Diagnosis for early stage knee osteoarthritis: probability stratification, internal and external validation; data from the CHECK and OAI cohorts

Wang, Q.K.; Runhaar, J.; Kloppenburg, M.; Boers, M.; Bijlsma, J.W.J.; Bierma-Zeinstra, S.M.A.; CREDO Expert Grp

Seminars in Arthritis and Rheumatism 55 (2022) 152007

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

Diagnosis for early stage knee osteoarthritis: probability stratification, internal and external validation; data from the CHECK and OAI cohorts

Qiuke Wang[a,*], Jos Runhaar[a,*], Margreet Kloppenburg[b], Maarten Boers[c], Johannes W.J. Bijlsma[d], Sita M.A. Bierma-Zeinstra[a,e], the CREDO expert group

[a] Department of General Practice, Erasmus MC University Center Rotterdam, Rotterdam, the Netherlands
[b] Department of Rheumatology, Leiden University Medical Center, Leiden, the Netherlands
[c] Department of Epidemiology & Biostatistics, Amsterdam UMC, Amsterdam, the Netherlands.
[d] Department of Rheumatology and Clinical Immunology, University Medical Centre Utrecht, Utrecht, the Netherlands
[e] Department of Orthopaedics & Sports Medicine, Erasmus MC University Center Rotterdam, Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Knee osteoarthritis
Early diagnosis
Internal validation
External validation

ABSTRACT

Objective: To internally and externally validate our diagnostic criteria of early stage knee osteoarthritis (OA) in the CHECK and OAI cohorts.
Design: We applied two previously developed diagnostic models to all knees in CHECK and OAI cohorts to calculate probabilities of early stage knee OA at baseline. Knees were categorized into three groups based on probability: 'no OA' (probability $\leq$ 30%), 'uncertain' (probability between 30% and 70%) and 'early stage OA' (probability $\geq$ 70%). To validate the diagnosis, we obtained OA related outcome measures at 10-year follow-up in the CHECK cohort, and at 8-9-year follow-up in the OAI cohort. We compared outcome measures between 'no OA' and 'early stage OA' knees, and between 'no OA' and 'uncertain' knees using generalized estimating equations.
Results: In CHECK (n = 1042 knees) both models showed 'early stage OA' knees presented with significant and clinically relevant higher WOMAC scores, higher Kellgren & Lawrence (KL) grade, and higher rates of joint space narrowing (JSN) progression after 10 years, compared to 'no OA' knees. In OAI (n = 2937 knees) both models showed 'early stage OA' knees presented with significant and clinically relevant higher WOMAC scores, higher KL grade, and higher rates of KL and JSN progression after 8-9 years, compared to 'no OA' knees. Smaller, but still significant differences between 'uncertain' and 'no OA' knees were observed in both cohorts.
Conclusions: These results support internal and external validity of the two sets of diagnostic criteria for early stage knee OA.

Introduction

Knee osteoarthritis (OA), a chronic disease characterized by long-term symptomatic and structural progression [1], is usually diagnosed in the advanced stage of disease. Interventions have limited effect in altering or delaying disease progression at this stage [2,3]. It may be more effective to intervene in an early disease stage – this has been demonstrated in other chronic diseases like lung cancer [4], chronic heart failure [5] and rheumatoid arthritis [6]. However, diagnostic criteria for early stage knee OA are not yet established, making it difficult to determine who may benefit from early intervention.

In a previous study, we developed diagnostic criteria for early stage knee OA using clinical expert diagnosis of clinically relevant knee OA (5 to 10 years after first presentation in primary care) as a reference standard [7]. These criteria can be used to quantify the probability of early stage knee OA in patients with initial presentations in primary care. In contrast to 'classification criteria' [8–10], diagnostic criteria should facilitate treatment decisions in the clinic; thus, a probability threshold is needed to rule early stage OA in or out [11–14]. This would allow earlier and more selective application of potentially effective interventions, such as education [1], physical therapy [15] and weight loss [16,17].

Further validation of our criteria and probability thresholds are warranted before implementation in clinical practice. An early diagnosis is valid when it predicts important clinical outcomes. This validation should be carried out within the dataset in which the criteria

* Corresponding authors.
E-mail addresses: q.wang@erasmusmc.nl (Q. Wang), j.runhaar@erasmusmc.nl (J. Runhaar).

https://doi.org/10.1016/j.semarthrit.2022.152007

were developed (internal validation), but in order to evaluate generalizability it must also be performed in another independent sample (external validation) [11,17].

In this study, we explored the appropriate probability thresholds for our early diagnostic models, and then internally and externally validated these using probability stratification. The primary aim of this study was to evaluate whether the knees diagnosed as 'early stage' knee OA at baseline had more severe symptoms and structural damage than 'no OA' knees after 10 years in the CHECK cohort (internal validation); and whether similar results were observed in the Osteoarthritis Initiative (OAI) cohort with 9 years follow-up (external validation).

Methods

CHECK cohort

We obtained data from the CHECK cohort (a longitudinal cohort study of patients with knee or hip complaints, suspected for early stage OA in Dutch primary care, followed for 10 years) for the internal validation. The inclusion criteria of CHECK were (1) non-traumatic knee or hip pain or stiffness, (2) aged 45−65 years old, (3) no previous consultation with a general practitioner, or the first consultation within 6 months before inclusion. Patients were excluded if the complaints could be explained by other diseases. Details are published elsewhere [18]. In the present study, we included CHECK participants who had knee complaints at baseline and had data available at 10 years follow-up (T10), and excluded knees lost to follow-up at T10. Included and excluded knees were compared to determine whether there were any apparent differences (>10%) in baseline characteristics.

OAI cohort

We used the data from OAI incident subcohort for external validation. The OAI incident subcohort is a multi-center, longitudinal, observational cohort study in the USA, which recruited participants at risk of developing knee OA and followed these for at least 9 years. Patients were recruited at university research units or tertiary hospitals. The age-specific inclusion and exclusion criteria are available on the website (https://oai.nih.gov). For the present study, we selected eligible participants' knees within the same age range as the CHECK cohort (45 to 65 years) and same Kellgren & Lawrence (KL) grade range at baseline (KL≤2). Knees were excluded if they were lost to follow-up at 9 years (T9) or had missing data in the baseline measures which our models required. Included and excluded knees were compared for apparent differences (>10%) in baseline characteristics.

Early diagnostic models and probability thresholds

We previously developed two multivariable models for the diagnosis of early stage knee OA in the CHECK cohort, using experts' diagnosis of clinically relevant knee OA as a reference standard [7]. We obtained the expert diagnoses by recruiting both general practitioners and secondary care physicians, who independently evaluated the longitudinal clinical and radiographic data acquired at the 5, 8 and 10-year follow-up visits of all knees [7,19]. All the experts were unaware of patient baseline characteristics, and the final diagnosis (OA, no OA or uncertain) was made upon agreement between experts [7]. Model 1 of the diagnostic criteria included baseline clinical factors only, and model 2 included both clinical and radiographic factors at baseline (see equations in **Fig. 1** and descriptions of the model development in the **supplementary material**) [7].

All selected knees in both cohorts were divided into three groups, based on the probability for early stage knee OA obtained from the two models: a 'no OA' group (with a low probability for early stage knee OA), 'uncertain' group (medium probability) and 'early stage OA' group (high probability). Based on the literature and clinical applicapability [11,12,20], we used two potential thresholds, including 30%/70% (knees with a probability below 30% were diagnosed as 'no OA', and knees with probability higher than 70% were diagnosed as 'early stage OA') and 40%/60%. Larger ranges of thresholds, e.g. 20%/80%, would leave too many knees as 'uncertain' to be clinically meaningful. Moreover, the choices were supported by clinical experts' diagnosis that knees with medium probabilities (30%-70%) were more likely to be diagnosed as 'uncertain' by the experts than those with higher or lower probabilities (**supplementary Fig. 1**).

Baseline measures

In both CHECK and OAI, we extracted all required data from baseline measures for each participant/knee, including age, sex, body mass index (BMI), physical examinations (joint line tenderness, joint effusion, crepitus, bony swelling), Western Ontario and McMaster

(Model 1) Logit = −6.934 + 0.030*(Age) + 0.583*(Female) + 0.113*(BMI) +
0.524*(WOMAC function descending positive) +
0.414*(WOMAC function rising positive) +
0.462*(WOMAC morning stiffness positive) +
0.515*(Joint line tenderness test positive) +
1.427*(Effusion test positive) + 0.440*(Crepitus test positive)

(Model 2) Logit = −5.230 + 0.572*(Female) + 0.109*(BMI) +
0.420*(WOMAC function descending positive) +
0.440*(WOMAC function rising positive) +
0.479*(WOMAC morning stiffness positive) +
0.544*(Joint line tenderness test positive) +
1.257*(Effusion test positive) + 0.363*(Crepitus test positive) +
0.720*(Bony swelling test positive) +
1.208*(Patellofemoral joint space narrowing positive) +
1.263*(Medial joint space narrowing positive)

Probability of early stage knee OA = 1/(1+e^(-logit))*100%

Fig. 1. Two model equations. Personal probability of early knee OA can be calculated by using one of the following formulas. Model 1 includes baseline clinical factors only; model 2 includes both clinical and radiographic factors. WOMAC subscales are defined as 'positive' when patient report 'moderate' or more severe symptoms. Joint space narrowing is defined as 'positive' when grade≥2 (equals≥ 'minimal').

Universities Osteoarthritis Index (WOMAC) questionnaires and two radiographic items (medial joint space narrowing (JSN) and patellofemoral (PF) JSN).

Although required by model 2, there was no record of bony swelling and patellofemoral JSN in the OAI cohort. Knowing that the prevalence of the two measures in the CHECK cohort was relatively low (4% and 3%), we imputed these two variables into the OAI dataset but set all observations to 'none'.

Outcome measures

To validate the early stage OA diagnosis, we collected final follow-up data, including symptoms and bony structural signs of OA. We used the total WOMAC scores (0−100) and its subscales for pain, function, and knee stiffness to assess symptoms. We assessed structural signs of OA using actual KL grades, KL progression (defined as yes/no based on an increase in KL grade of ≥ 1 from baseline to final follow-up), and JSN progression (defined as yes/no based on an increase in lateral or medial JSN grade of ≥ 1 from baseline to final follow-up). Besides, we collected data on the cumulative surgical rate (total knee replacement (TKR) for OA) during the entire follow-up.

In the CHECK cohort, we used T10 measures as outcomes. In the OAI cohort, we used T9 WOMAC scores, T9 surgery information (cumulative surgical rate over the 9 years) and T8 radiographic measures as the outcomes, because no radiographs were taken at T9 in OAI. Additionally, T8 KL grades were not scored for all the knees in the OAI cohort, approximately 30% of knees in OAI had missing values in T8 for KL grade as well as KL progression, though JSN grade and JSN progression were scored for all the knees.

Statistical analysis

We applied the two diagnostic model equations in both cohorts by using baseline data to calculate the probability of early stage knee OA for each knee[7]. Applying thresholds of 30%/70%, we stratified all knees into three subgroups: 'no OA' (probability $\leq$ 30%), 'uncertain' (probability between 30% and 70%) and 'early stage OA' (probability $\geq$ 70%).

First, we performed internal validation in the CHECK cohort. Using the expert diagnosis as the reference standard, we assessed positive predictive value (PPV) and negative predictive value (NPV) and their 95% confidence intervals (CI) for the 'early stage OA' and 'no OA' diagnoses, respectively. We also assessed PPV and NPV when excluding the knees diagnosed as 'uncertain' by experts, because these knees were excluded while developing the two models [7]. To compare the symptoms between 'early stage OA' knees and 'no OA' knees, as well as between 'uncertain' knees and 'no OA' knees at final follow-up, we calculated mean differences (MD) and 95% CI on T10 WOMAC scores using generalized estimating equation (GEE). We chose GEE because it adjusts for repeated measures within patients where both knees were included. To evaluate the clinical relevance, we adopted the concept of minimal clinically important difference (MCID) for WOMAC scores, and considered the differences as clinically relevant when MD exceeded the MCID. We employed MCID values for WOMAC total score of 7; pain subscale, 9; function subscale, 6; and stiffness subscale, 7 [3,21]. Similarly, we used GEE for testing inter-group differences on T10 KL grade, KL progression, JSN progression and cumulative surgical rate. Odds ratios (OR) and 95% CI were calculated.

Next, for external validation, we performed the same tests on the same outcome measures in all knees selected from OAI incident subcohort. We then performed sensitivity analyses using a subset of symptomatic knees from OAI participants, defined as participants who reported knee pain/stiffness within one month prior to baseline. We did this because most OAI knees (82%) were asymptomatic at baseline, which limited our ability to externally validate our model because it had been developed in a sample of symptomatic knees. Specifically, the prevalence of early stage OA can be higher among symptomatic knees, possibly affecting model performance [11].

Finally, we repeated all analyses using the thresholds of 40%/60%. All analyses were performed using SPSS V.25.0. We defined a p-value < 0.05 as statistically significant.

Results

Participants

670 participants (1042 knees) from the CHECK cohort were included in this study, 81% were female, and mean (SD) age was 56 (5) years. 1548 participants (2937 knees) from the OAI incident subcohort were included, 59% were female, and mean (SD) age was 55 (6) years (see patient selection details in **supplementary figure s2**). Baseline characteristics of all knees are presented in **Table 1.** Only minor differences were observed in baseline characteristics between included and excluded knees within both cohorts (**supplementary table 1 and 2**). The model-based probability distributions in the two cohorts are presented in **Fig. 2**.

Internal validation

With the thresholds of 30%/70%, model 1 diagnosed 293 (28%) knees as 'no OA', 556 (53%) as 'uncertain' and 193 (19%) as 'early stage OA' at baseline; and model 2 diagnosed 315 (30%) knees as 'no OA', 530 (51%) as 'uncertain' and 197 (19%) as 'early stage OA' at baseline. PPV and NPV were similar for the two models (model 1: PPV=0.68 and NPV=0.70; model 2: 0.71 and 0.69), and both were somewhat improved after excluding knees diagnosed as 'uncertain' by the experts (model 1: PPV=0.80 and NPV=0.78; model 2: 0.84 and 0.80) (**Table 2**).

Table 3 summarizes the inter-group differences in outcome measures. For both model 1 and model 2, 'early stage OA' knees presented with statistically and clinical relevantly higher WOMAC scores at 10 years follow-up (MD ranges from 16.3 to 20.5, all $p < 0.001$) compared to 'no OA' knees. Knees with 'early stage OA' also had significantly worse T10 KL grade and JSN compared to knees with 'no OA'. None of the 'no OA' knees underwent surgery within 10 years, whereas 10% and 15% of 'early stage OA' knees did. The differences between 'uncertain' and 'no OA' knees were smaller but still statistically significant and in the same direction. KL progression over 10 years did not differ between any group.

Table 1
Baseline characteristics of knees in CHECK and OAI cohorts.

	CHECK cohort (N=1042)	OAI cohort (N=2937)
Age, mean (SD)	56 (5)	55 (6)
Gender (female), %	81	59
BMI, mean (SD)	26.4 (4.2)	28.2 (4.8)
WOMAC pain, 0-100, mean (SD)	25.6 (17.3)	7.7 (12.6)
WOMAC function, 0-100, mean (SD)	24.3 (17.0)	7.3 (11.9)
WOMAC stiffness, 0-100, mean (SD)	34.5 (20.9)	14.0 (17.2)
WOMAC total, 0-100, mean (SD)	25.4 (16.4)	7.6 (11.3)
KL grade, %		
0	59	55
1	28	22
2	13	23

SD, standard deviation; BMI, body mass index; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index; KL grade, Kellgren & Lawrence grade.

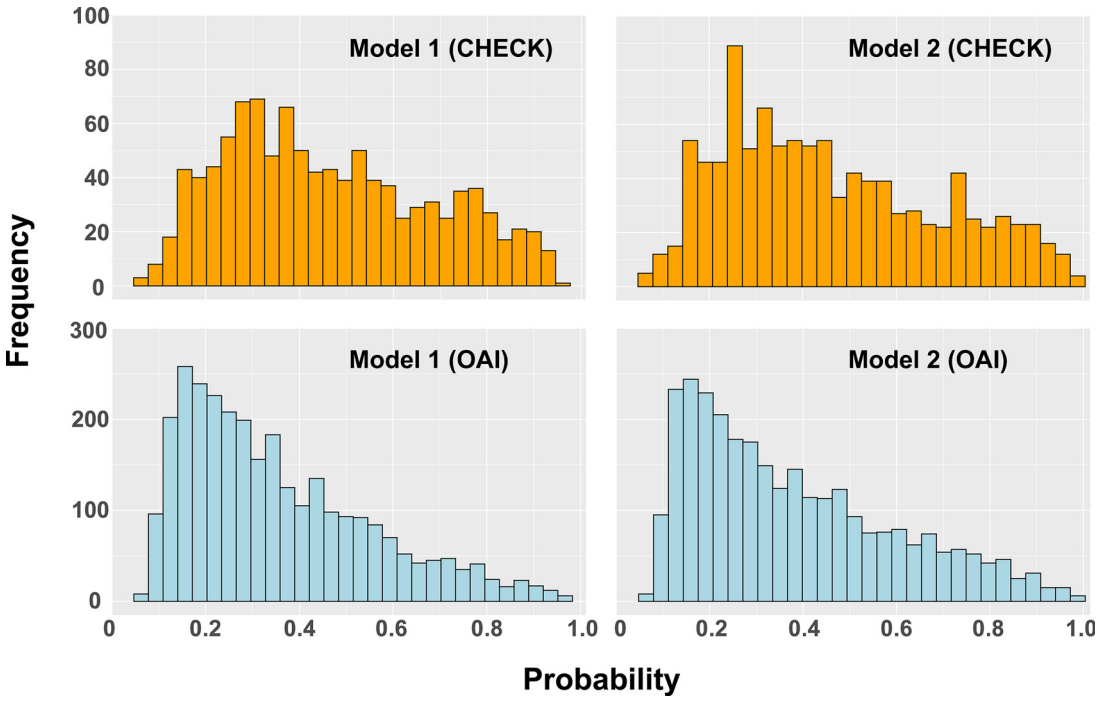

Fig. 2. Two model-based probability distribution in the CHECK and OAI cohorts.

Table 2
Predictive performances of the two models against expert diagnosis in the CHECK cohort (with thresholds of 30%/70%).

		Expert diagnosis (N = 1042)			PPV(95%CI)	NPV(95%CI)	PPV*(95%CI)	NPV* (95%CI)
		No OA, %	Uncertain, %	OA, %				
Model 1	No OA	20	3	5	0.68	0.70	0.80	0.78
	Uncertain	21	12	20	(0.64–0.72)	(0.66–0.74)	(0.76–0.84)	(0.74–0.82)
	Early stage OA	3	3	13				
Model 2	No OA	21	4	5	0.71	0.69	0.84	0.80
	Uncertain	20	12	19	(0.67–0.75)	(0.65–0.73)	(0.80–0.88)	(0.76–0.84)
	Early stage OA	3	3	13				

OA, osteoarthritis; PPV, positive predictive value; NPV, negative predictive value.Percentages were calculated by using the total number (1042) as denominator stratified by model.
* PPV and NPV calculated with excluding expert diagnosed 'uncertain' knees.

External validation

With the thresholds of 30%/70%, model 1 diagnosed 1462 (50%) knees as 'no OA', 1248 (42%) as 'uncertain' and 227 (8%) as 'early stage OA' at baseline; and model 2 diagnosed 1347 (46%) knees as 'no OA', 1276 (43%) as 'uncertain' and 314 (11%) as 'early stage OA' at baseline.

Table 4 summarizes the inter-group differences in outcome measures. For both model 1 and model 2, 'early stage OA' knees presented with statistically and clinical relevantly higher WOMAC scores after 9 years (MD ranges from 11.1 to 14.8, all $p < 0.001$) compared to 'no OA' knees. Knees with 'early stage OA' also had significantly worse KL grade, KL progression, and JSN progression over 8 years compared to 'no OA'. The differences between 'uncertain' and 'no OA' knees were smaller but still statistically significant and in the same direction. Cumulative surgical rate was not well discriminated by the two models.

Sensitivity analysis in the OAI cohort

365 participants with 523 knees reported knee symptoms at baseline. Using model 1, 201 (38%) knees were diagnosed as 'no OA', 243 (47%) as 'uncertain' and 79 (15%) as 'early stage OA' at baseline; and using model 2, 180 (34%) knees were diagnosed as 'no OA', 254 (49%) as 'uncertain' and 89 (17%) as 'early stage OA' at baseline.

In general, the results of differences between 'early stage OA' and 'no OA' were robust to this sensitivity analysis, while the differences between 'uncertain' and 'no OA' knees were non-significant regarding JSN progression, WOMAC pain and function scores (**supplementary table 3**).

Internal and external validation with thresholds of 40%/60%

With the thresholds of 40%/60%, fewer knees were diagnosed as 'uncertain', and inter-group differences were smaller but still statistically significant and remained in the same direction for both the internal and external validation. Analytical results were similar to those for the 30%/70% threshold (**supplementary table 4-6**).

Discussion

Knees diagnosed as 'uncertain' and 'early stage OA' had worse prognosis for symptoms and structural OA features than 'no OA' knees, in both CHECK and OAI populations. This suggests our diagnostic criteria can potentially be applied in real practice.

To establish clinically meaningful diagnostic criteria, we used clinical experts' diagnosis of clinically relevant knee OA as the reference standard when developing the two models. Although patients' symptoms do not always correspond with structural alterations [22,23],

Table 3
Internal validation with thresholds of 30%/70% in 1042 knees from the CHECK cohort.

	Model 1			Uncertain vs. no OA, MD/OR¶ (95%CI)	Early stage OA vs. no OA,MD/OR¶ (95%CI)	Model 2			Uncertain vs. no OA, MD/OR¶ (95%CI)	Early stage OA vs. no OA,MD/OR¶ (95%CI)
	No OA (*N* = 293)	Uncertain (*N*=556)	Early stage OA (*N*=193)			No OA (*N*=315)	Uncertain (*N*=530)	Early stage OA (*N*=197)		
T10 WOMAC pain, 0-100, mean (SD)	15.3 (14.3)	24.8 (19.5)	31.5 (20.4)	9.5** (6.7-12.3)	16.3** (12.2-20.3)	16.1 (15.1)	24.7 (19.3)	31.5 (20.8)	8.6** (5.7-11.4)	15.4** (11.2-19.5)
T10 WOMAC function,0-100, mean (SD)	15.6 (13.8)	25.9 (19.4)	36.1 (21.5)	10.2** (7.4–13.0)	20.5** (16.2–24.7)	16.7 (14.8)	25.7 (19.4)	35.7 (21.6)	9.1** (6.2–11.9)	19.0** (14.7–23.2)
T10 WOMAC stiffness, 0-100, mean (SD)	23.7 (18.4)	34.3 (24.5)	43.1 (24.8)	10.7** (7.1–14.2)	19.4** (14.3–24.5)	24.3 (19.0)	34.3 (24.3)	43.3 (25.2)	10.0** (6.5–13.6)	19.1** (14.1–24.1)
T10 WOMAC total, 0–100, mean (SD)	16.2 (13.4)	26.3 (18.9)	35.7 (20.7)	10.1** (7.4–12.8)	19.5** (15.4–23.6)	17.2 (14.3)	26.2 (18.9)	35.4 (20.9)	9.0** (6.3–11.8)	18.2** (14.1–22.4)
T10 KL grade, %				2.2** (1.6-3.0)	4.7** (3.0-7.2)				2.3** (1.7–3.0)	6.7** (4.2–10.6)
0	8	3	1			7	3	1		
1	44	31	23			45	30	21		
2	47	57	57			47	59	51		
3	1	6	9			1	6	12		
4	0	0	0			0	0	0		
TKR	0	3	10			0	2	15		
KL progression, %	80	82	83	1.2 (0.8–1.7)	1.3 (0.8–2.1)	79	83	84	1.3 (0.9–1.8)	1.4 (0.9–2.3)
JSN progression, %	72	80	83	1.5* (1.0–2.2)	1.9* (1.1–3.3)	73	80	82	1.5* (1.0–2.2)	1.7* (1.0–3.0)
Surgery§, %	0	3	10	#	#	0	2	15	#	#

OA, osteoarthritis; SD, standard deviation; T10, data obtained at 10-year follow-up; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index; KL grade, Kellgren & Lawrence grade; TKR, total knee replacement (treated as KL5 while doing the analysis); JSN, joint space narrowing; MD, mean difference; OR, odds ratio; CI, confidence interval.
¶ MD was calculated for WOMAC scores; OR was for KL grade, KL progression, JSN progression and cumulative surgical rate; § Cumulative surgery rate over 10 years; # OR cannot be calculated as no knee had taken surgery in no OA group; *$p<0.05$; **$p<0.001$.

Table 4
External validation with thresholds of 30%/70% in 2937 knees from the OAI cohort.

	Model 1			Uncertain vs. no OA, MD/OR (95%CI)¶	Early stage OA vs. no OA,MD/OR (95%CI)¶	Model 2			Uncertain vs. no OA, MD/OR (95%CI)¶	Early stage OA vs. no OA, MD/OR (95%CI)¶
	No OA (*N*=1462)	Uncertain (*N*=1248)	Early stage OA (*N*=227)			No OA (*N*=1347)	Uncertain (*N*=1276)	Early stage OA (*N*=314)		
T9 WOMAC pain, 0-100, mean (SD)	6.0 (9.6)	9.6 (14.0)	18.5 (19.9)	3.5** (2.5-4.6)	12.5** (9.6-15.4)	5.6 (9.8)	9.5 (14.1)	16.7 (18.7)	3.9** (2.9-4.9)	11.1** (8.7-13.5)
T9 WOMAC function, 0-100, mean (SD)	5.5 (10.3)	9.7 (13.8)	18.5 (18.5)	4.2** (3.2–5.3)	13.1** (10.2–15.9)	5.1 (9.7)	9.4 (13.6)	17.3 (17.9)	4.3** (3.3–5.3)	12.2** (9.8–14.6)
T9 WOMAC stiffness, 0–100, mean (SD)	10.4 (15.1)	16.0 (18.5)	25.1 (21.9)	5.7** (4.2–7.1)	14.7** (11.3–18.1)	10.0 (14.8)	15.3 (18.2)	24.8 (21.3)	5.3** (3.8–6.7)	14.8** (11.9–17.7)
T9 WOMAC total, 0–100, mean (SD)	5.8 (9.7)	9.8 (12.9)	18.3 (17.5)	4.0** (3.1–5.0)	12.6** (9.9–15.2)	5.4 (9.2)	9.5 (12.9)	17.1 (16.8)	4.1** (3.2–5.1)	11.7** (9.5–14.0)
T8 KL grade, %				2.2#** (1.8–2.6)	4.9#** (3.3–7.2)				2.3#** (1.9–2.8)	4.3#** (3.0–6.1)
0	48	31	16			50	31	19		
1	18	18	14			18	19	13		
2	6	11	15			6	11	12		
3	2	4	8			1	4	7		
4	0.1	0.2	1			0	0.5	0		
TKR	1	2	2			1	2	2		
Missing†	25	34	44			24	32	47		
KL progression†, %	15	31	46	2.5#** (2.0–3.2)	4.8#** (3.1–7.3)	15	30	43	2.4#** (1.9–3.0)	4.3#** (2.9–6.3)
JSN progression, %	8	15	26	2.1** (1.6–2.8)	4.3** (2.9–6.3)	7	14	24	2.1** (1.6–2.7)	4.1** (2.8–5.8)
Surgery§, %	1	2	2	2.1 (0.9–4.4)	2.7 (0.9–8.2)	1	2	2	1.7 (0.8–3.4)	2.5* (1.0–6.4)

OA, osteoarthritis; SD, standard deviation; T9/8, data obtained at 9/8-year follow-up; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index, (0-100, 0 indicates no symptom); KL grade, Kellgren & Lawrence grade; TKR, total knee replacement (treated as KL5 while doing the analysis); JSN, joint space narrowing; MD, mean difference; OR, odds ratio; CI, confidence interval. Model 1 includes clinical factors only; model 2 includes both clinical and radiographic factors.
¶ MD was calculated for WOMAC scores; OR was for KL grade, KL progression, JSN progression and cumulative surgical rate; † T8 KL grades were not scored for all the knees in the OAI cohort, KL progression had same proportion of missing values and percentages were calculated by excluding missing values; § Cumulative surgery rate over 9 years; # OR were calculated after excluding missing values; *$p < 0.05$; **$p < 0.001$.

our model based diagnostic criteria for early stage OA predicted, on average, worse symptoms and radiographic findings at follow-up. Many previous models were developed for predicting 'structural OA', while these had poor or uncertain performances in discriminating knee symptoms [24–27]. In the present study, the differences in WOMAC scores between 'early stage OA' and 'no OA' knees in the two cohorts were consistently beyond the MCID, however we acknowledge that use of MCID is controversial and may not be generalizable [21].

The ideal way to externally validate the current models would have been to use expert consensus-based diagnosis (i.e., the criterion reference standard) as the outcome measure. However, this type of diagnosis is generally not available in cohort studies, and this prevented us from performing an ideal external validation. Instead, we used WOMAC scores and radiographic features because these were the features used by clinical experts to diagnose knees in the CHECK cohort. We assumed the knees with more severe symptoms (higher WOMAC scores) and more structural damage would be more likely to be diagnosed as knee OA by experts, and this was confirmed during internal validation of this study. On the other hand, experts' diagnoses could vary from one region to another and could be influenced by their specialties (e.g. primary vs. secondary care)[19]. Nevertheless, using these commonly available features will facilitate further replication of our results in other studies, as well as in clinical practice.

Setting thresholds for the probabilities derived from specific models is somewhat subjective and arbitrary, according to previous studies [11,12,20]. Laure *et al.* stated that the choice of threshold should reflect the potential for harm of a false positive and the potential benefit of a true positive [12]. Since no study has reported on the harms and benefits of early stage knee OA diagnosis, our choice was based on the assumption that potential harm is approximately the same as benefit. We adopted the category of 'uncertain' in the probability stratification because our previous study showed that several knees were difficult to diagnose and were marked as 'uncertain' by the clinical experts [19]. Moreover, this form of stratification was also frequently applied in other studies [13,20,28,29]. Using thresholds at 30%/70% and 40%/60%, predictive values of the two models reached $\pm$0.70 and $\pm$0.60 in the CHECK cohort, respectively. Since the predictive value of 0.60 is slightly lower than the values reported for clinical diagnosis in other chronic musculoskeletal diseases [30,31], we decided to use 30%/70% as the primary thresholds. However, further studies are needed to assess the cost-effectiveness and clinical impact of these thresholds, before deciding which one to implement.

Early diagnosis should facilitate treatment decisions. Nevertheless, since no intervention has been tested for its effectiveness in 'early stage OA' population, our early diagnostic criteria may, therefore, help open the 'early treatment window' and facilitate future intervention studies. As no diagnostic criteria are perfect, when effective early interventions become more widely available, the benefit of the treatment in true positive cases should be well balanced against the harm (over-treatment) in false positive cases. For instance, in the CHECK cohort, if a 70% effective intervention (deduced based on the efficacy of disease-modifying drug in rheumatoid arthritis [32]) for preventing clinically relevant OA were to be implemented in all 'early stage OA' knees, about 50% to 60% of knees would benefit from such treatment, and 20% to 30% of the knees might be overtreated. All those treated would be exposed to potential side-effects of the treatment. Moreover, based on the results of this study, 'uncertain' knees should be considered more likely to develop into OA than 'no OA' knees, so interventions might also be appropriate for these 'uncertain' knees. If so, these interventions would need to be of low cost with few side effects. In addition, we have also developed 2-year disease course based early diagnostic criteria, which can provide a re-evaluation (especially when symptoms and structural damage get worse) within 2 years [33]. Thus, a suggestion of regular follow-up can be delivered to these 'uncertain' knees for monitoring disease status and probably adjusting treatment strategies.

Despite having no baseline radiographic feature in model 1, the model still predicted bony structural changes. Given that use of radiographs is discouraged for patients suspected of having knee OA in primary care [19], model 1 may be a more useful tool in clinical practice. In our previous study we found that radiographic features were of limited added value for model performance [7]. These findings were confirmed in the present study: model 2 only increased the PPV by 3-4% and NPV by 1-2%. These findings were further supported by using patient symptoms and structural damage as outcomes since the two models' discriminative abilities among symptomatic and structural features were very similar in both CHECK and OAI cohorts.

Model performance always varies from one dataset to another, because of heterogeneity in study designs. This heterogeneity, in turn, can help to assess the 'reliability' of the model when applied in different situations. For the current study, there are several noteworthy differences in patient characteristics, data obtaining and research background between CHECK and OAI cohorts. First, CHECK recruited individuals who reported knee complaints but had not yet, or had only recently, sought medical care, while the OAI incident subcohort recruited individuals at high risk of developing knee OA, and who were mostly asymptomatic at baseline. Moreover, in CHECK the two knees of patients from the CHECK shared one series of WOMAC scores, while in OAI participants completed WOMAC questionnaires separately for each knee [34]. This could explain differences between the two cohorts among WOMAC scores. As expected, the baseline WOMAC scores in OAI were generally lower than that in the CHECK. According to the model equations, lower WOMAC scores would result in lower probabilities. This could be a reason why fewer knees were diagnosed as 'early stage OA' in OAI. Despite the difference, the results of external validation and sensitivity analyses support the use of our criteria in the OAI knees. Second, differences in patient characteristics are also reflected in radiographic findings. Most knees in the CHECK progressed in KL/JSN grade, but only a few knees did in OAI, despite the fact that the methods of KL grading were similar in the two cohorts [34]. In this case, our study results suggest that our models remain valid in low pre-test probability populations, in terms of radiographic progression. Third, the sex balance is different between the two cohorts. This can be explained by the fact that the CHECK cohort was designed without a pre-specified goal on sex balance, while the OAI cohort aimed for equal numbers of men and women in each age group. Finally, the social, cultural and healthcare system differences between the USA and the Netherlands cannot be ignored. For example, the decision for knee joint surgery may be influenced by regional factors and may have influenced surgical rates between the two cohorts. Given that model's discriminative ability can be influenced by the prevalence [11], the small number of surgical cases may be the cause of the insignificant results in the OAI cohort.

This study has limitations. First, only one cohort was used for external validation. Even though several other knee OA cohorts were identified by our research group [27,35,36], they all lacked several factors required for our models, especially the factors of physical examination. Second, since there is no standard expert OA diagnosis for the knees in the OAI cohort, we could not evaluate the two models' external diagnostic abilities, such as assessing receiver operating characteristic curve, PPV and NPV. However, based on the similar patterns of symptoms and radiographic features over the three categories across the two cohorts, we believe it remains reasonable to consider those in OAI with 'early stage OA' to be similar to those in CHECK, and therefore likely candidates for early intervention. Third, radiographic PF OA was included in the CHECK cohort and was taken into consideration while developing the two models, but there were no data regarding radiographic PF OA in OAI. Therefore, this study couldn't validate the models' performance in regards to PF OA.

Fourth, both knees of each person were included in this study, which might have influenced the results. At baseline, knees from a unilateral (early stage) OA patient shared or had correlated WOMAC scores, which could have increased the probability of having 'early stage OA' in the 'healthy' knee. However, unilateral knee OA is very likely (up to 80% probability) to progress to bilateral OA over time [37,38]. Thus, it seems fair to assign a slightly over-estimated probability for the '(temporarily) healthy' knee. When comparing outcome measures, although GEE was applied to adjust for possible correlated measures between knees within patients, residual effects cannot be fully ruled out. Fifth, validating diagnostic criteria in cohort studies has its natural limitations, because participants included in the cohorts (selected by multiple inclusion and exclusion criteria) may not represent patients seen in clinical practice. This study included a group of patients who were suspected of early stage knee OA, and aged 45 to 65 years with knee KL grade $\leq$ 2. Therefore, results may not be applicable to younger or older patients, or patients with higher KL grades. The age range was selected based on prior knowledge that the population prevalence of knee OA starts to increase from 45 years and the two models were originally developed in a population aged 45 to 65 years [1,19]. Besides, most knees suspected of having early OA present with KL$\leq$ 2 [34]. Sixth, patients lost to final follow-up were excluded from this study, which might have introduced selection bias, although the baseline characteristics of excluded knees were similar to those of the included knees. We didn't use imputation since it would use similar mathematical algorithms as prediction models, which may have unrealistically optimized our results.

In conclusion, these results support the internal and external validity of the two model-based early diagnostic criteria for knee OA. 'Early stage' OA knees were found to have worse prognosis than 'no OA' knees in both CHECK and OAI populations. Further studies on transforming the two diagnostic models into clinically feasible diagnostic tools and evaluating the impact of clinical application are warranted.

Author Contribution

Q.W.: Conception and design, analysis and interpretation of the data, drafting of the article, final approval of the article. **J.R.**: Conception and design, analysis and interpretation of the data, critical revision of the article for important intellectual content, final approval of the article, collection and assembly of data, obtaining of funding. **M.K.**: Interpretation of the data, critical revision of the article for important intellectual content, final approval of the article. **M.B.**: Statistical expertise, table and figure formatting, interpretation of the data, critical revision of the article for important intellectual content, final approval of the article. **J.W.J.B.**: Interpretation of the data, critical revision of the article for important intellectual content, final approval of the article. **S.B.Z**: Conception and design, Interpretation of the data, critical revision of the article for important intellectual content, final approval of the article, obtaining of funding, administrative support.

Declaration of Competing Interest

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), grants from IMI-APPROACH, Board member OARSI, President Dutch Society Rheumatology and Member EULAR Council.

Acknowledgments

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. We thank the OAI cohort research team for sharing patient data. We thank Erin Macri for checking and editing the language.

Role of the funding source

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.semarthrit.2022.152007.

References

[1] Hunter DJ, Osteoarthritis Bierma-Zeinstra S. Lancet 2019;393(10182):1745–59 doi: S0140-6736(19)30417-9.
[2] Brandt KD, Mazzuca SA. Lessons learned from nine clinical trials of disease-modifying osteoarthritis drugs. Arthritis Rheum 2005;52(11):3349–59. doi: 10.1002/art.21409.
[3] Hochberg MC, Guermazi A, Guehring H, Aydemir A, Wax S, Fleuranceau-Morel P, et al. Effect of intra-articular sprifermin vs placebo on femorotibial joint cartilage thickness in patients with osteoarthritis: the FORWARD randomized clinical trial. JAMA 2019;322(14):1360–70 doi: 2752470.
[4] Cainap C, Pop LA, Balacescu O, Cainap SS. Early diagnosis and screening in lung cancer. Am J Cancer Res 2020;10(7):1993–2009.
[5] Kelder JC, Cramer MJ, van Wijngaarden J, van Tooren R, Mosterd A, Moons KG, et al. The diagnostic value of physical examination and additional testing in primary care patients with suspected heart failure. Circulation 2011;124(25):2865–73 CIRCULATIONAHA.111.019216.
[6] Sparks JA. Rheumatoid arthritis. Ann Intern Med 2019;170(1):ITC1–ITC16 doi: 2719975.
[7] Runhaar J, Kloppenburg M, Boers M, Bijlsma JWJ, Bierma-Zeinstra SMA. Towards developing diagnostic criteria for early knee osteoarthritis: data from the CHECK study. Rheumatology (Oxford) 2021;60(5):2448–55. doi: 10.1093/rheumatology/keaa643.
[8] Luyten FP, Bierma-Zeinstra S, Dell'Accio F, Kraus VB, Nakata K, Sekiya I, et al. Toward classification criteria for early osteoarthritis of the knee. Semin Arthritis Rheum 2018;47(4):457–63. doi: 10.1016/j.semarthrit.2017.08.006.
[9] Aggarwal R, Ringold S, Khanna D, Neogi T, Johnson SR, Miller A, et al. Distinctions between diagnostic and classification criteria? Arthritis Care Res (Hoboken) 2015;67(7):891–7. doi: 10.1002/acr.22583.
[10] Mahmoudian A, Lohmander LS, Jafari H, Luyten FP. Towards classification criteria for early-stage knee osteoarthritis: a population-based study to enrich for progressors. Semin Arthritis Rheum 2021;51(1):285–91. doi: 10.1016/j.semarthrit.2020.11.002.
[11] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162(1):W1–73 doi: 2088542.
[12] Wynants L, van Smeden M, McLernon DJ, Timmerman D, Steyerberg EW, Van Calster B, et al. Three myths about risk thresholds for prediction models. BMC Med 2019;17(1):192. doi: 10.1186/s12916-019-1425-3.
[13] Manola J, Royston P, Elson P, McCormack JB, Mazumdar M, Negrier S, et al. Prognostic model for survival in patients with metastatic renal cell carcinoma: results from the international kidney cancer working group. Clin Cancer Res 2011;17(16):5443–50 doi: 1078-0432.CCR-11-0553.

[14] Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ 2012;344:e4181. doi: 10.1136/bmj.e4181.
[15] Deyle GD, Allen CS, Allison SC, Gill NW, Hando BR, Petersen EJ, et al. Physical therapy versus glucocorticoid injection for osteoarthritis of the knee. N Engl J Med 2020;382(15):1420–9. doi: 10.1056/NEJMoa1905877.
[16] Runhaar J, van Middelkoop M, Reijman M, Willemsen S, Oei EH, Vroegindeweij D, et al. Prevention of knee osteoarthritis in overweight females: the first preventive randomized controlled trial in osteoarthritis. Am J Med 2015;128(8):888-95 e4. doi: S0002-9343(15)00244-2.
[17] Roos EM, Arden NK. Strategies for the prevention of knee osteoarthritis. Nat Rev Rheumatol 2016;12(2):92–101. doi: 10.1038/nrrheum.2015.135.
[18] Wesseling J, Boers M, Viergever MA, Hilberdink WK, Lafeber FP, Dekker J, et al. Cohort profile: cohort hip and cohort knee (CHECK) study. Int J Epidemiol 2016;45(1):36–44. doi: 10.1093/ije/dyu177.
[19] Wang Q, Runhaar J, Kloppenburg M, Boers M, Bijlsma JWJ, Bierma-Zeinstra SMA, 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(10). doi: 10.3390/jcm9103374.
[20] Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American heart association. Circulation 2009;119(17):2408–16. doi: 10.1161/CIRCULATIONAHA.109.192278.
[21] Bellamy N, Hochberg M, Tubach F, Martin-Mola E, Awada H, Bombardier C, et al. Development of multinational definitions of minimal clinically important improvement and patient acceptable symptomatic state in osteoarthritis. Arthritis Care Res (Hoboken) 2015;67(7):972–80. doi: 10.1002/acr.22538.
[22] Schiphof D, Runhaar J, Waarsing JH, van Spil WE, van Middelkoop M, Bierma-Zeinstra SMA. 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(10):1491–500. doi: 10.1016/j.joca.2019.06.002.
[23] van Berkel AC, Schiphof D, Waarsing JH, Runhaar J, van Ochten JM, Bindels PJE, et al. 10-Year natural course of early hip osteoarthritis in middle-aged persons with hip pain: a CHECK study. Ann Rheum Dis 2021. doi: 10.1136/annrheumdis-2020-218625.
[24] Sharma L, Hochberg M, Nevitt M, Guermazi A, Roemer F, Crema MD, et al. Knee tissue lesions and prediction of incident knee osteoarthritis over 7 years in a cohort of persons at higher risk. Osteoarthritis Cartilage 2017;25(7):1068–75. doi: 10.1016/j.joca.2017.02.788.
[25] van Oudenaarde K, Jobke B, Oostveen AC, Marijnissen AC, Wolterbeek R, Wesseling J, et al. Predictive value of MRI features for development of radiographic osteoarthritis in a cohort of participants with pre-radiographic knee osteoarthritis-the CHECK study. Rheumatology (Oxford) 2017;56(1):113–20. doi: 10.1093/rheumatology/kew368.
[26] Kerkhof HJ, Bierma-Zeinstra SM, Arden NK, Metrustry S, Castano-Betancourt M, Hart DJ, et al. Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. Ann Rheum Dis 2014;73(12):2116–21. doi: 10.1136/annrheumdis-2013-203620.
[27] Camacho-Encina M, Balboa-Barreiro V, Rego-Perez I, Picchi F, VanDuin J, Qiu J, et al. Discovery of an autoantibody signature for the early diagnosis of knee osteoarthritis: data from the osteoarthritis initiative. Ann Rheum Dis 2019;78(12):1699–705. doi: 10.1136/annrheumdis-2019-215325.
[28] Krupp NL, Weinstein G, Chalian A, Berlin JA, Wolf P, Weber RS. Validation of a transfusion prediction model in head and neck cancer surgery. Arch Otolaryngol Head Neck Surg 2003;129(12):1297–302. doi: 10.1001/archotol.129.12.1297.
[29] Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest 2009;27(3):235–43. doi: 10.1080/07357900802572110.
[30] Cook CJ, Cook CE, Reiman MP, Joshi AB, Richardson W, Garcia AN. Systematic review of diagnostic accuracy of patient history, clinical findings, and physical tests in the diagnosis of lumbar spinal stenosis. Eur Spine J 2020;29(1):93–112. doi: 10.1007/s00586-019-06048-4.
[31] Petersen T, Laslett M, Juhl C. Clinical classification in low back pain: best-evidence diagnostic rules based on systematic reviews. BMC Musculoskelet Disord 2017;18(1):188. doi: 10.1186/s12891-017-1549-6.
[32] Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. Lancet 2016;388(10055):2023–38. doi: 10.1016/S0140-6736(16)30173-8.
[33] Wang Q, Runhaar J, Kloppenburg M, Boers M, Bijlsma JWJ, Bierma-Zeinstra SMA, et al. Diagnosis of early stage knee osteoarthritis based on early clinical course: data from the CHECK cohort. Arthritis Res Ther 2021;23(1):217. doi: 10.1186/s13075-021-02598-5.
[34] Wesseling J, Dekker J, van den Berg WB, Bierma-Zeinstra SM, Boers M, Cats HA, et al. CHECK (Cohort Hip and Cohort Knee): similarities and differences with the Osteoarthritis Initiative. Ann Rheum Dis 2009;68(9):1413–9. doi: 10.1136/ard.2008.096164.
[35] Segal NA, Nevitt MC, Gross KD, Hietpas J, Glass NA, Lewis CE, et al. The Multicenter Osteoarthritis Study: opportunities for rehabilitation research. PM R 2013;5(8):647–54. doi: 10.1016/j.pmrj.2013.04.014.
[36] Mallen CD, Peat G, Thomas E, Lacey R, Croft P. Predicting poor functional outcome in community-dwelling older adults with knee pain: prognostic value of generic indicators. Ann Rheum Dis 2007;66(11):1456–61. doi: 10.1136/ard.2006.067975.
[37] Metcalfe AJ, Andersson ML, Goodfellow R, Thorstensson CA. Is knee osteoarthritis a symmetrical disease? Analysis of a 12 year prospective cohort study. BMC Musculoskelet Disord 2012;13:153. doi: 10.1186/1471-2474-13-153.
[38] Gunther KP, Sturmer T, Sauerland S, Zeissig I, Sun Y, Kessler S, et al. Prevalence of generalised osteoarthritis in patients with advanced hip and knee osteoarthritis: the Ulm Osteoarthritis Study. Ann Rheum Dis 1998;57(12):717–23. doi: 10.1136/ard.57.12.717.