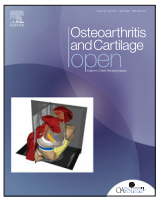
Osteoarthritis and Cartilage Open 6 (2024) 100506

Contents lists available at ScienceDirect

Osteoarthritis and Cartilage Open

journal homepage: www.elsevier.com/journals/osteoarthritis-and-cartilage-open/2665-9131



Diagnostic models to predict structural spinal osteoarthritis on lumbar radiographs in older adults with back pain: Development and internal validation

Mirna Chamoro [a,*], Martijn W. Heymans [b], Edwin H.G. Oei [c], Sita M.A. Bierma-Zeinstra [a,d], Bart W. Koes [a,e], Alessandro Chiarotto [a]

[a] Department of General Practice, Erasmus MC, University Medical Center, Rotterdam, the Netherlands
[b] Department of Epidemiology and Biostatistics, Amsterdam University Medical Center, Amsterdam, the Netherlands
[c] Department of Radiology & Nuclear Medicine, Erasmus MC, University Medical Center, Rotterdam, the Netherlands
[d] Department of Orthopedics and Sports Medicine, Erasmus MC, University Medical Center, Rotterdam, the Netherlands
[e] Research Unit of General Practice, Department of Public Health & Center for Muscle and Joint Health, University of Southern Denmark, Odense, Denmark

ARTICLE INFO

Handling Editor: H Madry

Keywords:
Clinical diagnostic prediction model
Spinal OA

ABSTRACT

Objective: It is difficult for health care providers to diagnose structural spinal osteoarthritis (OA), because current guidelines recommend against imaging in patients with back pain. Therefore, the aim of this study was to develop and internally validate multivariable diagnostic prediction models based on a set of clinical and demographic features to be used for the diagnosis of structural spinal OA on lumbar radiographs in older patients with back pain.

Design: Three diagnostic prediction models, for structural spinal OA on lumbar radiographs (i.e. multilevel osteophytes, multilevel disc space narrowing (DSN), and both combined), were developed and internally validated in the 'Back Complaints in Older Adults' (BACE) cohort (N = 669). Model performance (i.e. overall performance, discrimination and calibration) and clinical utility (i.e. decision curve analysis) were assessed. Internal validation was performed by bootstrapping.

Results: Mean age of the cohort was 66.9 years (±7.6 years) and 59% were female. All three models included age, gender, back pain duration and duration of spinal morning stiffness as predictors. The combined model additionally included restricted lateral flexion and spinal morning stiffness severity, and exhibited the best model performance (optimism adjusted c-statistic 0.661; good calibration with intercept −0.030 and slope of 0.886) and acceptable clinical utility. The other models showed suboptimal discrimination, good calibration and acceptable decision curves.

Conclusion: All three models for structural spinal OA displayed lesuboptimal discrimination and need improvement. However, these internally validated models have potential to inform primary care clinicians about a patient with risk of having structural spinal OA on lumbar radiographs. External validation before implementation in clinical care is recommended.

1. Introduction

Low back pain (LBP) is a very common condition with a large burden on society and health care systems [1]. Over 619 million people (95% CI [554–694 million]) are affected worldwide and this number is expected to increase further to 843 million of prevalent cases by 2050 [2]. The majority of patients (~80–90%) with LBP are labelled as having nonspecific LBP, usually diagnosed by excluding specific underlying conditions, such as spinal fractures or malignancies [3]. However, within the large group of patients with nonspecific LBP, there may be distinct diagnostic subgroups of patients with similar characteristics, that may benefit from a separate, more accurate, diagnosis, and have a different

* Corresponding author. Department of General Practice, Erasmus MC, University Medical Center, PO Box 2040, 3000 CA Rotterdam, the Netherlands.
E-mail addresses: m.chamoro@erasmusmc.nl (M. Chamoro), mw.heymans@amsterdamumc.nl (M.W. Heymans), e.oei@erasmusmc.nl (E.H.G. Oei), s.bierma-zeinstra@erasmusmc.nl (S.M.A. Bierma-Zeinstra), b.koes@erasmusmc.nl (B.W. Koes), a.chiarotto@erasmusmc.nl (A. Chiarotto).

https://doi.org/10.1016/j.ocarto.2024.100506
Received 17 July 2024; Accepted 21 July 2024

prognostic course as compared to the majority of patients with nonspecific LBP. One of these diagnostic subgroups may be represented by patients with symptomatic spinal osteoarthritis (OA) [4,5].

Until now, there are no agreed diagnostic criteria for spinal OA, making it difficult to diagnose and to further study this condition. It is unclear when and how nonspecific LBP may be reflective of symptomatic spinal OA. Nevertheless, some research has been undertaken to reach consensus on definitions of spinal OA. This includes a Delphi study, seeking consensus among experts on statements regarding symptomatic and structural spinal OA [6]. Consensus was reached on spinal pain duration, spinal pain intensity, self-reported spinal morning stiffness and back-related physical function limitations as clinical features to be considered in a symptomatic definition of spinal OA. Subsequently, a systematic review looked into the associations between the clinical features on which consensus was reached and imaging findings suggestive of structural spinal OA [7]. In that systematic review high-quality evidence (GRADE-approach) was found for various associations, with small to moderate magnitudes. These included an association between LBP intensity and disc space narrowing (on lumbar spine radiographs), and the duration of self-reported spinal morning stiffness and disc space narrowing on lumbar radiographs. However, most associations between clinical features and structural findings of spinal OA were based on very low- to moderate-quality evidence, meaning that further high-quality research is needed on these associations [7].

Current clinical practice guidelines recommend to not offer routine imaging to patients with nonspecific LBP [8], meaning that clinicians in daily practice cannot rely on imaging to diagnose spinal OA. One manner for clinicians to identify patients with the presence of structural spinal OA without using imaging is by using valid multivariable diagnostic prediction models that helps them to identify patients with spinal complaints at risk of having spinal osteoarthritis. A multivariable diagnostic prediction model aims to provide an individual with a risk of presence of a disease [9]. This helps to inform patients and health care professionals, aids medical decision making and improves health outcomes. In the field of spinal pain, there is already a series of diagnostic prediction models developed to aid early detection of disc degeneration [10], to enable automatic evaluation of disc degeneration on imaging [11], or to predict progression of (lumbar) disc degeneration on MRI [12]. However, none of these models attempted to identify symptomatic spinal OA on lumbar radiographs, taking into account expert-agreed clinical features, such as spinal morning stiffness or back pain- related physical function limitation, without the use of imaging. Therefore, our objective is to explore which of the aforementioned clinical features (i.e. back pain intensity, back pain duration, self-reported spinal morning stiffness, back pain related physical function limitation and limited or painful range of motion) as well as some demographic features (i.e. age, sex, body mass index, number of comorbidities and education level) are strongly related to spinal OA structural features on radiographs such as multilevel disc space narrowing and the presence of multilevel osteophytes. Hence, the aim of this study was to develop and internally validate multivariable diagnostic prediction models based on a set of clinical and demographic features to be used for the diagnosis of structural spinal OA on lumbar radiographs in older patients with back pain. For this study we chose to focus on older adults with back pain, since this is also the population who has a high probability of having spinal OA and there is limited research looking into this group of patients.

2. Methods

A research protocol was drawn up in advance and registered in Open Science Framework (https://osf.io/uc7sm/?view_only=ce85c93a91d648749bb938beb696a103). The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement reporting guidelines were used for the reporting of this study [13], see Appendix A.

2.1. Development sample

For the development and internal validation of the diagnostic prediction models, we used the baseline data from the "Back Complaints in Older Adults" (BACE) cohort, a prospective cohort which recruited participants from 2009 up to 2011 in the Netherlands [14,15]. Patients over 55 years of age who consulted their general practitioner (GP) with a new episode of back complaints were included. Patients should not have had a previous episode of back pain in the preceding six months. The exclusion criteria were: patients who were unable to fill in the questionnaires as a result of language problems or a cognitive disorder, and patients unable to undergo a physical examination (e.g. wheelchair-bound patients). Patient characteristics, demographic and back complaint features, as well as information on comorbidities and physical function limitations, were collected using questionnaires, and participants underwent physical examination and lumbar spine radiographs at baseline. More information on the BACE cohort and the characteristics of the included patients is provided elsewhere [15]. Ethical approval for the BACE cohort was received from the Medical Ethics Committee of the Erasmus Medical Center in the Netherlands (NL24829.078.08).

2.2. Potential predictors

A selection of the most relevant candidate predictors, based on prior literature and clinical expertise [6,7,15], was carried out for the development of the diagnostic prediction models. The following candidate predictors were evaluated through questionnaires: duration of the current back pain episode in days, back pain intensity over the last week measured with a 0–10 numeric rating scale (NRS; with 0 representing no pain at all, and 10 the worst pain ever) [16], back-related physical function limitation assessed with the 24-item Roland-Morris Disability Questionnaire (RMDQ, where higher scores represent higher levels of back pain-related disability) [17], self-reported duration of spinal morning stiffness assessed using a 3-point Likert scale (i.e. none, $\leq$30 min, >30min), self-reported spinal morning stiffness severity assessed with a 5-point Likert scale (i.e. none, mild, moderate, severe, extreme); additionally, the following demographic characteristics were assessed: age (in years), sex (assessed as male or female), body mass index (BMI, kg/m^2), number of comorbidities (more or less than three, including cardiovascular disease, musculoskeletal disease, depression, malignancy, gout or neurological disease), and education level (high, middle or low; based on the International Standard Classification of Education (ISCED) classification). The following included candidate predictors were assessed during physical examination by a trained research assistant: restricted lateral flexion (i.e. inability to bend sideways using the fingertips and reach further than the knee while standing), restricted rotation (asymmetry in rotation while sitting), restricted lumbar flexion (measured as the finger-to-floor distance in centimeters) [18].

2.3. Outcomes of the diagnostic prediction models

The outcome predicted by the model was the presence or absence of structural spinal OA on lateral lumbar spine radiographs. Since there are no widely accepted criteria or a clinical reference standard to define structural spinal OA, we used the definitions previously described by Van den Berg et al. [18,19]. Structural spinal OA was defined as the presence of multilevel disc space narrowing and/or osteophytes on lumbar radiographs. The presence of osteophytes and disc space narrowing in the lumbar region was scored using the grading system as described in the Lane atlas, with 0 = none, grade 1 = mild, grade 2 = moderate and grade 3 = severe lumbar disc degeneration [20]. Multilevel disc space narrowing was defined as grade $\geq$1 narrowing at 2 or more levels between L1-2 and L5-S1, and multilevel osteophytes as grade $\geq$2 at 2 or more levels [18]. The radiographs in the BACE cohort were assessed by two independent observers, who were trained by an experienced

musculoskeletal radiologist. They were blinded for the patient characteristics. The interrater agreement was good for both osteophytes identification (kappa = 0.65) and disc space narrowing (kappa = 0.70) [18,19].

Three separate models were developed, one predicting the risk for multilevel osteophytes on lumbar radiographs, one for multilevel disc space narrowing, and a combined model, predicting the risk for the combination of multilevel disc space narrowing and osteophytes on lumbar radiographs (as defined above). For the development of all three models the same candidate predictors were used.

2.4. Statistical analysis

In line with methodological guidance to develop diagnostic prediction models, multivariable logistic regression analyses were used and among the a-priori selected literature and expert based predictors the most relevant were selected by backward selection, based on a p-value of <0.05 [21]. Potential non-linearity of continuous predictors (i.e. pain duration, pain intensity, BMI, age and lumbar flexion) was evaluated using restricted cubic spline functions. Multicollinearity was checked with the use of a correlation matrix, where a value of 0.80 or higher indicated a strong correlation between predictor variables. We also calculated the Variance Inflation Factor (VIF), with a cut-off value of 5 [22]. A sample size calculation was performed as described by Riley et al. [23] for the development of prediction models (shrinkage 0.75, c-statistic 0.70). From this calculation, we estimated that a maximum of 27 candidate parameters could be included in the models. To ascertain the best-fitted and most stable models, the models were internally validated by use of bootstrapping, a resampling method that helps to reduce overfitting and improve model accuracy. One hundred bootstrap samples were created for the internal validation procedure. Missing data was evaluated in relation to observed data and as a result multiple imputation on predictors as well as outcomes was performed for the primary analyses [24]. Fifty imputed datasets were generated by use of the Multivariate Imputation by Chained Equations (MICE) method [25]. Regression coefficients, standard errors and model performance estimates were pooled using Rubin's Rules. Complete case analysis was performed as a sensitivity analysis.

2.5. Model performance

A diagnostic prediction model should be able to distinguish diseased from non-diseased individuals correctly (discrimination), and, at the same time, it should produce predicted probabilities that are in line with the actual outcome probabilities (calibration) [26]. Discrimination was expressed as the C-statistic, comparable to the area under the receiver-operating curve for a logistic model. The C-statistic represents the chance that in two individuals, one with and one without the outcome, the predicted outcome probability will be higher for the individual with the outcome compared with the one without. A C-statistic of ≥ 0.70 was considered satisfactory [27,28]. To assess the calibrating potential of the models, we performed calibration-in-the-large and calibration plots were constructed to visually display the agreement between the predicted outcomes of the models and the observations in the data [27]. Calibration was considered good if the slope-value was close to one and intercept-value close to zero, and if the calibration plot visually showed good calibration. Overall performance was expressed by the Nagelkerke's R and the Brier score [27]. Clinical decision curve analysis in complete cases was carried out to decide upon the use of the most suitable model in practice by assessing the net benefit of the model at different threshold probabilities for intervention [29], such as education and lifestyle changes.

All analysis were conducted in SPSS version 26 and Rstudio (pmsampsize [30], mice [25], rms [31], dcurves [32], and psfmi packages [33]).

2.6. Deviations from the protocol

Some changes were made from the registered protocol. We stated in the protocol that we would develop a combined model, looking at multilevel osteophytes *and/or* multilevel disc space narrowing. However, we adjusted this and we developed a combined model which predicted the risk for multilevel osteophytes *and* disc space narrowing combined. This choice was made, because on hindsight, it seemed more clinically relevant to see what the added value of a combined model was, compared to a separate model predicting the risk for multilevel osteophytes and a separate diagnostic prediction model for multilevel disc space narrowing.

3. Results

3.1. Model development

The BACE cohort included 669 older patients with back pain; mean age was 66.9 years (±7.6 years) and 59% were female. Other population characteristics are shown in Table 1.

Table 1
Baseline characteristics in the Back Complaints in Older Adults (BACE) cohort (N = 669).

General characteristics and demographics		Missing (N, %)
Age (years), mean ± SD	67.0 ± 7.7	0 (0.0)
Sex, female N (%)	394 (59)	0 (0.0)
BMI, mean ± SD	27.5 ± 4.7	6 (0.9)
Education level, N (%)		1 (0.2)
Low	464 (68.6)	
Middle	90 (13.3)	
High	114 (16.9)	
Number of comorbidities, N (%)		9 (1.3)
≤3 comorbidities	296 (44.8)	
>3 comorbidities	394 (55.2)	
Clinical features		
Duration of current back pain episode (days), mean ± SD	248 ± 974	74 (11)
Back pain intensity (NRS, 0–10), mean ± SD	4.6 ± 2.6	2 (0.3)
Duration of spinal morning stiffness, N (%)		8 (1.2)
No morning stiffness	163 (24.7)	
≤30 min	338 (51.1)	
>30 min	160 (24.2)	
Severity of spinal morning stiffness, N (%)		1 (0.2)
None	92 (13.8)	
Mild	155 (23.2)	
Moderate	229 (34.3)	
Severe	147 (22)	
Extreme	45 (6.7)	
Physical functioning (RMDQ, 0–24), mean ± SD	9.9 ± 5.8	
Restricted spinal lateral flexion, N (%)		9 (1.3)
No	205 (31.1)	
Yes	455 (68.9)	
Restricted spinal rotation, N (%)		6 (0.9)
No	507 (76.5)	
Yes	156 (23.5)	
Lumbar flexion (FFD in cm), mean ± SD	10.9 ± 11.9	14 (2.1
Pain during spinal lateral flexion	418 (63.3)	9 (1.3)
Pain during spinal rotation	257 (38.7)	6 (0.9)
Pain during lumbar flexion	272 (42.2)	23 (3.7)
Outcomes		
Multilevel osteophytes, N (%)	258 (40.9)	38 (5.7)
Multilevel disc space narrowing, N (%)	453 (71.7)	37 (5.5)
Combined (multilevel osteophytes & disc space narrowing), N (%)	226 (35.8)	38 (5.7)

NRS: Numeric Rating Scale; RMDQ: Roland-Morris Disability Questionnaire; FFD: finger-to-floor distance.

3.2. Missing values

Several baseline characteristics had more than 5% missing values and a few up to 11% (Table 1). The outcome variables all displayed approximately 6% missing values (Table 1). We chose to use 50 imputed datasets and convergence plots of all imputed variables were constructed. The convergence plots of all imputed variables showed no irregular patterns, indicating healthy convergence and no multicollinearity between the variables (appendix B) [24].

3.3. Developed models

The combined model predicted the risk of having the outcomes multilevel osteophytes and multilevel disc space narrowing based on the following predictors: age, gender, back pain duration, restricted spinal lateral flexion, duration of spinal morning stiffness and severity of spinal morning stiffness (Table 2). The results show that, according to this model, the higher the age (OR 1.05), being female (OR 1.79), having longer back pain duration (OR 1.001), having restricted spinal lateral flexion (OR 1.60), having (self-reported) spinal morning stiffness for longer than 30 min (OR 1.43) and having moderate (self-reported) spinal morning stiffness severity (OR 2.41), the higher an individual's risk is for the combined presence multilevel osteophytes and disc space narrowing on lumbar radiographs.

The osteophytes model predicted the risk of multilevel osteophytes on lumbar radiographs using age, gender, back pain duration and duration of spinal morning stiffness as predictors. The multilevel disc space narrowing model contained the same predictors as the multilevel osteophytes model. Full information of the multivariable diagnostic prediction models predicting multilevel osteophytes, and multilevel disc space narrowing separately, are shown in Table 2.

3.4. Model performance

The model performance after internal validation of all three models is shown in Table 2 and Fig. 1 (calibration). Overall, model performance was similar for all three models. However, the combined model had the best discriminative value (AUC of 0.682, 95% CI 0.637–0.723, optimism-

Table 2
Multivariable diagnostic prediction models for structural spinal osteoarthritis outcomes in older adults with back pain (BACE cohort, N = 669).

Predictors	Coefficients	OR (95% CI)	R^2 *optimism-adjusted	Brier Score *optimism-adjusted	C-statistic (95% CI)	Optimism-adjusted C-statistic
Combined model*			0.125	0.092	0.682 (0.637–0.723)	0.661
Intercept	−4.505		*0.095	*0.066		
Age	0.044	1.046 (1.022–1.069)				
Gender	0.580	1.785 (1.253–2.545)				
Back pain duration	0.001	1.001 (1.000–1.003)				
Restricted lateral flexion	0.468	1.596 (1.064–2.394)				
Spinal morning stiffness ≤30min	−0.470	0.625 (0.350–1.115)				
Spinal morning stiffness >30min	0.357	1.429 (0.715–2.855)				
Spinal morning stiffness severity - mild	0.398	1.489 (0.755–2.935)				
Spinal morning stiffness severity - moderate	0.881	2.413 (1.137–5.121)				
Spinal morning stiffness severity - severe	0.489	1.630 (0.715–3.716)				
Spinal morning stiffness severity - extreme	0.022	1.022 (0.358–2.912)				
Osteophytes model*			0.105	0.089	0.663 (0.619–0.704)	0.651
Intercept	−3.860		*0.086	*0.062		
Age	0.046	1.047 (1.024–1.070)				
Gender	0.757	2.131 (1.524–2.981)				
Back pain duration	0.001	1.001 (1.000–1.003)				
Spinal morning stiffness ≤30min	−0.059	0.942 (0.630–1.410)				
Spinal morning stiffness >30min	0.658	1.931 (1.192–3.129)				
Disc space narrowing model*			0.104	0.078	0.662 (0.618–0.703)	0.652
Intercept	−4.592		*0.087	*0.063		
Age	0.045	1.046 (1.023–1.069)				
Gender	0.763	2.145 (1.529–3.009)				
Back pain duration	0.001	1.001 (1.000–1.003)				
Spinal morning stiffness ≤30min	−0.046	0.955 (0.638–1.431)				
Spinal morning stiffness >30min	0.664	1.942 (1.202–3.139)				

Back pain duration of current episode in days; age in years; gender: male/female; restricted lateral flexion: inability to bend sideways using the fingertips and reach further than the knee while standing.

*Penalized combined model (with adjusted coefficients): $-4.446 + 0.038 \times$ age $+ 0.499 \times$ gender $+ 0.001 \times$ back pain duration $+ 0.403 \times$ restricted lateral flexion $+ -0.405 \times$ spinal morning stiffness duration ≤30 min $+ 0.307 \times$ spinal morning stiffness duration >30 min $+ 0.343 \times$ spinal morning stiffness severity mild $+ 0.758 \times$ spinal morning stiffness severity moderate $+ 0.421 \times$ spinal morning stiffness severity severe $+ 0.019 \times$ spinal morning stiffness severity extreme.

*Penalized osteophytes model (with adjusted coefficients): $-4.388 + 0.041 \times$ age $+ 0.761 \times$ gender $+ 0.001 \times$ back pain duration $+ -0.021 \times$ spinal morning stiffness duration ≤30 min $+ 0.607 \times$ spinal morning stiffness duration >30 min.

*Penalized disc space narrowing model (with adjusted coefficients): $-4.318 + 0.042 \times$ age $+ 0.714 \times$ gender $+ 0.001 \times$ back pain duration $+ -0.043 \times$ spinal morning stiffness duration ≤30 min $+ 0.621 \times$ spinal morning stiffness duration >30 min.

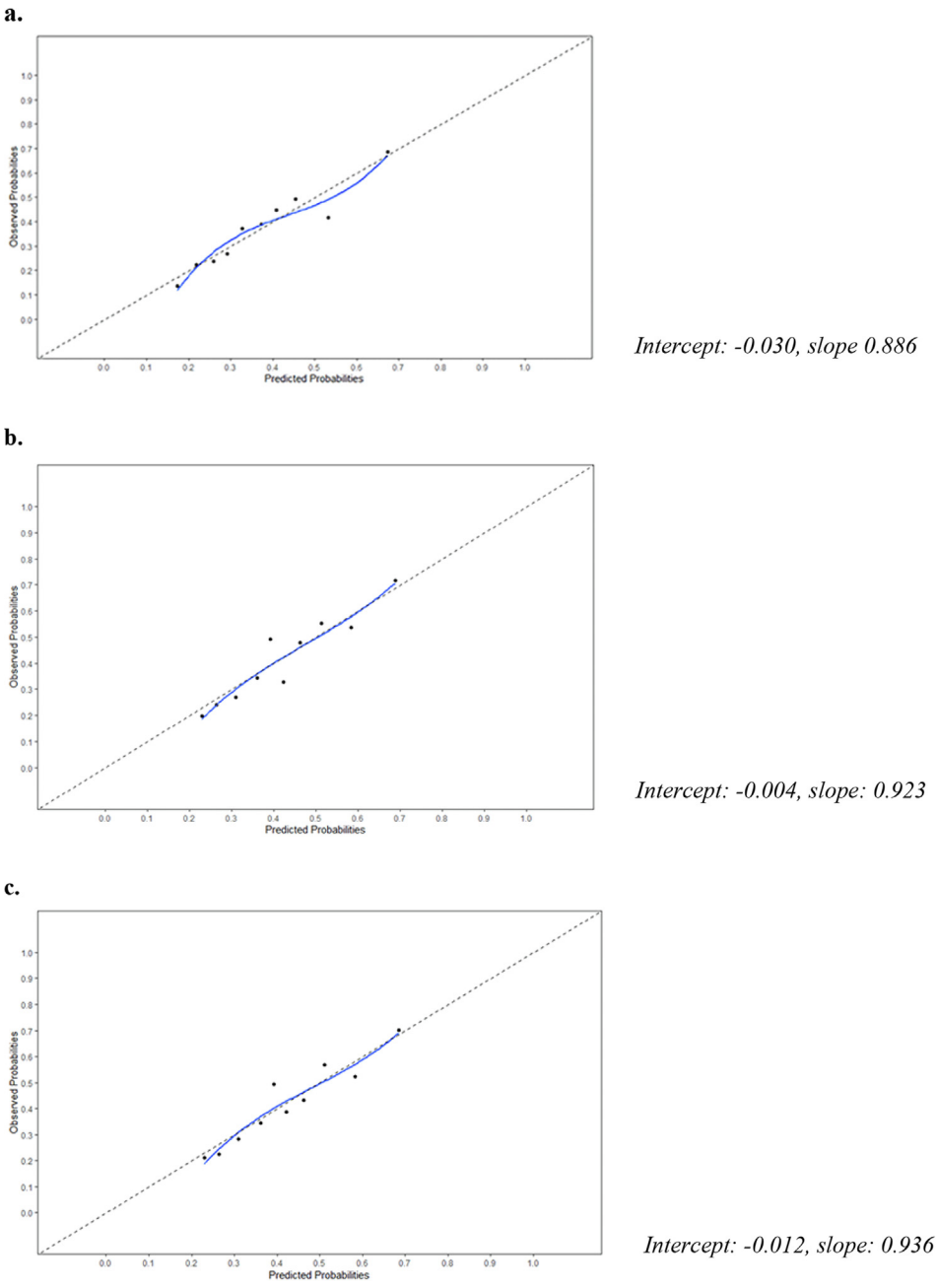


Fig. 1. Calibration curves of the developed models. a. Combined model (multilevel osteophytes and dis space narrowing), b. Osteophytes model (multilevel osteophytes), c. Disc space narrowing model (multilevel disc space narrowing). Calibration curves with intercept (0 is perfect) and calibration slope (1 is perfect).

adjusted: 0.661) and a good calibration (slope of 0.886). The multilevel osteophytes model and multilevel disc space narrowing model had poorer discriminative values (optimism-adjusted AUC of 0.651 and 0.652, respectively), but showed good calibration (Fig. 1).

3.5. Decision curve analyses

Decision curves for all three models are shown in Fig. 2. The combined model showed a gain in net benefit when that model was used, compared to giving all patients an intervention, from a threshold probability of approximately 15% and above. Similar results were obtained for the osteophytes model, but the disc space narrowing model showed additional advantage from a threshold probability of approximately 50%.

3.6. Sensitivity analyses

Sensitivity analyses on complete cases (multilevel osteophytes, multilevel disc space narrowing and combined models, n = 536, n = 537, and n = 536; respectively) showed comparable performance measure values (see appendix B). The multilevel osteophytes and the combined models derived in complete case data yielded the same or almost the same predictors. In addition, the multilevel disc space narrowing model

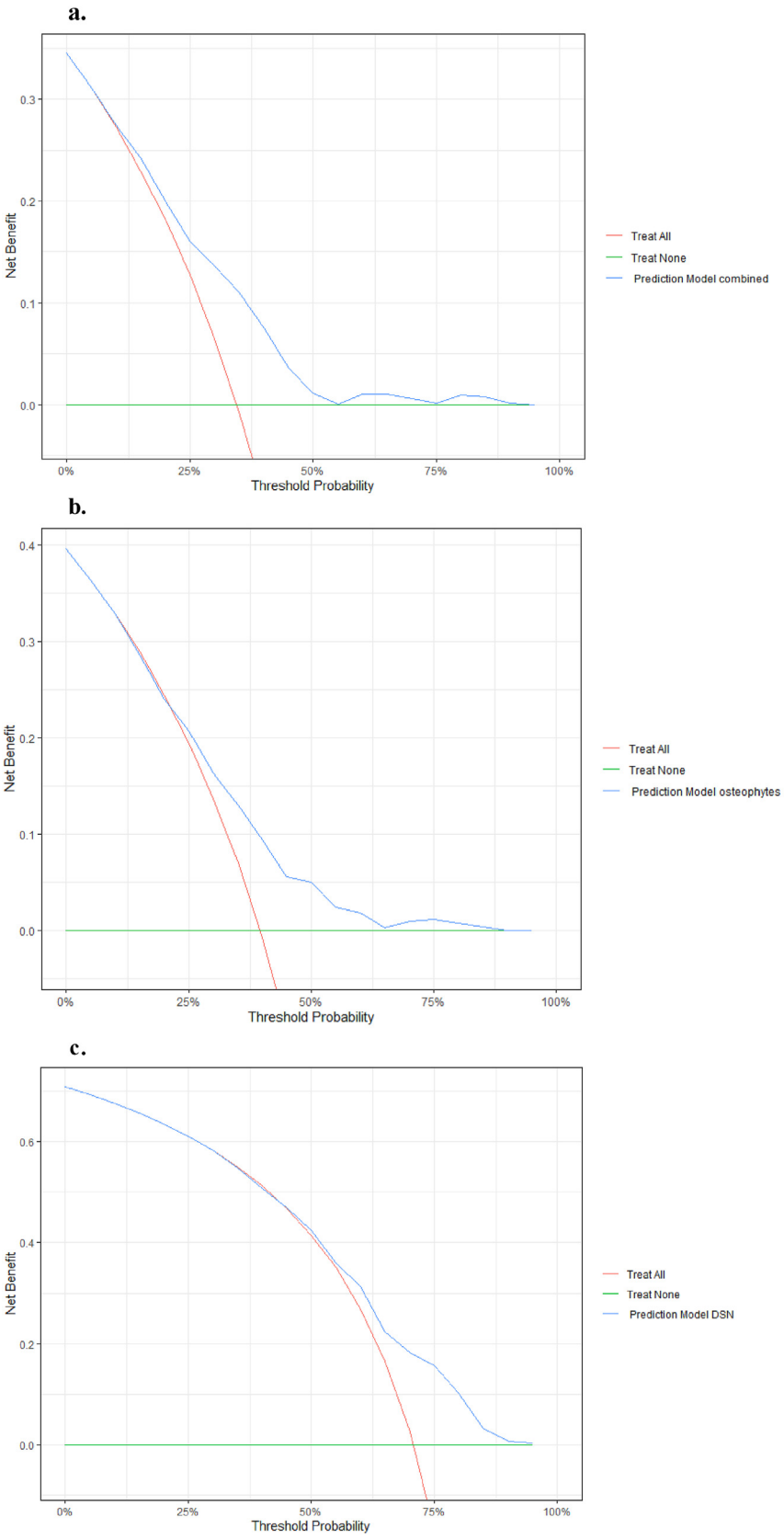


Fig. 2. Decision curves of the developed models. a. Combined model (multilevel osteophytes and disc space narrowing), b. Osteophytes model (multilevel osteophytes), c. Disc space narrowing model (multilevel disc space narrowing).

in the complete cases contained additionally severity of spinal morning stiffness and restricted lateral flexion as predictors.

4. Discussion

This study developed and internally validated multivariable diagnostic prediction models for the presence of structural spinal OA on lumbar radiographs. The combined model (i.e. multilevel osteophytes and disc space narrowing) included age, gender, back pain duration, restricted spinal lateral flexion and spinal morning stiffness duration and severity as predictors, exhibiting the best overall performance, discrimination and calibration. However, the discriminative ability of this model showed to be less than satisfactory (i.e. c-statistic <0.7), after adjusting for optimism; calibration was good. The other two models displayed the same trend, where calibration was good, but discrimination did not exceed the threshold for acceptable discriminative ability. A possible explanation is that the chosen candidate predictors may not be explanatory enough of structural spinal OA, leading to poor model performance of models containing these predictors and we have to look into other (sets of) characteristics. These results indicate that these models need further improvement, before external validation and implementation in clinical practice.

4.1. Important results predictors and comparison to existing literature

A surprising outcome of this study is that, according to all developed diagnostic prediction models, having spinal morning duration for longer than 30 min gives an individual a higher risk of structural spinal OA, i.e. having multilevel osteophytes and/or disc space narrowing, whereas spinal morning stiffness shorter than 30 min gives a lower risk of structural spinal OA, compared to having no spinal morning stiffness. This finding has been reported before in a study by van den Berg et al., looking into the association between self-reported spinal morning stiffness and lumbar disc degeneration in a different primary care cohort [34]. Prolonged spinal morning stiffness (>30 min) is typically associated with inflammatory spondylarthropathies, such as axial spondyloarthritis [35], whereas spinal morning stiffness duration of 30 min or less is usually associated with OA [36–38].

A possible explanation for the fact we found that prolonged spinal morning stiffness was associated with a higher risk of structural spinal OA could be that spinal OA might have a longer or larger inflammatory component than what has been assumed thus far. Another explanation could lie in that the data on spinal morning stiffness duration in our study was self-reported by participants and derived through a questionnaire, which could have led to some degree of detection bias. However, there are differences in the duration of morning stiffness related to OA when looking at peripheral joints. For knee OA the criterion for morning stiffness duration is indeed 30 min or less [39], whereas for hip OA this is 60 min or less [40] and for hand OA no duration of (morning) stiffness is specified [41]. It is possible that the mere presence of morning stiffness is more explanatory of OA than the duration of it, which could also be the case in our study.

4.2. Model comparison to existing models in the literature

In existing literature there are many prediction models for LBP [42–44], and for structural spinal OA [10–12,45]. Most of these models predict the risk or prognosis for LBP outcomes, taking several demographic, clinical and imaging features into account, or aim to improve the diagnosis of structural spinal OA (on imaging) by enabling automatic evaluation of disc degeneration on imaging. A promising new approach is the use of deep machine learning algorithms to predict the risk of spinal OA using patients' medical information [10], but this is still in the early phases. Clinical diagnostic prediction models predicting the risk for structural spinal OA based on demographic and clinical features are scarce. To our knowledge, there are no diagnostic prediction models for structural spinal OA, that include the selected set of candidate predictors (e.g. back pain duration, back pain intensity, duration and severity of self-reported spinal morning stiffness, back pain related physical disability and restricted range of motion) and the outcomes evaluated in this study.

In comparison, prediction models for structural knee OA-features, such as joint space narrowing and osteophytes, have shown similar results. Ramazanian et al. identified 26 prediction models for knee OA, with mostly median samples sizes of less than one thousand. They report poor to moderate model performance of the identified prediction models, with an AUC range between 0.6 and 0.9 and of 0.6–0.8 when externally validated [46]. Only 11 of the 26 included studies reported the calibration, mostly with a Hosmer and Lemeshow test (range p-values 0.19–0.90) [46]. Most of these models included similar demographic and clinical predictors as our models, such as age, sex, BMI, pain and morning stiffness.

4.3. Implications for practice and further research

Further research looking into other clinical features to be considered as spinal OA predictors, e.g. biomarkers or genetic markers, is recommended. Also, more research on the reliability and uniform measurement of some of the candidate predictors, such as range of motion assessment, or self-reported morning stiffness evaluation, is needed. This information would benefit the field and could lead to more accurate (future) prediction models, and ultimately aid health care providers in the care for patients with spinal OA. Furthermore, different (combinations of) definitions of structural spinal OA (e.g. endplate changes) on various modalities (e.g. MRI) should be considered and further investigated to see which definition is most suitable and accurate. Lastly, external validation and updating of the developed models in different datasets is required, before considering their use in clinical practice.

4.4. Strengths and limitations

This study has several strengths. Firstly, the predictors in the derived models are easy to assess in clinical practice and the models can aid health care professionals to identify patients with spinal complaints at risk of having spinal osteoarthritis and, hopefully, improve clinical management and health outcomes of these patients. Secondly, the TRIPOD statement guidelines were followed and the study protocol was made openly available prior to conduction of the study. Thirdly, the results of this study can be used a stepping stone in the process of developing diagnostic criteria for spinal OA.

However, there are also some limitations. Firstly, there are limitations to how the included predictors are measured, due to varying definitions and measuring instruments [47]. For example, there is no data on the reliability of some of the candidate predictors, such as range of motion assessment, or self-reported morning stiffness evaluation. Self-reported severity or duration of clinical features, derived through questionnaires, could have led to (some degree of) detection bias. Nevertheless, the measurement tools used in this study are appropriate for use in busy clinical settings, such as general practice in the Netherlands. Secondly, the outcome definitions that we used can be a limitation. We specified the outcome predicted by the models to be multilevel osteophytes and/or disc space narrowing on lateral lumbar radiographs, as there is a lack of definition of structural spinal OA, due to lack of accepted criteria or clinical reference standards to define structural spinal OA. Even though these are widely used definitions of structural spinal OA [18,48], one could differ that other degenerative features might be more declarative, such as the presence of endplate changes or other structural features. Furthermore, it needs to be mentioned that spinal OA imaging features are also prevalent in patients without LBP, and the patients' clinical condition needs to be taken into account when interpreting these features [49]. Thirdly, there is no methodology developed to pool decision curves across multiply imputed datasets and therefore we applied the developed

models in the observed data for the decision curve analysis. However, we think that the results will be close to the imputed data, as the results of the sensitivity analyses we performed in the observed data for the other performance measures did not largely differ from the multiply imputed data.

5. Conclusions

All three internally validated diagnostic prediction models for structural spinal OA on lumbar radiographs displayed suboptimal discrimination and need improvement, with the combined model (i.e. multilevel osteophytes and disc space narrowing) exhibiting the best model performance, compared to the other two models (i.e. multilevel osteophytes, and multilevel disc space narrowing, separately). Although these models need improvement, the models do have the potential to inform and aid primary care clinicians about a patient's risk of having structural spinal OA. However, external validation and updating is required before the models can be implemented in clinical care.

Author contributions

All authors conceived and designed the study. The analysis was conducted by MC and checked by MH. AC, BK and SBZ supervised (and when necessary corrected) the assessment.

MC, MH, BK, SBZ and AC discussed the results and interpretation. MC drafted the manuscript and carried out the first round of revisions. All authors have read and approved the final version of the manuscript and contributed to its critical revision. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding

Individual funding was received by M. Chamoro from the Stichting Beroepsopleiding Huisartsen (SBOH). The SBOH played no role in the design, conduct or reporting of this study.

The BACE-Dutch cohort was funded by the department of General Practice, Erasmus MC, Rotterdam, Netherlands and the Coolsingel foundation, Rotterdam, The Netherlands.

Conflicts of interest

There were no competing interests of any of the authors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ocarto.2024.100506.

References

[1] J. Hartvigsen, M.J. Hancock, A. Kongsted, Q. Louw, M.L. Ferreira, S. Genevay, et al., What low back pain is and why we need to pay attention, Lancet 391 (2018) 2356–2367.
[2] Collaborators GBDLBP, Global, regional, and national burden of low back pain, 1990-2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021, Lancet Rheumatol. 5 (2023) e316–e329.
[3] A. Chiarotto, B.W. Koes, Nonspecific low back pain, N. Engl. J. Med. 386 (2022) 1732–1740.
[4] A.P. Goode, T.S. Carey, J.M. Jordan, Low back pain and lumbar spine osteoarthritis: how are they related? Curr. Rheumatol. Rep. 15 (2013) 305.
[5] N. Fine, S. Lively, C.A. Séguin, A.V. Perruccio, M. Kapoor, R. Rampersaud, Intervertebral disc degeneration and osteoarthritis: a common molecular disease spectrum, Nat. Rev. Rheumatol. 19 (2023) 136–152.
[6] K. de Luca, A. Chiarotto, F. Cicuttini, L. Creemers, E. de Schepper, P.H. Ferreira, et al., Consensus for statements regarding a definition for spinal osteoarthritis for use in research and clinical practice: a Delphi study, Arthritis Care Res. 75 (2023) 1095–1103.
[7] M. Chamoro, K. de Luca, O. Ozbulut, E.H.G. Oei, C.L.A. Vleggeert-Lankamp, B.W. Koes, et al., Association between clinical findings and the presence of lumbar spine osteoarthritis imaging features: a systematic review, Osteoarthr. Cartilage 31 (2023) 1158–1175.
[8] C.B. Oliveira, C.G. Maher, R.Z. Pinto, A.C. Traeger, C.C. Lin, J.F. Chenot, et al., Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview, Eur. Spine J. 27 (2018) 2791–2803.
[9] M. van Smeden, J.B. Reitsma, R.D. Riley, G.S. Collins, K.G. Moons, Clinical prediction models: diagnosis versus prognosis, J. Clin. Epidemiol. 132 (2021) 142–145.
[10] J.R.,S. Bradley, Developing predictive models for early detection of intervertebral disc degeneration risk, Healthcare Analy. 2 (2022).
[11] F. Niemeyer, F. Galbusera, Y. Tao, A. Kienle, M. Beer, H.J. Wilke, A deep learning model for the accurate and reliable classification of disc degeneration based on MRI data, Invest. Radiol. 56 (2021) 78–85.
[12] J.P.Y. Cheung, X. Kuang, M.K.L. Lai, K.M. Cheung, J. Karppinen, D. Samartzis, et al., Learning-based fully automated prediction of lumbar disc degeneration progression with specified clinical parameters and preliminary validation, Eur. Spine J. 31 (2022) 1960–1968.
[13] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, Bmj 350 (2015) g7594.
[14] J. Scheele, P.A. Luijsterburg, M.L. Ferreira, C.G. Maher, L. Pereira, W.C. Peul, et al., Back complaints in the elders (BACE); design of cohort studies in primary care: an international consortium, BMC Musculoskelet Disord 12 (2011) 193.
[15] J. Scheele, W.T. Enthoven, S.M. Bierma-Zeinstra, W.C. Peul, M.W. van Tulder, A.M. Bohnen, et al., Characteristics of older patients with back pain in general practice: BACE cohort study, Eur. J. Pain 18 (2014) 279–287.
[16] A. Chiarotto, L.J. Maxwell, R.W. Ostelo, M. Boers, P. Tugwell, C.B. Terwee, Measurement properties of visual analogue scale, numeric rating scale, and pain severity subscale of the brief pain inventory in patients with low back pain: a systematic review, J. Pain 20 (2019) 245–263.
[17] R. Smeets, A. Köke, C.W. Lin, M. Ferreira, C. Demoulin, Measures of function in low back pain/disorders: low back pain rating scale (LBPRS), oswestry disability index (ODI), progressive isoinertial lifting evaluation (PILE), quebec back pain disability scale (QBPDS), and roland-morris disability questionnaire (RDQ), Arthritis Care Res. 63 (Suppl 11) (2011) S158–S173.
[18] R. van den Berg, L.M. Jongbloed, N.O. Kuchuk, L.D. Roorda, J.C.M. Oostveen, B.W. Koes, et al., The association between self-reported low back pain and radiographic lumbar disc degeneration of the cohort hip and cohort knee (CHECK) study, Spine (Phila Pa 1976) 42 (2017) 1464–1471.
[19] R. van den Berg, A. Chiarotto, W.T. Enthoven, E. de Schepper, E.H.G. Oei, B.W. Koes, S.M.A. Bierma-Zeinstra, Clinical and radiographic features of spinal osteoarthritis predict long-term persistence and severity of back pain in older adults, Ann Phys Rehabil Med 65 (2022) 101427.
[20] N.E. Lane, M.C. Nevitt, H.K. Genant, M.C. Hochberg, Reliability of new indices of radiographic osteoarthritis of the hand and hip and lumbar disc degeneration, J. Rheumatol. 20 (1993) 1911–1918.
[21] E.W. Steyerberg, Y. Vergouwe, Towards better clinical prediction models: seven steps for development and an ABCD for validation, Eur. Heart J. 35 (2014) 1925–1931.
[22] F.E. Harrell Jr., Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis, Springer International Publishing, 2015.
[23] R.D. Riley, J. Ensor, K.I.E. Snell, F.E. Harrell Jr., G.P. Martin, J.B. Reitsma, et al., Calculating the sample size required for developing a clinical prediction model, Bmj 368 (2020) m441.
[24] Y. Vergouwe, P. Royston, K.G. Moons, D.G. Altman, Development and validation of a prediction model with missing predictor data: a practical approach, J. Clin. Epidemiol. 63 (2010) 205–214.
[25] S.G.-O.,K. van Buuren, Mice: multivariate imputation by chained Equations in R, J. Stat. Software 45 (2011) 1–67.
[26] J.M. Hendriksen, G.J. Geersing, K.G. Moons, J.A. de Groot, Diagnostic and prognostic prediction models, J Thromb Haemost 11 (Suppl 1) (2013) 129–141.
[27] E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, et al., Assessing the performance of prediction models: a framework for traditional and novel measures, Epidemiology 21 (2010) 128–138.
[28] Jr Hosmer, S. Dwl, R.X. Sturdivant, Applied Logistic Regression, 3rd Edition, Wiley, 2013.
[29] A.J. Vickers, E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, Med. Decis. Making 26 (2006) 565–574.
[30] J. Ensor, Pmsampsize: Sample Size for Development of a Prediction Model, 2024, The Comprehensive R Archive Network, 2023.
[31] F.E. Harrell Jr., Package 'rms', 2024, The Comprehensive R Archive Network, 2023.
[32] D.D. Sjoberg, Dcurves: Decision Curve Analysis for Model Evaluation, 2024, The Comprehensive R Archive Network, 2022.
[33] M.W. Heymans, psfmi: Prediction Model Pooling, Selection and Performance Evaluation across Multiply Imputed Datasets, 2024, The Comprehensive R Archive Network, 2023.
[34] R. van den Berg, E.M. Jongbloed, N.O. Kuchuk, B.W. Koes, E.H.G. Oei, S.M.A. Bierma-Zeinstra, P.A.J. Luijsterburg, Association between self-reported spinal morning stiffness and radiographic evidence of lumbar disk degeneration in participants of the cohort hip and cohort knee (CHECK) study, Phys. Ther. 100 (2020) 255–267.
[35] J. Sieper, D. van der Heijde, R. Landewé, J. Brandt, R. Burgos-Vagas, E. Collantes-Estevez, et al., New criteria for inflammatory back pain in patients with chronic back pain: a real patient exercise by experts from the Assessment of SpondyloArthritis international Society (ASAS), Ann. Rheum. Dis. 68 (2009) 784–788.

[36] Q. Wang, J. Runhaar, M. Kloppenburg, M. Boers, J.W.J. Bijlsma, S.M.A. Bierma-Zeinstra, et al., Evaluation of the giagnostic performance of American College of Rheumatology, EULAR, and National Institute for Health and Clinical Excellence criteria against clinically relevant knee osteoarthritis: data from the CHECK Cohort, Arthritis Care Res (Hoboken) 76 (4) (2024 Apr) 511–516, https://doi.org/10.1002/acr.25270. Epub 2024 Jan 23. PMID: 37933434.
[37] W. Zhang, M. Doherty, G. Peat, M.A. Bierma-Zeinstra, N.K. Arden, B. Bresnihan, et al., EULAR evidence-based recommendations for the diagnosis of knee osteoarthritis, Ann. Rheum. Dis. 69 (2010) 483–489.
[38] G. Wood, J. Neilson, E. Cottrell, S.P. Hoole, C. Guideline, Osteoarthritis in people over 16: diagnosis and management-updated summary of NICE guidance, Bmj 380 (2023) 24.
[39] R. Altman, E. Asch, D. Bloch, G. Bole, D. Borenstein, K. Brandt, 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. 29 (1986) 1039–1049.
[40] R. Altman, G. Alarcón, D. Appelrouth, D. Bloch, D. Borenstein, K. Brandt, et al., The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip, Arthritis Rheum. 34 (1991) 505–514.
[41] R. Altman, G. Alarcón, D. Appelrouth, D. Bloch, D. Borenstein, K. Brandt, et al., The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hand, Arthritis Rheum. 33 (1990) 1601–1610.
[42] R. Haskins, P.G. Osmotherly, D.A. Rivett, Diagnostic clinical prediction rules for specific subtypes of low back pain: a systematic review, J. Orthop. Sports Phys. Ther. 45 (61–76) (2015) A61–A64.
[43] T.R. Stanton, M.J. Hancock, C.G. Maher, B.W. Koes, Critical appraisal of clinical prediction rules that aim to optimize treatment selection for musculoskeletal conditions, Phys. Ther. 90 (2010) 843–854.
[44] F.G. Silva, L.O. Costa, M.J. Hancock, G.A. Palomo, L.C. Costa, T. da Silva, No prognostic model for people with recent-onset low back pain has yet been demonstrated to be suitable for use in clinical practice: a systematic review, J. Physiother. 68 (2022) 99–109.
[45] F. Li, X. Sun, Y. Wang, L. Gao, J. Shi, K. Sun, Development and validation of a novel nomogram to predict the risk of intervertebral disc degeneration, Med. Inflamm. 2022 (2022) 3665934.
[46] T. Ramazanian, S. Fu, S. Sohn, M.J. Taunton, H.M. Kremers, Prediction models for knee osteoarthritis: review of current models and future directions, Arch Bone Jt Surg. 11 (2023) 1–11.
[47] R.W. Wingbermühle, A. Chiarotto, B. Koes, M.W. Heymans, E. van Trijffel, Challenges and solutions in prognostic prediction models in spinal disorders, J. Clin. Epidemiol. 132 (2021) 125–130.
[48] E.I. de Schepper, J. Damen, J.B. van Meurs, A.Z. Ginai, M. Popham, A. Hofman, et al., The association between lumbar disc degeneration and low back pain: the influence of age, gender, and individual radiographic features, Spine (Phila Pa 1976) 35 (2010) 531–536.
[49] W. Brinjikji, P.H. Luetmer, B. Comstock, B.W. Bresnahan, L.E. Chen, R.A. Deyo, et al., Systematic literature review of imaging features of spinal degeneration in asymptomatic populations, AJNR Am J Neuroradiol 36 (2015) 811–816.