approved the protocol of the CHECK cohort and all patients have signed informed consent. The current report follows the MI-CLAIM guideline for machine learning papers (see Checklist in Supplementary Data S1, available at *Rheumatology* online) [20].

We obtained patient T5, T8 and T10 follow-up data, and categorized these data into two parts: clinical and radiographic data. Clinical data included features of demographics [sex, age, body mass index (BMI), racial background, marital status, menopausal status, educational level, chronic diseases, occupation, smoking status and alcohol usage], medical history (comorbidities, quadriceps tendinitis, intra-articular fracture, Baker's cyst, ligament or meniscus damage, osteochondritis dissecans, plica syndrome and septic arthritis), symptoms [qualitative items of knee pain and stiffness, quantitative items measured with WOMAC total and subscale scores [21] and numeric rating scale (NRS) pain score] and physical examinations (knee warmth, bony tenderness, crepitus, range of motion, knee pain on extension and flexion). Radiographic data included standardized grades for tibial attrition (yes/no), medial/lateral joint space narrowing (0–3), femoral/tibial sclerosis (yes/no), and medial/lateral and tibial/femoral/patellar osteophytes (0–3), and the whole tibiofemoral joint according to the Kellgren and Lawrence (KL) grading system (0–4) [22]. The radiographic grades were obtained for each knee from independent reading by trained observers (blinded for clinical data) on weight-bearing posterior-anterior fixed flexion and lateral knee radiographs [19, 23].

Clinicians

We recruited clinicians who had a degree in general practice, orthopaedics, rheumatology or sports medicine for >2 years, or were in training in these specialties combined with a PhD in OA research. We then assessed their characteristics by querying on experience in OA treatment (years), number of OA patients treated per week, and personal perception on the importance of radiographs in making the OA diagnosis.

Diagnosis of clinically relevant knee OA

We stored clinical and radiographic data in special software (built in-house) for optimal presentation. Details of software training and diagnosis making process have been described in our previous studies [15, 24]. A brief description of the diagnosis-making process is given below.

Clinicians were divided into pairs; each pair consisted of one GP and one SP, and assessed the same subset of knees (from 40–50 patients). First, the software only presented T5 to T10 clinical data to the clinicians. Each clinician assessed these independently and, for each knee, chose between 'yes, clinically relevant OA has developed' and 'no, clinically relevant OA has not developed'. Next, the software activated T5 to T10 radiographic data access. Clinicians did the assessments independently again and answered the same questions. At this stage, clinicians had read-only access to the clinical data and their own previous diagnoses. Besides, the actual radiographic films were also available, and the software recorded the access to the films. After the procedure, each knee had four diagnoses: GPs' and SPs' diagnosis before and after viewing radiographic data.

Statistical analysis

We first built a nested 5-fold cross-validation (CV) enhanced machine learning pipeline to test candidate machine learning algorithms and select the algorithm with the best performance for further analysis. We totally tested eight candidate algorithms in a sub-dataset from our full dataset: two feature selectors [recursive feature elimination (RFE) and Relief] along with four classifiers [Logistic Regression, Random Forest (RF), Support Vector Classifier and Xgboost]. RFE-RF model was selected for further analysis as it turned out the highest area under receptor operating curve (AUC). The final machine learning pipeline is visualized in Fig. 1, with detailed descriptions of the pipeline in the figure legend.

In total, 48% of knees had incomplete data and 13% of features had >10% (up to 14%) missing values, which was mainly caused by loss to follow-up. To reduce the risk of overfitting, we impute missing values by simple imputation (using the mean value for continuous variable and the most frequent value for categorical variable) and incorporated it into the 5-fold CV. Besides, we restricted the maximum depth of the forest to 5, number of trees to 1000 and allowed the model to include 10 up to 50 variables.

We applied the pipeline on the data using GPs' diagnosis based on clinical data as the outcome factor, and included all

Downloaded from https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keac707/6936417 by Universiteit Leiden - LUMC user on 27 February 2023

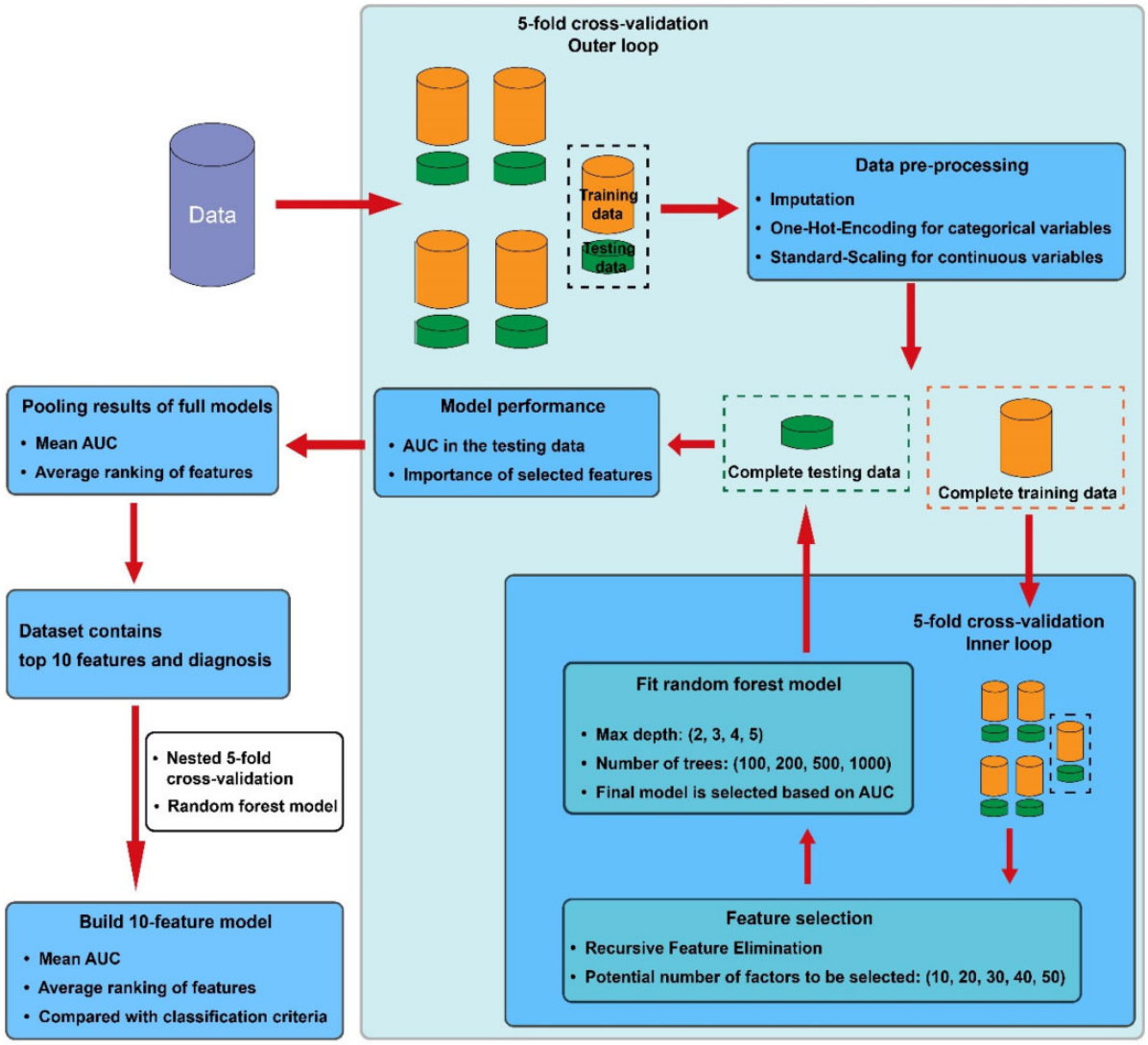

Figure 1. A nested 5-fold cross-validation (CV) enhanced machine learning pipeline. First, a full dataset goes into the outer loop of 5-fold CV, then it is randomly split into five equally sized subsets (folds); four of the folds are combined as training dataset while the remaining one fold is used as testing dataset. Next, training and testing datasets were pre-processed. Then the training dataset of the outer loop goes into the inner loop which consists of another 5-fold CV for developing and testing random forest models in combination with recursive feature elimination. The model tested with the highest AUC is selected to be output into the outer loop, and then model performance is assessed in the independent testing dataset. The 5-fold CV outer loop requires the procedure to repeat five times until every fold has been used as testing dataset. Therefore, five full models are developed, and mean AUC and average ranks of features were calculated among the five models. Finally, a dataset with containing the top 10 features and outcome measure enters the same pipeline to build 10-feature models

patient (T5, T8 and T10) clinical features as predictors. We then used GPs' diagnosis based on clinical and radiographic data as the outcome factor, and patient (T5, T8 and T10) clinical and radiographic features as predictors. We did the same for the diagnosis of SP. Finally, we obtained four bunches of full models (because of the nested 5-fold CV, each bunch consists of five models) for four diagnoses: $model_{GP}$, $model_{GP+radiographs}$, $model_{SP}$ and $model_{SP+radiographs}$. We calculated the mean AUC and its 95% CI for assessing model performance of each bunch. We ranked features included in the models by the Gini index and calculated the average rank of each feature (among the five models within a bunch) for each diagnosis. We pre-specified top 10 features as the highly-ranked ones related to the diagnosis, as we had allowed all the models to include at least 10 features. Next, we further developed models containing the top 10 features only (via the same machine learning pipeline) and calculated the corresponding AUC. To better evaluate 10-feature model's performance, we applied commonly used clinical criteria (EULAR, ACR and NICE criteria) in T10 follow-up data to identify 'clinical OA' knees (see detailed descriptions of these criteria in Supplementary Table S1, available at *Rheumatology* online). Using the clinician's diagnosis as the reference standard, we compared the AUC between 10-feature models and the three criteria by *Delong's* method [25]. Among the top 10 features, we further identified the top five features to highlight the most important ones.

To further examine the robustness of our findings regarding the involvement of missing values, we performed a sensitivity analysis with building the same machine models in the

Downloaded from https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keac707/6936417 by Universiteit Leiden - LUMC user on 27 February 2023

complete datasets (knees with missing values in any of the predictors were excluded). We then evaluated the top 10 features and 10-feature models' AUC.

For multiple comparisons between AUC, we adjusted *P*-values by the *Bonferroni's* method (multiplies the raw *P*-values by the number of tests), and a *P*-value <0.05 was considered statistically significant. All the analysis was performed in Python 3.6 (package scikit-learn, Numpy, pandas and seaborn) and R software 4.0 (package pROC).

Results

Patients and clinicians

This study included 716 patients with 1106 symptomatic knees; 79% female, mean (S.D.) age at T10 was 66 (5) years, mean (S.D.) BMI at T10 was 27 (4) kg/m^2. Clinical data contained 139 clinical features, radiographic data contained 36 radiographic features, see all the features and their descriptive characteristics in Supplementary Table S2, available at *Rheumatology* online.

A total of 17 GPs and 17 SPs were recruited to form 17 clinician pairs; among the SPs, seven were orthopaedists, eight rheumatologists and two sports physicians. Clinician characteristics are presented in Table 1. SPs averagely treated more OA patients per week and valued radiographs more than GPs.

Clinically relevant knee OA

GPs diagnosed clinically relevant OA in 42% and 43% knees, before and after viewing radiographic data, respectively. SPs in 43% and 51% of the knees. Both GPs and SPs somewhat modified their diagnoses after viewing radiographic data, while generally they agreed on 70% diagnoses regardless of whether radiographic data were available or not (Fig. 2). During the procedure, GPs viewed 45% of the actual radiographic knee films and SPs viewed 75%.

Machine learning models and model performance

All the RFE-RF full models contained 50 features and had good performance for explaining clinicians' diagnoses: model$_{GP}$, mean AUC of 0.87 (95% CI, 0.85, 0.89); model$_{GP+radiographs}$, 0.84 (95% CI, 0.82, 0.86); model$_{SP}$, 0.83 (95% CI, 0.80, 0.86); model$_{SP+radiographs}$, 0.79 (95% CI, 0.76, 0.82).

Models containing the top 10 features presented similarly good performance: 10-feature model$_{GP}$, mean AUC of 0.83 (95% CI, 0.85, 0.85); 10-feature model$_{GP+radiographs}$, 0.82 (95% CI, 0.80, 0.84); 10-feature model$_{SP}$, 0.77 (95% CI, 0.75, 0.79); 10-feature model$_{SP+radiographs}$, 0.76 (95% CI, 0.73, 0.78). Clinicians' diagnoses were poorly explained by the three commonly used clinical criteria (AUC ranged from 0.62–0.68 for GPs, and 0.58–0.65 for SPs). Mean AUC of the 10-feature models were all significantly higher than the three criteria (Fig. 3).

Table 1. Characteristics of recruited clinicians

	General practitioner ($n=17$)	Secondary care physician ($n=17$)
Experience of treating OA patients, years, mean (S.D.)	12 (9)	15 (9)
Number of OA patients treated per week, mean (S.D.)	5 (3)	27 (30)
Importance of radiographs[a], median (range)	2 (1–4)	4 (2–4)

[a] Perceived importance of radiography for making the diagnosis of knee OA: 1, not important; 2, minor important; 3, somewhat important; 4, very important.

Top features

Top 10 and top five features selected by the RFE-RF models are presented in Table 2. Before viewing radiographic data, patient symptom features, especially quantitative measures (WOMAC scores), were the most important ones related to the diagnosis of both GPs and SPs. Besides, none of the physical examination items was identified as important for GPs' diagnosis, while T10 joint line tenderness was for SPs' diagnosis.

After viewing radiographic data, the top five features for GPs' diagnosis remained the same, but were moderately changed for SPs' diagnosis with incorporating three radiographic features. Among the top 10 features, two medial compartment structural features (T10 medial joint space narrowing grade and T5 medial tibial osteophyte grade) were found related to GPs' diagnosis, while the grades for the whole tibia-femoral joint (T5 to T10 KL grade) and T10 medial femoral osteophyte were found related to SPs' diagnosis.

None of the demographic and medical history features were identified as important in any of the RFE-RF models.

Sensitivity analysis

The AUC of the models built in the complete datasets were found slightly lower (about 2%) than those obtained in our main analysis, while the top 10 features were generally similar (Supplementary Tables S3 and S4, available at *Rheumatology* online). Overall, the sensitivity analysis supports the robustness of our main findings.

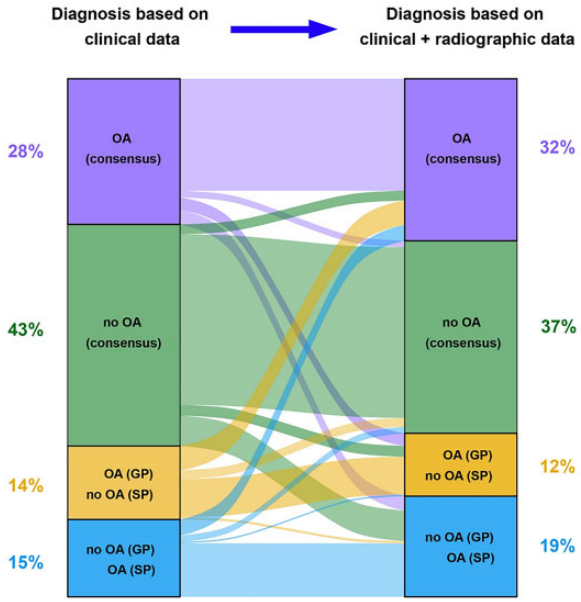

Figure 2. Clinicians' diagnosis before and after viewing radiographic data. 'Consensus' means GP and SP made the same diagnosis. Percentages indicates proportions of knees in each category and are calculated against total number of knees (1106). GP: general practitioner; SP: secondary care physician

Downloaded from https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keac707/6936417 by Universiteit Leiden - LUMC user on 27 February 2023

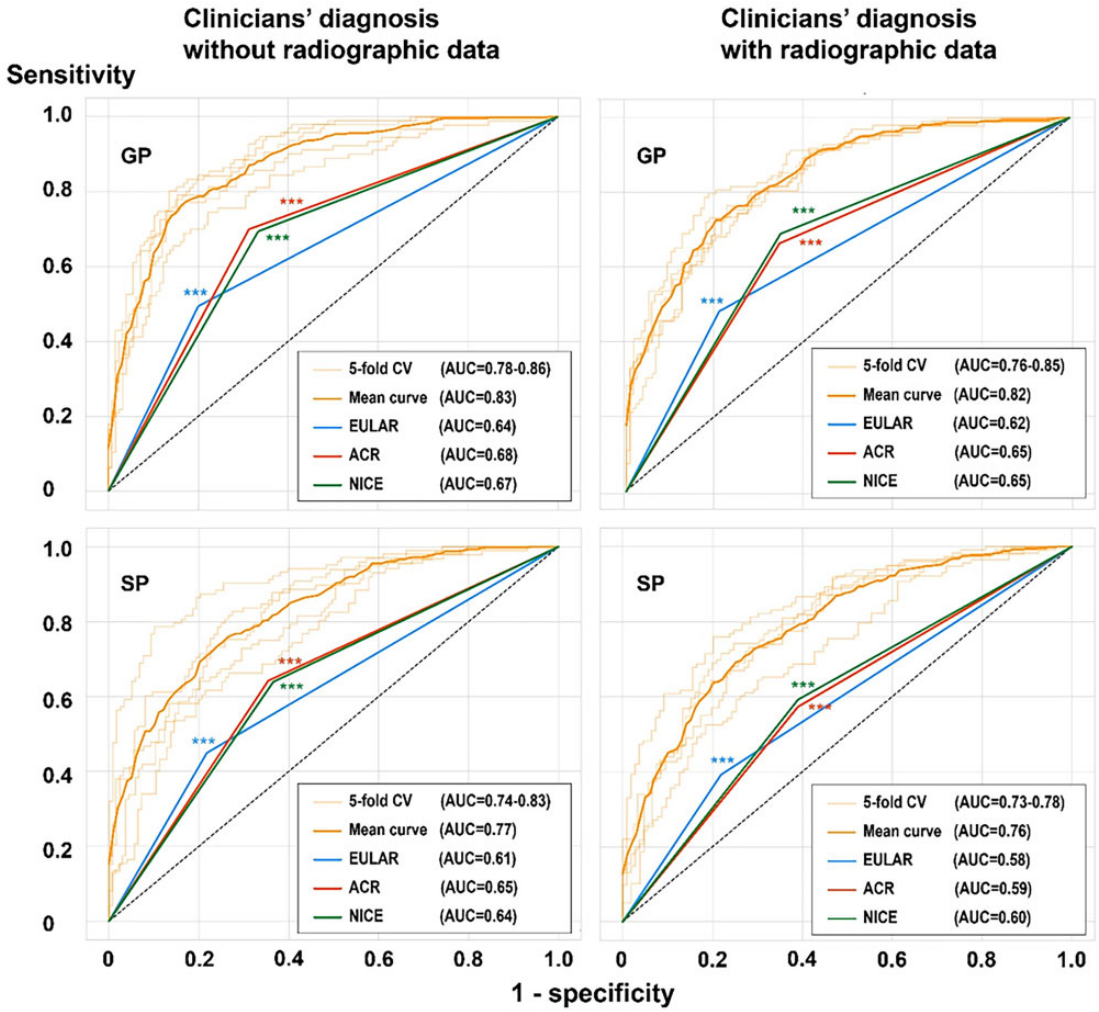

Figure 3. Receiver operating curves for top 10-feature models and clinical diagnostic/classification criteria against clinicians' diagnosis. Mean curve was generated based on the five curves of the 5-fold cross-validation (CV). AUC: area under the curve; NICE: National Institute for Health and Care Excellence. ***$P < 0.001$, comparing mean AUC of 10-feature models with EULAR, ACR and NICE criteria

Discussion

In this study, we developed RFE-RF models with good performance to explain GPs' and SPs' diagnosis of clinically relevant knee OA. The patient features identified by these models suggest typical characteristics of the patients who would likely receive a diagnosis of clinically relevant knee OA from clinicians.

Although GPs and SPs agreed on most diagnoses (70%), both before and after viewing radiographic data, variations existed in the patient features related to the diagnoses in different situations. When radiographs were unavailable, patients with more severe symptoms were more likely to receive the OA diagnosis from both GP and SP. Additionally, only SP seemed to have taken joint line tenderness examination into account, which might be one of the reasons for the 30% discrepancies in the diagnosis between the two kinds of clinicians. Another reason could be that the threshold of symptom severity for making an OA diagnosis was different between GP and SP, as suggested by a real-world report from Jordan *et al.* In that study, GPs tended not to diagnose knee OA in patients with mild symptoms only [12]. Unfortunately, there is no similar study available in secondary care.

When radiographs were available, radiographic features appeared in the top-10 lists for both GPs and SPs. Whereas, focusing on the top five features only, GPs seemed to still make diagnoses mainly based on patient symptoms; SPs shifted to a combination of symptomatic and structural features. This is consistent with actual clinical practice that SPs tend to check radiographic films for assessing the structural severity and then plan further treatments (e.g. orthopaedic surgeons assess the necessity/suitability of surgery based on radiographs), while GPs are advised to not obtain radiography for OA diagnoses [4] and are likely to provide symptom relief treatments.

Patient demographic and medical history features were found unimportant for diagnosis in the four 10-feature models. This is inconsistent with previous reports that patients with older age, more comorbidities or obesity were more likely to be diagnosed as knee OA by GPs [12, 14]. Besides, the EULAR, ACR and NICE criteria, which were developed based on the consensus of clinical and research experts, all treated age as an important indicator for identifying OA knees. A possible reason could be that the CHECK cohort recruited patients above 45 years at baseline [19], so all the

Downloaded from https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keac707/6936417 by Universiteit Leiden - LUMC user on 27 February 2023

Table 2. Top 10 and five (bold items) features related to clinicians' diagnosis on clinically relevant knee OA

	GP diagnosis based on clinical data	SP diagnosis based on clinical data	GP diagnosis based on clinical and radiographic data	SP diagnosis based on clinical and radiographic data
Demographics and medical history	None	None	None	None
Symptoms	**T10 WOMAC total score**	**T10 WOMAC total score**	**T10 WOMAC total score**	**T8 WOMAC total score**
	T10 WOMAC function score	**T10 WOMAC function score**	**T10 WOMAC function score**	**T5 WOMAC total score**
	T8 WOMAC total score	**T8 WOMAC total score**	**T8 WOMAC total score**	T8 WOMAC function score
	T8 WOMAC function score	**T8 WOMAC function score**	**T8 WOMAC function score**	T5 WOMAC function score
	T5 WOMAC total score	**T5 WOMAC total score**	**T5 WOMAC total score**	
	T10 WOMAC pain score	T10 WOMAC pain	T10 WOMAC stiffness score	
	T5 WOMAC function score	T5 WOMAC function score	T5 WOMAC function score	
	T5 WOMAC stiffness score	T5 WOMAC pain score		
	T10 knee stiffness-No			
	T10 knee stiffness-Yes			
Physical examination	None	T10 Joint line tenderness – Negative[a]	T10 knee flexion degree	T10 Joint line tenderness – Positive
		T10 Joint line tenderness – Positive[a]		T8 knee flexion degree
Radiographic features	–	–	T10 medial joint space narrowing grade	**T10 KL grade**
			T5 medial tibial osteophyte grade	**T10 medial femoral osteophyte grade**
				T8 KL grade
				T5 KL grade

[a] In the random forest model, joint line tenderness tested positive indicates the knee is more likely to have OA; negative indicates have no OA.
GP: general practitioner; SP: secondary care physician; T5 (8,10): 5 (8,10)-year follow-up; KL: Kellgren and Lawrence.

patients in this study (after 5 years follow-up) had already fulfilled the age requirement of the three criteria. On the other hand, our analysis focused on identifying the most important features, could have missed features with weak associations. Hence, the findings should be interpreted as clinicians rely more on symptoms, physical examinations or radiographic features than on risk factors (e.g. older age, comorbidities and higher BMI) in diagnosing knee OA.

The differences in the top features between GPs and SPs suggests that researchers using registry diagnoses to assess OA disease burden should be aware of the situation under which the diagnosis is made. For instance, diagnoses of GPs (with and without radiographs) mainly reflect patient symptoms, which could provide hints towards the demands for symptom relief management (e.g. pain medication) in primary care. Global disease burden studies defined OA patients by the combination of knee pain and KL grade ($\geq$2), which seems to reflect the disease burden similar to the perspective from SP [26].

In this study, none of the three commonly used clinical criteria adequately captured the knees recognized with clinically relevant knee OA by clinicians. It should be noted that ACR criteria were originally developed as classification criteria to be used in research, and EULAR and NICE criteria were developed for diagnosis in clinical settings. It might be 'unfair' to test the ACR criteria against clinicians' diagnosis, while our results suggest the diagnostic performance of ACR was similar to the NICE. Meanwhile, the knee OA definition by the NICE criteria is exactly the same as the Dutch healthcare practice guideline for GPs (https://richtlijnen.nhg.org/standaarden/niet-traumatische-knieklachten), which reveals an inconsistency between guidelines and GPs' actual clinical practice. This is in line with a previous report that indicated only moderate adherence to practice guidelines by clinicians [27]. On the other hand, the presented RFE-RF models performed well in the discrimination of the diagnosis by clinicians, which showed the feasibility of applying machine learning models in similar research problems. Meanwhile, it may also imply the feasibility of simulating human diagnosis through machine learning models.

The design of this study has several strengths. First, GPs and SPs were paired to (independently) review the same sample of knees, which made it robust to compare the diagnoses and features between the two kinds of clinicians. Second, clinicians' diagnoses were made on patient longitudinal data, which is more similar to the actual situation than on cross-sectional data; as suggested by a previous study, GPs would record patients with 'joint pain' at the initial consultations and then diagnose OA after 6–7 years follow-up [12]. Third, nested 5-fold CV was used to improve model stability and allowed all model performances to be evaluated in an independent testing dataset. Fourth, data imputation was incorporated into the 5-fold CV, meaning that the imputation was done on the random sample of the whole dataset for five times. Similar to the merits of the multiple imputation, this created more uncertainty in the imputed values and thus increased the standard error to obtain a better estimation of the correct value. Fifth, because patient features are likely to be inter-correlated, RFE, by iteratively training a model with removing the lowest ranking features, was used in combination

Downloaded from https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keac707/6936417 by Universiteit Leiden - LUMC user on 27 February 2023

with the RF model. Previous studies have demonstrated the robustness of developing the RFE-RF model in data with correlated features [17, 24, 28].

This study has limitations. First, despite the strengths regarding the internal validity, the findings should be treated cautiously when implemented externally. For example, clinicians were asked to diagnose clinically relevant OA; this could be different from the case where diagnosis of pre-clinical or early-stage OA is included. The CHECK cohort excluded the patients with potential differential diagnoses (e.g. rheumatoid arthritis) at baseline, so the findings could not apply to patients with these conditions. Moreover, only Dutch clinicians were recruited, which calls for future studies on evaluating the generalizability of our results in other regions. Second, we could have missed some radiographic features (e.g. tibiofemoral alignment) which were not listed in the dataset, but could have been captured by clinicians when viewing the actual radiographic films. Third, though RFE-RF models take feature interactions into account [16], trajectories of the features over the period (T5 to T10) might not have been well reflected in the analysis. Patients with worsened symptoms and structural progression might be more likely to be diagnosed with knee OA. While it had been shown that the majority of the knees had stable symptoms in the CHECK cohort from T5 to T10 [6], we assumed this issue is only relative in the minority of patients. For structural progression, we did an explorative analysis on testing correlation between KL progression (T10-T5 ≥ 1) and diagnoses (after viewing radiographic data); no statistically significant correlation was found for either GPs or SPs. In our final models, features from different time points were included, which should be interpreted as patients with consistent severe symptoms and structural damages were more likely to receive the diagnosis. Forth, missing values of the features were presented as 'blank box' to the clinicians but were imputed in our statistical analysis, which caused discrepancies between the two scenarios. We did the imputation because the RFE-RF model does not tolerate missing values. To reduce the risk of overfitting, we used simple imputation which may lead the data to be more similar and result in the increasement in the type I error (false-positive correlation). Whereas the sensitivity analysis validated the robustness of our main findings. We interpret this as missing values should have somewhat biased our results (e.g. AUC values), but it seemed the extent is not large enough to have influenced our main conclusions significantly. Fifth, information leakage could have occurred while selecting top five features, because they were selected from the top 10 features which were determined with access to the whole dataset.

In conclusion, RFE-RF models developed in this study had good performance in explaining clinicians' diagnosis of clinically relevant knee OA. Patients' (severity of) symptoms are the most important features related to the diagnosis of GPs, and of SPs when there is no access to radiographs. Although related to the diagnosis of both, radiographic features seem to be more important for SPs than GPs. The study findings helped to illustrate typical vignettes of patients recognized as clinically relevant knee OA by experts from two different care settings.

Supplementary material

Supplementary material is available at *Rheumatology* online.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Funding

This work was supported by the Dutch Arthritis Society (Project ID 15–1-301); Q.W. was financed by China Scholarship Council (CSC) (grant number: 201906230308).

Disclosure statement: J.R. and M.K. received research grants from the Dutch Arthritis Society; M.K. reports fee for consultancy (Abbvie, Pfizer, Levicept, GlaxoSmithKline, Merck-Serono, Kiniksa, Flexion, Galapagos, Jansen, CHDR, Novartis, UCB) and local investigator of industry-driven trial (Abbvie), from Wolters Kluwer (UptoDate), Springer Verlag (Reumatologie en klinische immunologie), board member for OARSI, president of the Dutch Society Rheumatology and member of the EULAR Council.

Acknowledgements

We would like to acknowledge the CREDO experts group (N.E. Aerts-Lankhorst, R. Agricola, A.N. Bastick, R.D.W. van Bentveld, P.J. van den Berg, J. Bijsterbosch, A. de Boer, M. Boers, A.M. Bohnen, A.E.R.C.H. Boonen, P.K Bos, T.A.E.J. Boymans, H.P. Breedveldt-Boer, R.W. Brouwer, J.W. Colaris, J. Damen, G. Elshout, P.J. Emans, W.T.M. Enthoven, E.J.M. Frölke, R. Glijsteen, H.J.C. van der Heide, A.M. Huisman, R.D. van Ingen, M.L. Jacobs, R.P.A. Janssen, P.M. Kevenaar, M.A. van Koningsbrugge, P. Krastman, N.O. Kuchuk, M.L.A. Landsmeer, W.F. Lems, H.M.J. van der Linden, R. van Linschoten, E.A.M. Mahler, B.L. van Meer, D.E. Meuffels, W.H. Noort-van der Laan, J.M. van Ochten, J. van Oldenrijk, G.H.J. Pols, T.M. Piscaer, J.B.M. Rijkels-Otters, N. Riyazi, J.M. Schellingerhout, H.J. Schers, B.W.V. Schouten, G.F. Snijders, W.E. van Spil, S.A.G. Stitzinger, J.J. Tolk, Y.D.M. van Trier, M. Vis, V.M.I Voorbrood, B.C. de Vos, and A. de Vries) for evaluating the medical files and their feedback on the manuscript.

References

1. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. Lancet 2019;393:1745–59.
2. Altman R, Asch E, Bloch D *et al.* Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. Arthritis Rheum 1986;29:1039–49.
3. Zhang W, Doherty M, Peat G *et al.* EULAR evidence-based recommendations for the diagnosis of knee osteoarthritis. Ann Rheum Dis 2010;69:483–9.
4. National Clinical Guideline Centre. Osteoarthritis: care and management in adults. UK: NICE, 2014.
5. Skou ST, Koes BW, Gronne DT, Young J, Roos EM. Comparison of three sets of clinical classification criteria for knee osteoarthritis: a cross-sectional study of 13,459 patients treated in primary care. Osteoarthritis Cartilage 2020;28:167–72.
6. Schiphof D, Runhaar J, Waarsing JH *et al.* The clinical and radiographic course of early knee and hip osteoarthritis over 10 years in CHECK (Cohort Hip and Cohort Knee). Osteoarthritis Cartilage 2019;27:1491–500.
7. Kluzek S, Rubin KH, Sanchez-Santos M *et al.* Accelerated osteoarthritis in women with polycystic ovary syndrome: a prospective

Downloaded from https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keac707/6936417 by Universiteit Leiden - LUMC user on 27 February 2023

nationwide registry-based cohort study. Arthritis Res Ther 2021; 23:225.
8. Misra D, Lu N, Felson D *et al.* Does knee replacement surgery for osteoarthritis improve survival? The jury is still out. Ann Rheum Dis 2017;76:140–6.
9. Yu D, Jordan KP, Snell KIE *et al.* Development and validation of prediction models to estimate risk of primary total hip and knee replacements using data from the UK: two prospective open cohorts using the UK Clinical Practice Research Datalink. Ann Rheum Dis 2019;78:91–9.
10. Fraenkel L, Buta E, Suter L *et al.* Nonsteroidal anti-inflammatory drugs vs cognitive behavioral therapy for arthritis pain: a randomized withdrawal trial. JAMA Intern Med 2020;180:1194–202.
11. Mol MF, Runhaar J, Bos PK *et al.* Effectiveness of intramuscular gluteal glucocorticoid injection versus intra-articular glucocorticoid injection in knee osteoarthritis: design of a multicenter randomized, 24 weeks comparative parallel-group trial. BMC Musculoskelet Disord 2020;21:225.
12. Jordan KP, Tan V, Edwards JJ *et al.* Influences on the decision to use an osteoarthritis diagnosis in primary care: a cohort study with linked survey and electronic health record data. Osteoarthritis Cartilage 2016;24:786–93.
13. Turkiewicz A, Petersson IF, Bjork J *et al.* Current and future impact of osteoarthritis on health care: a population-based study with projections to year 2032. Osteoarthritis Cartilage 2014;22:1826–32.
14. Arslan IG, Damen J, de Wilde M *et al.* Incidence and prevalence of knee osteoarthritis using codified and narrative data from electronic health records: a population-based study. Arthritis Care Res 2022;74:937–44.
15. Wang Q, Runhaar J, Kloppenburg M *et al.* The added value of radiographs in diagnosing knee osteoarthritis is similar for general practitioners and secondary care physicians; data from the CHECK early osteoarthritis cohort. J Clin Med 2020;9:3374.
16. Jamshidi A, Pelletier JP, Martel-Pelletier J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. Nat Rev Rheumatol 2019;15:49–60.
17. Lazzarini N, Runhaar J, Bay-Jensen AC *et al.* A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. Osteoarthritis Cartilage 2017;25:2014–21.
18. Runhaar J, Kloppenburg M, Boers M *et al.* Towards developing diagnostic criteria for early knee osteoarthritis: data from the CHECK study. Rheumatology 2021;60:2448–55.
19. Wesseling J, Boers M, Viergever MA *et al.* Cohort Profile: cohort hip and cohort knee (CHECK) study. Int J Epidemiol 2016;45: 36–44.
20. Norgeot B, Quer G, Beaulieu-Jones BK *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320–4.
21. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833–40.
22. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16:494–502.
23. Macri EM, Runhaar J, Damen J, Oei EH, Bierma-Zeinstra SM. Kellgren/Lawrence grading in cohort studies: methodological update and implications illustrated using data from the CHECK cohort. Arthritis Care Res 2022;74:1179–87.
24. Yang R, Zhang C, Gao R, Zhang L. A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data. Int J Mol Sci 2016;17:218.
25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
26. Safiri S, Kolahi AA, Cross M *et al.* Prevalence, deaths, and disability-adjusted life years due to musculoskeletal disorders for 195 countries and territories 1990-2017. Arthritis Rheumatol 2021;73:702–14.
27. DeHaan MN, Guzman J, Bayley MT, Bell MJ. Knee osteoarthritis clinical practice guidelines – how are we doing? J Rheumatol 2007; 34:2099–105.
28. Jiang H, Deng Y, Chen HS *et al.* Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics 2004;5:81.

Downloaded from https://academic.oup.com/rheumatology/advance-article/doi/10.1093/rheumatology/keac707/6936417 by Universiteit Leiden - LUMC user on 27 February 2023