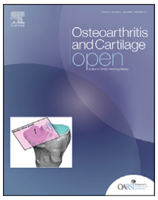
Contents lists available at ScienceDirect

Osteoarthritis and Cartilage Open

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



Prognostic value of B-score for predicting joint replacement in the context of osteoarthritis phenotypes: Data from the osteoarthritis initiative

F. Saxer [a,b,h], D. Demanse [c,h], A. Brett [d], D. Laurent [e], L. Mindeholm [a], P.G. Conaghan [f], M. Schieker [a,g,*]

[a] *Novartis Biomedical Research, Novartis Campus, 4002, Basel, Switzerland*
[b] *Medical Faculty, University of Basel, 4002, Basel, Switzerland*
[c] *Novartis Pharma AG, 4002, Basel, Switzerland*
[d] *Imorphics, Worthington House, Towers Business Park, Wilmslow Road, Manchester, M20 2HJ, UK*
[e] *Novartis Biomedical Research, Biomarker Development, 4002, Basel, Switzerland*
[f] *Leeds Institute of Rheumatic & Musculoskeletal Medicine, University of Leeds and NIHR Leeds Biomedical Research Centre, UK*
[g] *Medical Faculty, Ludwig-Maximilians-University, Munich, 80336, Germany*

ARTICLE INFO

Handling Editor: Professor H Madry

Keywords:
Bone shape
Cluster analysis
OA imaging
Patient stratification

ABSTRACT

Objective: Developing new therapies for knee osteoarthritis (KOA) requires improved prediction of disease progression. This study evaluated the prognostic value of clinical clusters and machine-learning derived quantitative 3D bone shape B-score for predicting total and partial knee replacement (KR).

Design: This retrospective study used longitudinal data from the Osteoarthritis Initiative. A previous study used patients' clinical profiles to delineate phenotypic clusters. For these clusters, the distribution of B-scores was assessed (employing Tukey's method). The value of both cluster allocation and B-score for KR-prediction was then evaluated using multivariable Cox regression models and Kaplan-Meier curves for time-to-event analyses. The impact of using B-score vs. cluster was evaluated using a likelihood ratio test for the multivariable Cox model; global performances were assessed by concordance statistics (Harrell's C-index) and time dependent receiver operating characteristic (ROC) curves.

Results: B-score differed significantly for the individual clinical clusters ($p < 0.001$). Overall, 9.4% of participants had a KR over 9 years, with a shorter time to event in clusters with high B-score at baseline. Those clusters were characterized clinically by a high rate of comorbidities and potential signs of inflammation. Both phenotype and B-score independently predicted KR, with better prediction if combined ($P < 0.001$). B-score added predictive value in groups with less pain and radiographic severity but limited physical activity.

Conclusions: B-scores correlated with phenotypes based on clinical patient profiles. B-score and phenotype independently predicted KR surgery, with higher predictive value if combined. This can be used for patient stratification in drug development and potentially risk prediction in clinical practice.

1. Introduction

Osteoarthritis (OA) is a major health care challenge affecting more than 500 million patients worldwide [1] associated with personal suffering [2], reduction in quality of life [3], loss of independence [4] and even mortality [5]. Despite decades of research since Kellgren and Lawrence (KL) first suggested a radiographic classification of disease severity in 1957 [6] and major advances in the understanding the pathology of the disease, there is still no medication licensed as disease modifying treatment for OA [7].

It is now widely accepted that OA may be triggered by various mechanisms culminating in the clinical and radiographic picture of OA [8]. Phenotyping has been employed to stratify patients based on observable traits, assuming that this could translate into the

* Corresponding author. Novartis Institutes for Biomedical Research, Novartis Campus, 4002, Basel, Switzerland.
E-mail addresses: franziska.saxer@novartis.com (F. Saxer), david.demanse@novartis.com (D. Demanse), Alan.Brett@stryker.com (A. Brett), didier.laurent@novartis.com (D. Laurent), linda.mindeholm@novartis.com (L. Mindeholm), P.Conaghan@leeds.ac.uk (P.G. Conaghan), matthias.schieker@novartis.com (M. Schieker).

[h] These authors contributed equally.

https://doi.org/10.1016/j.ocarto.2024.100458
Received 12 July 2023; Received in revised form 27 January 2024; Accepted 29 February 2024

Table 1
Summary of Bone shape score at baseline and type of knee replacement events across clusters.

Clusters	D1 (N = 619)	D2 (N = 849)	D3 (N = 860)	D4 (N = 785)	D5 (N = 1551)	total	M1 (N = 1524)	M2 (N = 2146)	M3 (N = 1004)	total
B-score (left)										
Number of missing	4	17	26	12	31	90	26	38	27	91
Mean (SD) (95% CI)	0.57 (1.45) (0.46, 0.69)	1.57 (1.98) (1.43, 1.7)	1.30 (1.93) (1.17, 1.43)	0.62 (1.57) (0.51, 0.74)	0.82 (1.66) (0.73, 0.9)	0.98 (1.77) (0.92, 1.03)	0.93 (1.77) (0.84, 1.02)	0.67 (1.57) (0.6, 0.74)	1.70 (1.98) (1.57, 1.82)	0.97 (1.77) (0.92, 1.03)
Median (Q1, Q3)	0.46 (−0.36, 1.29)	1.24 (0.19, 2.63)	0.95 (−0.05, 2.27)	0.43 (−0.35, 1.34)	0.60 (−0.26, 1.60)	0.71 (−0.18, 1.83)	0.71 (−0.20, 1.72)	0.47 (−0.32, 1.39)	1.42 (0.24, 2.91)	0.71 (−0.18, 1.83)
Min - Max	−3.30–6.35	−3.46–8.69	−3.30–8.96	−3.22–8.54	−3.13–8.40	−3.46–8.96	−3.36–8.54	−2.93–8.96	−3.46–8.69	−3.46–8.96
B-score (right)										
Number of missing	6	14	11	10	29	70	17	30	23	70
Mean (SD) (95% CI)	0.66 (1.53) (0.54, 0.78)	1.61 (1.96) (1.47, 1.74)	1.42 (2.02) (1.29, 1.56)	0.72 (1.52) (0.61, 0.82)	0.88 (1.69) (0.8, 0.97)	1.06 (1.80) (1, 1.11)	1.05 (1.79) (0.95, 1.14)	0.78 (1.60) (0.71, 0.85)	1.67 (2.03) (1.55, 1.8)	1.06 (1.80) (1, 1.11)
Median (Q1, Q3)	0.51 (−0.35, 1.50)	1.27 (0.21, 2.67)	1.06 (0.00, 2.53)	0.55 (−0.30, 1.44)	0.65 (−0.23, 1.67)	0.76 (−0.14, 1.94)	0.76 (−0.11, 1.86)	0.59 (−0.32, 1.58)	1.41 (0.15, 2.91)	0.76 (−0.14, 1.94)
Min - Max	−3.26–6.61	−3.28–8.28	−3.09–9.79	−3.12–8.29	−3.41–9.97	−3.41–9.97	−3.41–8.56	−3.28–9.97	−2.80–9.79	−3.41–9.97
B-score (maximum of left or right)										
Number of missing	0	0	1	1	3	5	0	2	3	5
Mean (SD)	0.97 (1.58)	2.15 (2.08)	1.84 (2.10)	1.02 (1.65)	1.23 (1.79)	1.44 (1.91)	1.45 (1.94)	1.08 (1.68)	2.20 (2.10)	1.44 (1.91)
(95% CI)	(0.85, 1.1)	(2.01, 2.29)	(1.7, 1.98)	(0.91, 1.14)	(1.14, 1.32)	(1.39, 1.5)	(1.36, 1.55)	(1.01, 1.15)	(2.07, 2.33)	(1.39, 1.5)
Median (Q1, Q3)	0.79 (−0.09, 1.79)	1.83 (0.69, 3.43)	1.48 (0.30, 3.07)	0.74 (−0.09, 1.73)	0.91 (0.07, 2.09)	1.09 (0.12, 2.42)	1.08 (0.16, 2.35)	0.82 (−0.03, 1.87)	1.93 (0.63, 3.54)	1.09 (0.12, 2.42)
Min - Max	−2.99–6.61	−3.01–8.69	−2.63–9.79	−2.56–8.54	−3.13–9.97	−3.13–9.97	−3.13–8.56	−2.93–9.97	−2.63–9.79	−3.13–9.97
First OA surgery related event										
Number of missing	578	689	758	748	1454	4227	1388	2003	844	4235
Total	35 (85%)	147 (92%)	98 (96%)	31 (84%)	89 (92%)	400 (92%)	121 (89%)	132 (92%)	149 (93%)	402 (92%)
Partial	6 (15%)	13 (8%)	4 (4%)	6 (16%)	8 (8%)	37 (8%)	15 (11%)	11 (8%)	11 (7%)	37 (8%)

identification of endotypes and thereby allow targeted treatment approaches [9,10].

One challenge in these efforts is the definition of structural disease severity, typically assessed using the KL classification, which has considerable limitations [11]. Another challenge is, despite an array of potential candidates, the lack of a reliable and validated imaging biomarker, especially with predictive validity for patient-relevant outcomes such as pain, function, or risk of joint replacement surgery [12,13]. Three-dimensional bone shape (termed B-score when referring to femoral shape) is a candidate imaging biomarker which can be derived from CT or MRI [14]. Previous analyses have demonstrated its linear metric provides a more granular assessment of OA severity compared to KL grading [15]. Bone shape has demonstrated an association with both radiographic disease progression and joint replacement [16–18]. In addition, responsiveness to investigational treatments in clinical trials via a flattening of the progression trajectory [19] and also an association with physical function have been suggested [20].

Despite promising data, the reliance of B-score on imaging techniques that are not commonly used for the assessment of OA [21,22], is a potential disadvantage for a potential risk assessment tool in clinical practice. Also, as a measure for screening and stratification in OA clinical trials, B-score requires 3D imaging and, in the case of CT scans, radiation exposure.

In a previous analysis of the Osteoarthritis Initiative (OAI) database, a multi-center, longitudinal, prospective observational cohort-study of knee OA including 4796 participants, we identified OA phenotypes based on clinical and performance characteristics [23].

These cluster allocations were associated with different levels of pain and joint space width at baseline, and over time clusters demonstrated slightly different patterns of disease progression; the relationship with joint replacement was not examined. Pain levels remained stable for clusters comprising patients with many comorbidities with a trend towards a pain increase in the other clusters. Joint space loss over time was more pronounced in a potentially inflammatory cluster characterized by a high rate of effusion compared to the other clusters [23].

Assuming an association between phenotypes and endotypes in OA, we hypothesized that there may be a relationship between B-score and the identified clinical clusters, allowing the use of one or the other for prediction of knee replacement (KR) in different settings (research or clinical practice), and dependent on available resources. The evaluation of B-score and clinical variables as potential predictors of KR is a first step in building a robust model for risk prediction, that could subsequently include other imaging or soluble biomarkers. Clusters based on clinical information and B-score have been chosen initially due to their availability at low cost (in the case of the clinical clusters), and the previous reported predictive value and sensitivity to change of B-score [15,24,25].

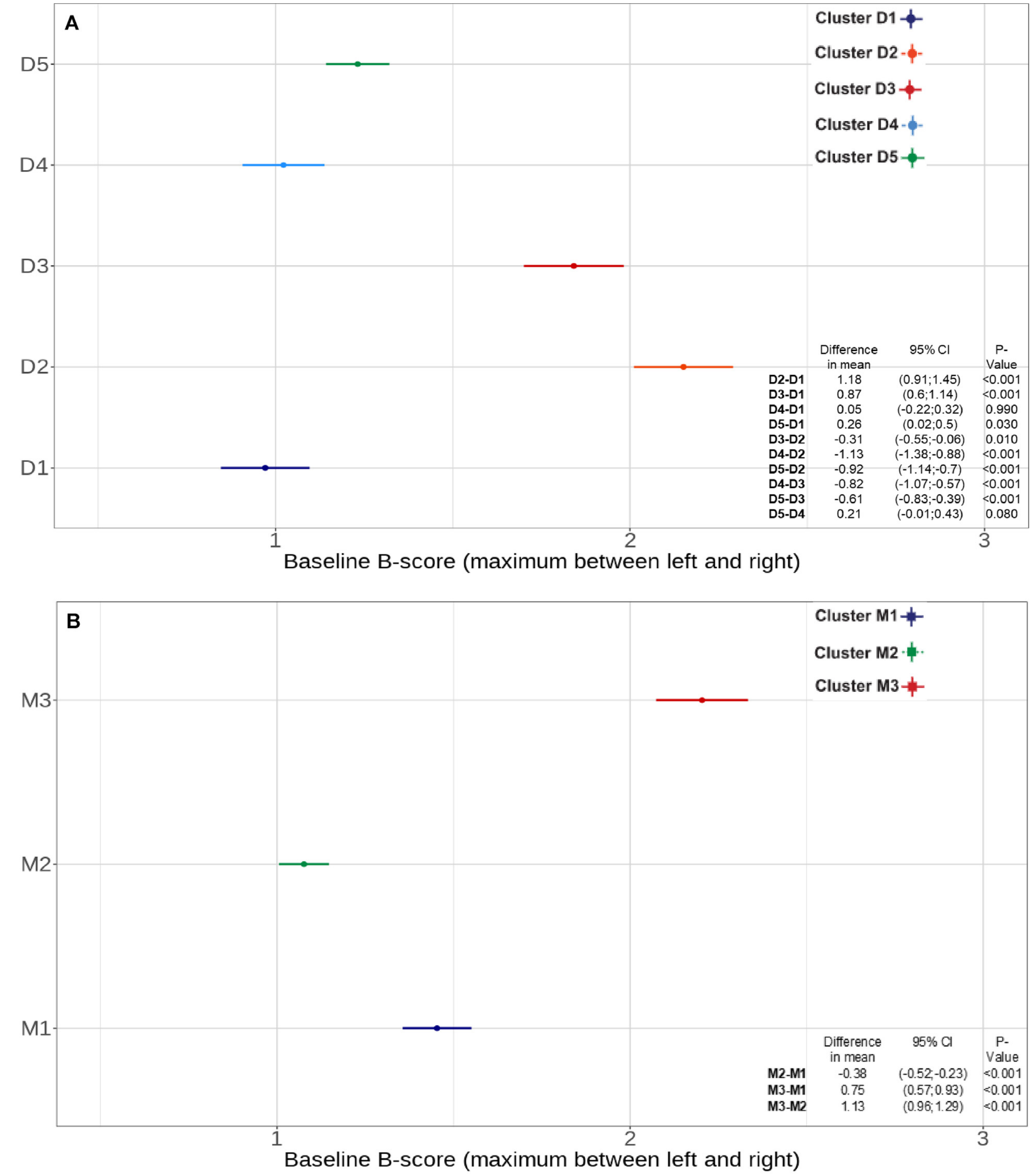
The present study therefore aimed at investigating the relationship between phenotypic cluster allocation and B-score and examined them both individually and in combination as predictors of total or partial KR.

2. Methods

The study was based on the data from the OAI. The initial study was approved by institutional review boards of the participating centres [26]. All patients gave informed consent to the collection of their data and to the secondary use. This study was approved by the Ethikkommission Nordwest-und Zentralschweiz (Basec-No. 2022-01979).

Data from the incident and progression cohorts [26] were used in this analysis. The progression-cohort (n = 1390) consisted of subjects with frequent knee symptoms and radiographic signs of tibio-femoral knee OA; the incidence-cohort (n = 3284) consisted of subjects with frequent knee symptoms without radiographic evidence of OA, plus at least 2 risk factors for OA.

In a previous evaluation, 157 baseline variables were selected for the analysis based on their clinical relevance and availability in the OAI as well as in clinical practice to allow a transfer of insights to various



	Difference in mean	95% CI	P-Value
D2-D1	1.18	(0.91;1.45)	<0.001
D3-D1	0.87	(0.6;1.14)	<0.001
D4-D1	0.05	(-0.22;0.32)	0.990
D5-D1	0.26	(0.02;0.5)	0.030
D3-D2	-0.31	(-0.55;-0.06)	0.010
D4-D2	-1.13	(-1.38;-0.88)	<0.001
D5-D2	-0.92	(-1.14;-0.7)	<0.001
D4-D3	-0.82	(-1.07;-0.57)	<0.001
D5-D3	-0.61	(-0.83;-0.39)	<0.001
D5-D4	0.21	(-0.01;0.43)	0.080

	Difference in mean	95% CI	P-Value
M2-M1	-0.38	(-0.52;-0.23)	<0.001
M3-M1	0.75	(0.57;0.93)	<0.001
M3-M2	1.13	(0.96;1.29)	<0.001

B-scores (mean and 95% CIs) in the different clusters based on DEC (A) and MFAC (B)

Fig. 1. Mean of B-scores by phenotypic cluster.

settings. B-score or any other imaging data was not part of the input variables for cluster definition in this setting:

Two different unsupervised machine-learning based approaches, namely deep embedded clustering (DEC) and multiple factor analysis and clustering (MFAC) were employed to stratify the population. The DEC model used an auto-encoder for dimensionality reduction and a clustering layer for cluster identification while MFAC used a weighted principal component and a hierarchical clustering on principal component for the cluster identification. The DEC model differentiated 5 clusters (D1-5), and with MFAC, 3 clusters were identified (M1-3) [23].

Briefly, based on DEC five clusters (D1-5) were identified. Patients in cluster D1 and D4 were slightly younger than the average age, their overall pain levels were low, while their activity levels were high (based on physical activity scale for the elderly and performance measures). Activity levels differentiated D1 from D4. Patients in D5 were older than

Table 2
Association between laterality of the selected baseline B-score and joint replacement surgery.

	Simultaneous bilateral KR	KR left side	KR right side
Left B-score > right (n; %)	12 (52%)	151 (72%)	39 (19%)
Only left B-score available (n; %)	0	7 (3%)	1 (0.5%)
Right B-score > left (n; %)	11 (48%)	46 (22%)	162 (78%)
Only right B-score available (n; %)	0	5 (2%)	5 (2%)
Total (n)	23	209	207

the average age, had the lowest activity and overall low pain levels. Patients in D2 and D3 had higher pain levels than the other clusters and higher impact on mobility as measured with the 400 m walk test. While D2 was specifically characterized by a high rate of knee joint effusion, D3 stood out based on comorbidities, depression and periarticular pain [23].

The three MFAC clusters (M1-3) separated as follows: M1 comprised predominantly male active patients with low levels of pain. Patients in M2 were older than the average, inactive (based on the lowest average physical activity scale for the elderly score) and reported overall low pain levels. Patients in M3 resembled those in D3 with a high burden of comorbidities, depression, periarticular pain and overall pain levels [23].

Pain levels remained stable for clusters D3/M3 with a trend towards a pain increase in the other clusters. Joint space loss over time was more pronounced in D2 compared to the other clusters [23].

For the current analysis, B-score has been added as input variable. B-score has previously been analyzed in the OAI as a predictor for disease progression [17] and in the current study was assessed for its relationship to the previously described phenotypic clusters. B-score is the statistical z-score that defines the bone shape value of an average non-osteoarthritic femur as 0 at the origin of a shape vector that has a positive value in the direction of increasing OA severity. Along this femur shape vector, a one point change in B-score equals one standard deviation of non-osteoarthritic femur shape [17]. Typically, the spectrum of B-score ranges from −2 to +7. Maximum B-score at baseline (left or right knee) was used in all statistical analyses in order to select the worst B-score and to avoid potential collinearity in the regression models, while maximising the number of observable KRs as outcome events. The model aimed at a prediction of knee joint replacement surgery at the subject level and not at disease progression at the joint level assuming that the less affected knee may also progress to KR in many patients. Descriptive statistics were used to evaluate the side-specific progression.

As an outcomes in this analysis, knee joint replacement surgery was used to capture a non-reversible marker of joint failure defined as total knee replacement (TKR, based on the variables V99ELKDAYS, V99ERKDAYS in the OAI outcomes dataset) or partial medial or lateral knee replacement (PKR, based on the variables V99ELKTLPR, V99ERKTLPR in the OAI outcomes dataset).

The time to onset of the first event of KR was defined as the time from the enrollment date to the first incidence of an event (TKR or PKR in either the left or right knee). In the absence of an event during the follow-up period, the censoring date applied was the earliest of the following dates: date of death, date of withdrawal of informed consent or date of last contact.

2.1. Statistical analysis

The data was summarized using descriptive statistics (quantitative data) and contingency tables (qualitative data). Categorical data was presented as frequencies and percentages. For continuous data, mean (along with 95% confidence interval (CI)), standard deviation, median, 25th and 75th percentiles, minimum and maximum were computed. Spearman correlations were used to assess the association between the left and right B-score at baseline, age and medial joint space width. The correlation between KL and B-score has previously been reported by Bowes et al. [15].

The distributions of baseline B-scores were compared between the clinical phenotypes by multiple (pair-wise) comparisons of the means using Tukey's method. Time to event first KR surgery (TKR or PKR) was presented descriptively using the Kaplan-Meier curve and was summarized by presenting the proportion of patients who are event-free at different time points (2, 5 and 8-years) along with the corresponding 95% CI.

The population was divided into a training set of 70% and a test set of 30% using random sampling and stratified by KR event. Cox proportional hazards regression models (univariable and multivariable) were used to investigate the predictors (baseline B-score and phenotypic clusters as predictor variables) of time to first OA KR surgery and to estimate hazard ratios and 95% CI on the training set. From the model coefficients and the baseline hazard, the cumulative hazard and survival at a specific time point was estimated for a range of B-score values and phenotypic clusters.

The Cox proportional hazards model is described as follows: $h(t|X) = h0(t)e^{\beta X}$.

where h(t) is the hazard rate at time t, h0(t) is the baseline hazard rate at time t, β is a vector of coefficients and *X* is a vector of covariates.

Then the survival curve of the Cox model is: $S(t|X) = S0(t)^{\exp(\beta X)}$

The likelihood ratio test was applied to multivariable Cox proportional hazard models and was used to test the added prognostic value of Cluster over B-score and vice-versa in the training set. Predictive performance of the Cox regression models was determined by time area under the receiver operating characteristic curve (AUC) metrics and discrimination by C-index using the test set. One thousand iteration resamples on the training and test data were performed to estimate the 95% CI for C-index.

No imputation of missing baseline B-score was performed.

All statistical computations were performed in R version 4.1.0 (2021-05-18), R Core Team (2021) using RStudio version 2022.07.3 + 585.pro1 environment RStudio Team (2021).

3. Results

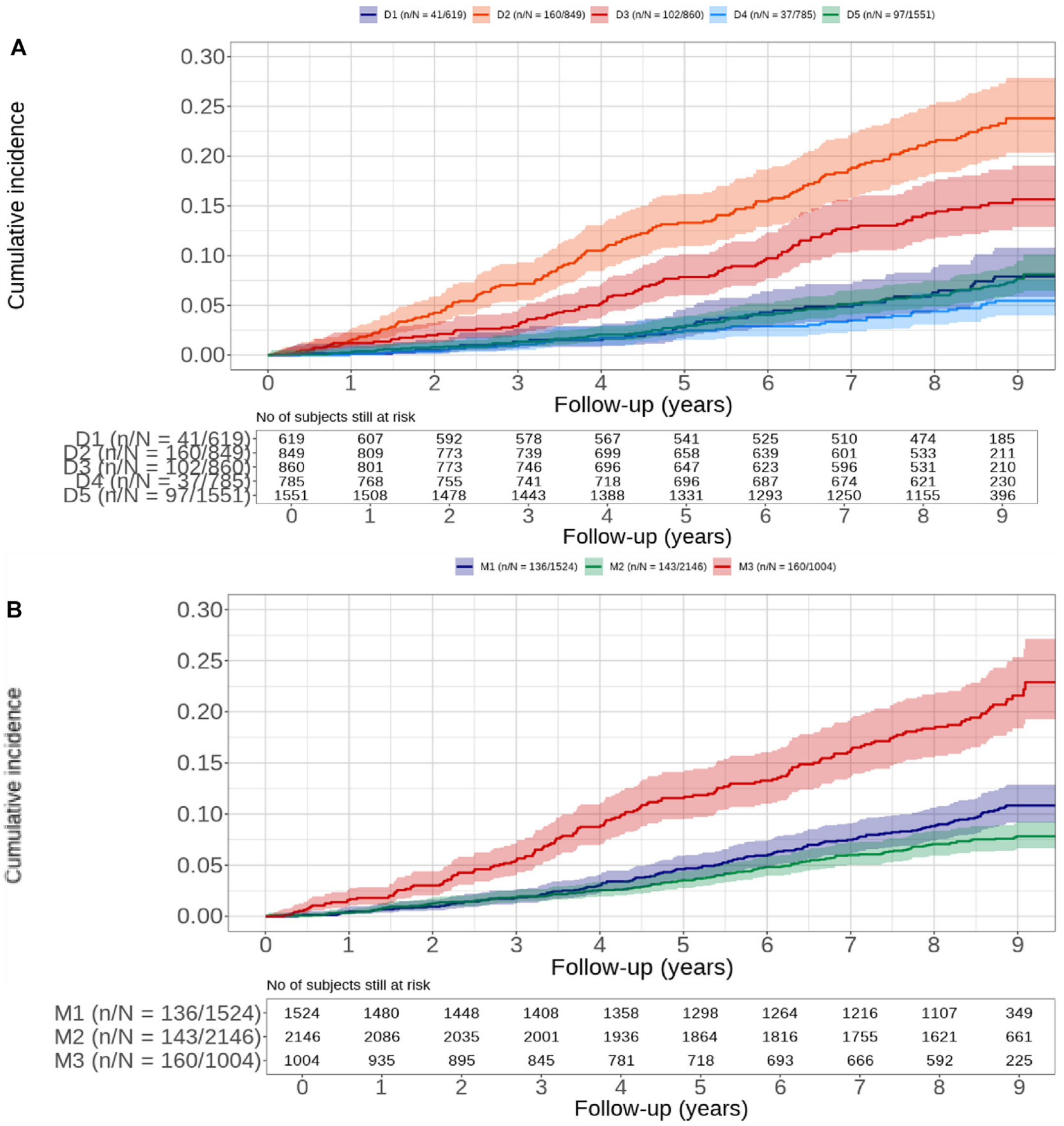
3.1. Relationship between bone shape and clinical clusters

Table 1 (and Supplemental Figure 1) summarize descriptively the distribution of B-scores in the incidence and progression cohorts of the OAI database. As expected, B-score was overall lower in the incidence cohort with a mean close to normal (between 0 and 1) and similar distributions for the right and left knees. There was an expected overlap of B-scores for the different KL groups, with KL grades spanning B-scores from −3 to +9 in these two cohorts (Supplemental Figure 2). Using a Spearman's rank correlation coefficient approach, we observed a negative weak association (Spearman's ρ = −0.26) between B-scores and medial JSW. Increasing KL grade was associated with increasing mean B-scores (Supplemental Figure 5), however, we observed a high overlap of KL-associated B-scores especially for KL grades 1–3 (Supplemental Figure 2).

There was a high correlation (Spearman's ρ = 0.77) between left and right knee B-scores (Supplemental Figure 3), but a weak correlation (Spearman's ρ = 0.10) with age (Supplemental Figure 4). B-score significantly differed between the different clusters as shown in Fig. 1.

3.2. Relationship to joint replacement

Knee joint replacement surgery was observed in 439 (402 TKR, 37 PKR) of 4674 patients. In 75–80% of cases, the knee with the higher baseline B-score progressed to surgery (see Table 2). The cumulative incidence of KR events differed significantly between the clusters as shown in Fig. 2. In a multivariable proportional hazards regression model using the training set and adjusted for B-score, the phenotypic cluster allocation to D2/D3 and M3 remained independent prognostic



C

Clusters	D1	D2	D3	D4	D5	M1	M2	M3
Number of events	41 (6.62%)	160 (18.85%)	102 (11.86%)	37 (4.71%)	97 (6.25%)	136 (8.92%)	143 (6.66%)	160 (15.94%)
Kaplan-Meier estimates (%) rate [95% CI] at 2 years	0.5 [0.2, 1.5]	4.2 [3, 5.8]	2 [1.2, 3.2]	0.4 [0.1, 1.2]	0.8 [0.5, 1.4]	0.9 [0.6, 1.6]	1.3 [0.9, 1.8]	3 [2.1, 4.3]
Kaplan-Meier estimates (%) rate [95% CI] at 5 years	2.8 [1.7, 4.5]	12.4 [10.3, 15]	7.5 [5.9, 9.7]	2.4 [1.5, 3.8]	2.8 [2.1, 3.8]	4.5 [3.5, 5.7]	3.5 [2.7, 4.4]	10.9 [9, 13.2]
Kaplan-Meier estimates (%) rate [95% CI] at 8 years	6.1 [4.4, 8.4]	19.3 [16.7, 22.3]	13.3 [11, 16]	4.3 [3, 6.1]	5.8 [4.7, 7.2]	8.4 [7.1, 10]	6.8 [5.8, 8]	16.8 [14.4, 19.5]

Kaplan-Meier curves for the incidence of joint replacement surgery per cluster based on A=DEC and B=MFAC, C shown the analysis of time to first knee replacement events by cluster using the Kaplan-Meier method

Fig. 2. Cumulative incidence of first KR events among OAI participants by cluster and time to KR based on a Kaplan-Meier approach.

factors for KR (Table 3), confirming the independent prognostic value of both cluster and B-score.

Eight-year KR estimates were 4.3% and 6.8% for D4 and M2 respectively versus 19.3% and 16.8% for D2 and M3 (Panel C Fig. 2). Based on the multivariate Cox regression model, the risk for OA KR events at 8 years is 28% for D2, 21% for D3 and 23% for M3 for a baseline B-score at 4 (Fig. 3).

The predictive value of cluster allocation was evaluated relative to the B-score as a predictor of KR using the likelihood ratio test for the multivariate Cox model using the training set. Adding phenotypic cluster information to the B-score significantly improved the predictive value ($P < 0.001$ for DEC clusters and $P = 0.04$ for MFA clusters; Table 4). Conversely, the B-score added significantly greater predictive information to cluster ($P < 0.001$). The area under the curve over time in the training and test sets and Harrel C-index were used to compare the performance of the Cox regression models (Fig. 4). Combining B-score and cluster (MFAC or DEC) in the Cox regression model, provided similar performances to the model with only B-score but greater performances than the model with cluster, as the C-statistic [95%CI] for the combined model was 0.76 [0.72, 0.80] versus 0.75 [0.71, 0.79]

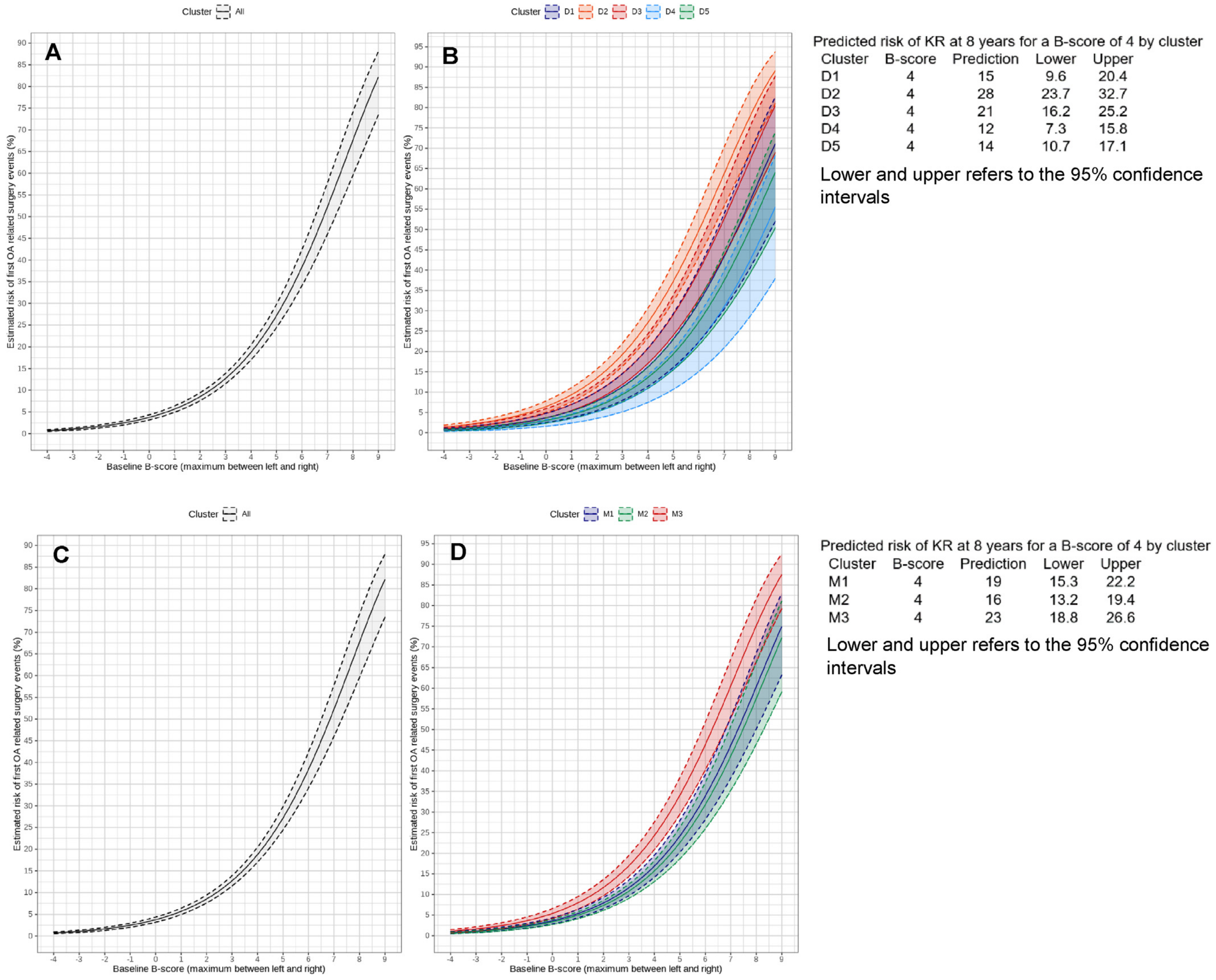
Table 3
Association between B-score, clusters and joint replacement surgery events in the training set.

		all	HR (univariable)	HR (multivariable)		all	HR (univariable)	HR (multivariable)
Cluster	**D4**	556 (17.0)	–	–	**M2**	1531 (46.8)	–	–
	D5	1085 (33.3)	1.34 (0.86–2.10, p = 0.197)	1.21 (0.77–1.89, p = 0.400)	**M1**	1046 (32.0)	1.42 (1.08–1.88, p = 0.012)	1.17 (0.88–1.54, p = 0.280)
	D1	435 (13.3)	1.31 (0.77–2.23, p = 0.321)	1.33 (0.78–2.26, p = 0.300)	**M3**	693 (21.2)	2.68 (2.04–3.51, p < 0.001)	1.45 (1.09–1.93, p = 0.011)
	D3	581 (17.8)	2.87 (1.83–4.48, p < 0.001)	1.88 (1.20–2.96, p = 0.006)				
	D2	605 (18.5)	4.54 (2.98–6.90, p < 0.001)	2.69 (1.76–4.11, p < 0.001)				
B-score	**Mean (SD)**	1.4 (1.9)	1.55 (1.48–1.62, p < 0.001)	1.50 (1.43–1.57, p < 0.001)		1.4 (1.9)	1.55 (1.48–1.62, p < 0.001)	1.52 (1.46–1.60, p < 0.001)

for B-score alone, and the area under the time-dependent receiver operating characteristic curve was superior to models with B-score or cluster alone. The area under the curve range between 0.8 (0.84) and 0.72 (0.73) from 6 months to 8 years when combining B-score and MFAC (or DEC) cluster versus 0.76 and 0.71 for B-score alone in the test set (Fig. 4). Among D5 and M2 clusters a trend for better performances of the B-score model was observed as the C-index in the test set were 0.78 [0.70, 0.86] and 0.81 [0.75, 0.87] respectively.

4. Discussion

These analyses demonstrate significant differences in B-scores between most of the pre-determined clusters based on clinical information, thereby associating clinical phenotypes with a structural biomarker. In addition, B-score and cluster allocation were found to be independent predictors of knee joint replacement surgery, while the combination of both increased the predictive value. B-score was the stronger predictor



Predicted risk of KR at 8 years for a B-score of 4 by cluster

Cluster	B-score	Prediction	Lower	Upper
D1	4	15	9.6	20.4
D2	4	28	23.7	32.7
D3	4	21	16.2	25.2
D4	4	12	7.3	15.8
D5	4	14	10.7	17.1

Lower and upper refers to the 95% confidence intervals

Predicted risk of KR at 8 years for a B-score of 4 by cluster

Cluster	B-score	Prediction	Lower	Upper
M1	4	19	15.3	22.2
M2	4	16	13.2	19.4
M3	4	23	18.8	26.6

Lower and upper refers to the 95% confidence intervals

Fig. 3. Predicted 8 years risk of first KR events versus baseline risk score overall and by cluster.

Table 4
Likelihood ratio (LR) analysis of Cluster and B-score for first joint replacement surgery events in the training set.

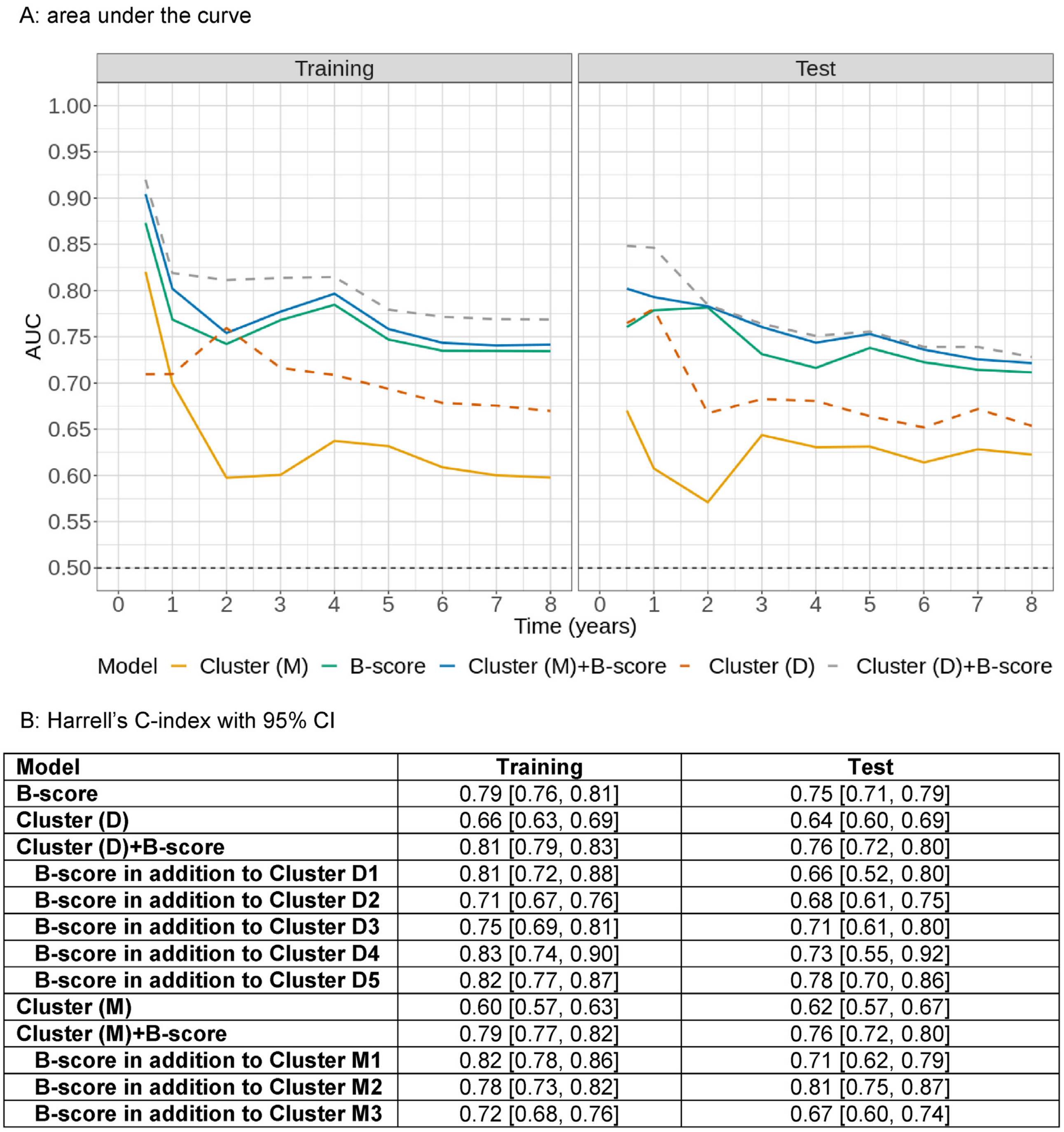
Cluster	N	Model 1	Model 2	LR Chi square	p-value
DEC (D)	3258	B-score	B-score + Cluster	38.5	< 0.001
		Cluster	Cluster + B-score	258.6	< 0.001
MFAC (M)	3266	B-score	B-score + Cluster	6.4	0.040
		Cluster	Cluster + B-score	278.9	< 0.001

Model 1 uses either B-score or cluster as basis for prediction, model 2 adds the respective second feature.

compared to cluster allocation.

Overlapping B-scores for KL-grades and weak JSW correlation confirm the difficulty of using radiographic evidence to define OA disease state. This was also demonstrated by the nearly normal distribution of B-scores in the incidence cohort contrasted with the skewed distributions in the progression cohort which reflects the heterogeneous nature of OA as a disease.

Age is a known risk factor for OA, in consequence changes in bone shape and therefore B-score may partly be driven by age. Indeed, a natural progression of bone shape has been described previously,



B: Harrell's C-index with 95% CI

Model	Training	Test
B-score	0.79 [0.76, 0.81]	0.75 [0.71, 0.79]
Cluster (D)	0.66 [0.63, 0.69]	0.64 [0.60, 0.69]
Cluster (D)+B-score	0.81 [0.79, 0.83]	0.76 [0.72, 0.80]
B-score in addition to Cluster D1	0.81 [0.72, 0.88]	0.66 [0.52, 0.80]
B-score in addition to Cluster D2	0.71 [0.67, 0.76]	0.68 [0.61, 0.75]
B-score in addition to Cluster D3	0.75 [0.69, 0.81]	0.71 [0.61, 0.80]
B-score in addition to Cluster D4	0.83 [0.74, 0.90]	0.73 [0.55, 0.92]
B-score in addition to Cluster D5	0.82 [0.77, 0.87]	0.78 [0.70, 0.86]
Cluster (M)	0.60 [0.57, 0.63]	0.62 [0.57, 0.67]
Cluster (M)+B-score	0.79 [0.77, 0.82]	0.76 [0.72, 0.80]
B-score in addition to Cluster M1	0.82 [0.78, 0.86]	0.71 [0.62, 0.79]
B-score in addition to Cluster M2	0.78 [0.73, 0.82]	0.81 [0.75, 0.87]
B-score in addition to Cluster M3	0.72 [0.68, 0.76]	0.67 [0.60, 0.74]

D refers to cluster information derived using DEC, M refers to cluster information derived using MFAC. The respective models use all clustering information

Fig. 4. Comparison of prediction performances across models.

discriminating patients with progression of OA from those with more physiologic age-related changes [20]. In this study though, the overall correlation between age and bone shape was weak.

The validity of a predictive or prognostic marker is ultimately related to the number of events of interest, and their reversibility. Kim et al. have suggested a composite endpoint for OA clinical trials based on "time to TKR or severe pain and/or severely impaired functioning" [27]. One challenge with this approach for clinical trial design may be the low incidence of these non-reversible events in a non-enriched cohort (for example, 138 TKRs in 1332 participants in the OAI progression cohort). Another problem may be the reversibility of pain or function outcomes in view of the fluctuating course of OA and the fact that pain is often seen to "flare" for a period of time [28]. While in a clinical setting with close patient-provider relationship these aspects can be validly assessed, registry or trial data with longer visit intervals make it difficult to differentiate between continuous symptom progression and fluctuation. Bone shape provides a biomarker for predicting disease progression toward joint replacement, an objective endpoint, which could qualify B-score as a surrogate marker for joint survival in future.

Bone shape has previously been demonstrated to be associated with the risk of TKR in the OAI database [15]. The concordance between B-score and the described phenotypes underlines the concept that phenotypes may be related to endotypes detectable with individual biomarkers. In addition, the predictive value of both phenotype and B-score may facilitate their use in routine clinical practice for prognostic estimates and patient information depending on the availability of information. From this analysis, the prediction for TKR improves from addition of B-score especially for the clinical clusters D5 and M2. Interestingly those clusters have the lowest pain levels of all clusters (KOOS pain average > 88/100) and only 11–16% are categorized as having a KL grade of 3–4 [23]. This may suggest that in these clusters the disease severity is underestimated based on clinical information and x-ray alone, so that the addition of B-score improves the predictive performance of the model.

Bone shape changes have also been shown to be reactive to trauma such as ACL injury or ACL surgery [25]. These changes may be triggered by inflammatory processes involved in musculoskeletal injury [29–31]. It might therefore also be expected that the potential inflammatory cluster D2 shows a faster progression than other clusters, and this was indeed observed in this evaluation. The relatively larger number of patients proceeding to KR in clusters M3 and D3 (both of which demonstrate increased comorbidities, depression and pain) is surprising since only 20–30% of participants have OA KL grades 3–4 at baseline. This fact may be pointing to an underestimation of disease severity by KL, or that subjects exhibiting increased pain tend to have TKR at an earlier stage of structural OA. This may also point to a third problem regarding the composite endpoint suggested by Kim et al., the fact that the timing of a TKR may be mediated by the general condition of the patient, and how that influences the surgeon's and patient's decision on whether surgery is appropriate.

5. Conclusion

B-scores in the OAI database were related to OA phenotypes based on clinical information, linking imaging structural changes to clinical patient profiles in this population. Given the independent predictive value for KR of both B-score and phenotypes, they can individually or in combination serve in clinical OA trial to enrich for patients likely to progress to KR, or contribute to risk prediction in clinical practice. The predictive value for KR highlights B-score as a potential surrogate marker for joint survival, especially in patients with less severe clinical presentation.

6. Limitations

The study describes a predictive model for KR in knee OA based on clinical variables and an imaging biomarker. The model has not been validated yet using a large independent test set. Validation in independent data sets is a clear prerequisite to ensure generalizability. However, the availability of comparable variables and access to large data sets can be challenging.

Author contributions

All authors have been involved in conception and design of the study, the analysis of the data was driven by DD, all authors contributed to the interpretation of data. FS and DD primarily drafted the manuscript, which was critically reviewed and approved by all authors.

Source of data and code

The clinical data, patient reported outcomes and additional analyses are publicly available after registration from https://nda.nih.gov/oai/ (access October 18th, 2022). The source code for DEC is available from https://github.com/piiswrong/dec (access October 24th, 2022) adapted to facilitate working with the Keras Package instead of the Caffe Package as reported by Xie et al. [32], for MFA refer to Le et al., 2008 [33].

Role of the funding source

The analysis was founded by the Novartis Institutes of Biomedical Research (BASICHR0042) under a research collaboration agreement with iMorphics who shared the results of their bone shape analysis in the OAI. The Funder had no influence on the study design, data interpretation or publication strategy.

Competing interests

Franziska Saxer is employee and shareholder of Novartis, she is affiliated to the University Basel and member of the European Union Medical Devices - Expert Panel section Orthopaedics, traumatology, rehabilitation, rheumatology.

David Demanse is employee and shareholder of Novartis.

Alan Brett is an employee and shareholder of Stryker.

Didier Laurent is employee and shareholder of Novartis.

Linda Mindeholm is consultant to Novartis and Versanis Bio and shareholder of Novartis.

Philip G Conaghan reports consultancies or speakers bureaus for AbbVie, EliLilly, Genascense, GlaxoSmithKline, Grunenthal, Janssen, Levicept, Merck, Moebius, Novartis, Stryker, Takeda and TrialSpark.

Matthias Schieker is employee and shareholder of Novartis, he is affiliated to LMU Munich and owner of LivImplant GmbH.

Acknowledgements

The authors like to thank the participants, investigators and funders of the OAI database, a public–private partnership comprising five contracts (N01-AR-2–2258; N01-AR-2–2259; N01-AR-2–2260; N01-AR-2–2261; N01-AR-2–2262) funded by the NIH, and conducted by the OAI Study Investigators. Data and/or research tools used in the preparation of this manuscript were obtained and analyzed from the controlled access datasets distributed from the Osteoarthritis Initiative (OAI), a data repository housed within the NIMH Data Archive (NDA).

The authors also like to thank Inga Ludwig for her critical review of the draft manuscript.

Philip G Conaghan is supported in part through the NIHR Leeds Biomedical Research Centre NIHR203331. The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ocarto.2024.100458.

References

[1] Collaborative-Global Burden of Disease Network. Global Burden of Disease Study 2019 results. vol. 2022. https://vizhub.healthdata.org/gbd-results/2020.

[2] D.J. Hunter, D. Schofield, E. Callander, The individual and socioeconomic impact of osteoarthritis, Nat. Rev. Rheumatol. 10 (2014) 437–441.

[3] I.L. Araujo, M.C. Castro, C. Daltro, M.A. Matos, Quality of life and functional independence in patients with osteoarthritis of the knee, Knee Surg Relat Res 28 (2016) 219–224.

[4] E.K. Nihtila, P.T. Martikainen, S.V. Koskinen, A.R. Reunanen, A.M. Noro, U.T. Hakkinen, Chronic conditions and the risk of long-term institutionalization among older people, Eur. J. Publ. Health 18 (2008) 77–84.

[5] E. Nuesch, P. Dieppe, S. Reichenbach, S. Williams, S. Iff, P. Juni, All cause and disease specific mortality in patients with knee or hip osteoarthritis: population based cohort study, BMJ 342 (2011) d1165.

[6] H.J. Kellgren, J.S. Lawrence, Radiological assessment of osteoarthrosis, Ann. Rheum. Dis. 16 (1957) 494.

[7] Y. Kim, G. Levin, N.P. Nikolov, R. Abugov, R. Rothwell, Concept end points informing design considerations for confirmatory clinical trials in osteoarthritis, Arthritis Care Res (Hoboken) 74 (2022) 1154–1162.

[8] L.A. Deveza, R.F. Loeser, Is osteoarthritis one disease or a collection of many? Rheumatology (Oxford) 57 (2018) iv34–iv42.

[9] A. Mobasheri, S. Saarakkala, M. Finnila, M.A. Karsdal, A.C. Bay-Jensen, W.E. van Spil, Recent advances in understanding the phenotypes of osteoarthritis, F1000Res 8 (2019).

[10] A. Mobasheri, W.E. van Spil, E. Budd, I. Uzieliene, E. Bernotiene, A.C. Bay-Jensen, et al., Molecular taxonomy of osteoarthritis for patient stratification, disease management and drug development: biochemical markers associated with emerging clinical phenotypes and molecular endotypes, Curr. Opin. Rheumatol. 31 (2019) 80–89.

[11] A. Guermazi, F.W. Roemer, D.T. Felson, K.D. Brandt, Motion for debate: osteoarthritis clinical trials have not identified efficacious therapies because traditional imaging outcome measures are inadequate, Arthritis Rheum. 65 (2013) 2748–2758.

[12] B. Antony, A. Singh, Imaging and biochemical markers for osteoarthritis, Diagnostics (Basel) 11 (2021).

[13] D.J. Hunter, J.E. Collins, L. Deveza, S.C. Hoffmann, V.B. Kraus, Biomarkers in osteoarthritis: current status and outlook - the FNIH Biomarkers Consortium PROGRESS OA study, Skeletal Radiol 52 (11) (2023 Nov) 2323–2339.

[14] A. Brett, M.A. Bowes, P.G. Conaghan, Comparison of 3D quantitative osteoarthritis imaging biomarkers from paired CT and MR images: data from the IMI-APPROACH study, BMC Muscoskel. Disord. 24 (2023) 76.

[15] M.A. Bowes, K. Kacena, O.A. Alabas, A.D. Brett, B. Dube, N. Bodick, et al., Machine-learning, MRI bone shape and important clinical outcomes in osteoarthritis: data from the Osteoarthritis Initiative, Ann Rheum Dis. 80 (4) (2021 Apr) 502–508.

[16] T. Neogi, M.A. Bowes, J. Niu, K.M. De Souza, G.R. Vincent, J. Goggins, et al., Magnetic resonance imaging-based three-dimensional bone shape of the knee predicts onset of knee osteoarthritis: data from the osteoarthritis initiative, Arthritis Rheum. 65 (2013) 2048–2058.

[17] M.A. Bowes, G.R. Vincent, C.B. Wolstenholme, P.G. Conaghan, A novel method for bone area measurement provides new insights into osteoarthritis and its progression, Ann. Rheum. Dis. 74 (2015) 519–525.

[18] A.J. Barr, B. Dube, E.M. Hensor, S.R. Kingsbury, G. Peat, M.A. Bowes, et al., The relationship between three-dimensional knee MRI bone shape and total knee replacement-a case control study: data from the Osteoarthritis Initiative, Rheumatology (Oxford) 55 (2016) 1585–1593.

[19] P.G. Conaghan, M.A. Bowes, S.R. Kingsbury, A. Brett, G. Guillard, B. Rizoska, et al., Disease-modifying effects of a novel Cathepsin K inhibitor in osteoarthritis: a randomized controlled trial, Ann. Intern. Med. 172 (2020) 86–95.

[20] D. McGuire, M. Bowes, A. Brett, N.A. Segal, M. Miller, D. Rosen, et al., Study TPX-100-5: intra-articular TPX-100 significantly delays pathological bone shape change and stabilizes cartilage in moderate to severe bilateral knee OA, Arthritis Res. Ther. 23 (2021) 242.

[21] G. Sakellariou, P.G. Conaghan, W. Zhang, J.W.J. Bijlsma, P. Boyesen, M.A. D'Agostino, et al., EULAR recommendations for the use of imaging in the clinical management of peripheral joint osteoarthritis, Ann. Rheum. Dis. 76 (2017) 1484–1494.

[22] 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.

[23] D. Demanse, F. Saxer, P. Lustenberger, L.B. Tankó, P. Nikolaus, I. Rasin, et al., Unsupervised machine-learning algorithms for the identification of clinical phenotypes in the Osteoarthritis Initiative database, Semin. Arthritis Rheum. 58 (2023).

[24] M. Bowes, C.B. Wolstenholme, G.R. Vincent, P.G. Conaghan, Bone shape does not cause OA – but OA does change bone shape; a study on data from 4654 knees from the osteoarthritis initiativeOARSI Annual Congress, Ann Rheum Dis 74 (3) (2015 Mar), 519-252016.

[25] M.A. Bowes, L.S. Lohmander, C.B.H. Wolstenholme, G.R. Vincent, P.G. Conaghan, R.B. Frobell, Marked and rapid change of bone shape in acutely ACL injured knees – an exploratory analysis of the Kanon trial, Osteoarthritis Cartilage 27 (2019) 638–645.

[26] M.C. Nevitt, D.T. Felson, G. Lester, The osteoarthritis initiative protocol for the cohort study, in: Diseases NIoAaMaS, 2006.

[27] Y. Kim, G. Levin, N.P. Nikolov, R. Abugov, R. Rothwell, Concept endpoints informing design considerations for confirmatory clinical trials in osteoarthritis, Arthritis Care Res (Hoboken) 74 (7) (2022 Jul) 1154–1162.

[28] M. Englund, A. Turkiewicz, Pain in clinical trials for knee osteoarthritis: estimation of regression to the mean, The Lancet Rheumatol 5 (6) (2023 Jun) e309–e311.

[29] O. Reikeras, P. Borgen, Activation of markers of inflammation, coagulation and fibrinolysis in musculoskeletal trauma, PLoS One 9 (2014) e107881.

[30] B.H.Y. Gibson, M.T. Duvernay, S.N. Moore-Lotridge, M.J. Flick, J.G. Schoenecker, Plasminogen activation in the musculoskeletal acute phase response: injury, repair, and disease, Res Pract Thromb Haemost 4 (2020) 469–480.

[31] J.W. MacKay, L. Watkins, G. Gold, F. Kogan, [(18)F]NaF PET-MRI provides direct in-vivo evidence of the association between bone metabolic activity and adjacent synovitis in knee osteoarthritis: a cross-sectional study, Osteoarthritis Cartilage 29 (2021) 1155–1162.

[32] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: B. Maria Florina, Q.W. Kilian (Eds.), Proceedings of the 33rd International Conference on Machine Learning, vol. 48, Proceedings of Machine Learning Research, 2016, pp. 478–487. PMLR.

[33] S. Le, J. Josse, F. Husson, FactoMineR: an R package for multivariate analysis, J. Stat. Software 25 (2008) 1–18.