RESEARCH ARTICLE

Towards novel osteoarthritis biomarkers: Multi-criteria evaluation of 46,996 segmented knee MRI data from the Osteoarthritis Initiative

Alexander Tack[1]*, **Felix Ambellan**[1], **Stefan Zachow**[1,2]

1 Zuse Institute Berlin, Berlin, Germany, **2** Charité – Universitätsmedizin Berlin, Berlin, Germany

* tack@zib.de

Abstract

Convolutional neural networks (CNNs) are the state-of-the-art for automated assessment of knee osteoarthritis (KOA) from medical image data. However, these methods lack interpretability, mainly focus on image texture, and cannot completely grasp the analyzed anatomies' shapes. In this study we assess the informative value of quantitative features derived from segmentations in order to assess their potential as an alternative or extension to CNN-based approaches regarding multiple aspects of KOA. Six anatomical structures around the knee (femoral and tibial bones, femoral and tibial cartilages, and both menisci) are segmented in 46,996 MRI scans. Based on these segmentations, quantitative features are computed, i.e., measurements such as cartilage volume, meniscal extrusion and tibial coverage, as well as geometric features based on a statistical shape encoding of the anatomies. The feature quality is assessed by investigating their association to the Kellgren-Lawrence grade (KLG), joint space narrowing (JSN), incident KOA, and total knee replacement (TKR). Using gold standard labels from the Osteoarthritis Initiative database the balanced accuracy (BA), the area under the Receiver Operating Characteristic curve (AUC), and weighted kappa statistics are evaluated. Features based on shape encodings of femur, tibia, and menisci *plus* the performed measurements showed most potential as KOA biomarkers. Differentiation between non-arthritic and severely arthritic knees yielded BAs of up to 99%, 84% were achieved for diagnosis of early KOA. Weighted kappa values of 0.73, 0.72, and 0.78 were achieved for classification of the grade of medial JSN, lateral JSN, and KLG, respectively. The AUC was 0.61 and 0.76 for prediction of incident KOA and TKR within one year, respectively. Quantitative features from automated segmentations provide novel biomarkers for KLG and JSN classification and show potential for incident KOA and TKR prediction. The validity of these features should be further evaluated, especially as extensions of CNN-based approaches. To foster such developments we make all segmentations publicly available together with this publication.

Citation: Tack A, Ambellan F, Zachow S (2021) Towards novel osteoarthritis biomarkers: Multi-criteria evaluation of 46,996 segmented knee MRI data from the Osteoarthritis Initiative. PLoS ONE 16(10): e0258855. https://doi.org/10.1371/journal.pone.0258855

Editor: Yuanyuan Wang, Monash University, AUSTRALIA

Received: April 19, 2021

Accepted: October 6, 2021

Published: October 21, 2021

Data Availability Statement: All MRI datasets are publicly available via the Osteoarthritis Initiative Database (https://nda.nih.gov/oai/). All 46'996 segmentation masks generated in this study will be made publicly available together with this publication at https://pubdata.zib.de.

Funding: The German Federal Ministry of Education and Research (BMBF) (https://www.bmbf.de/) research network on musculoskeletal diseases supported AT (grant no. 01EC1408B, Overload/PrevOP) and FA (grant no. 01EC1406E, TOKMIS). AT was further supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) (https://www.dfg.de/) research campus MODAL (grant no. 3FO18501) and DFG research project ZA 592/4-1. FA was further supported by the DFG under Germany's Excellence Strategy - The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689). SZ reports funding by the DFG cluster of excellence Cluster of Excellence "Matters of Activity. Image Space Material" under Germany's Excellence Strategy – EXC 2025 – 390648296. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Medical imaging has become the standard diagnostic means for assessing osteoarthritis. Substantial efforts have been made in the past decades to identify image-based biomarkers and to develop methods for image-based assessment of knee osteoarthritis (KOA) from conventional radiographs and tomographic image data. To rate KOA from X-Rays with the knee being in a load-bearing situation, the current gold standard is the Kellgren-Lawrence grading (KLG) [1], where e.g. radiographic joint narrowing (JSN) is measured. To three-dimensionally (3-D) assess an arthritic anatomy, 3-D imaging methods are compulsory [2] (or at least reliable 3-D reconstruction methods from 2-D images [3]). Compared to computed tomography, magnetic resonance imaging (MRI) offers the advantages of no radiation exposure and significantly better differentiation of soft tissues. Various procedures have been proposed for KOA diagnostics from MRI data, such as manual image reading based on semi-quantitative scoring systems [4, 5], computerized quantitative analysis based on manual definitions of regions of interest (ROI) [6–10], up to fully automated methods based on machine learning [11, 12].

In order to gain more insight into the pathogenesis of osteoarthritis and the underlying phenotypes, a big amount of data needs to be studied. Time efficient processing of thousands of MRI scans, however, seems to be feasible only by employing automated methods, which additionally could give rise to a more objective and holistic support of KOA scoring by incorporating multiple data sources. Promising results in view of automated processing were achieved recently using deep learning. However, most of these approaches were designed to perform a single task only, e.g. diagnosis of cartilage degeneration [13, 14] or meniscal lesions [15–17] in MRI data. Methods of deep learning are complex to design, their decision process is hardly explainable, and huge burdens need to be overcome with respect to a generalizability for different imaging modalities [18].

In this study, we evaluate various KOA aspects using different quantitative characteristics of individual structures of the knee in order to investigate their potential as biomarkers. Therefore, we apply an automated segmentation approach for six anatomical structures (femoral and tibial bones, femoral and tibial cartilages, and both menisci) to almost all subjects contained in the Osteoarthritis Initiative (OAI) database (https://nda.nih.gov/oai) and perform thorough quality assurance of our segmentation results. We investigate which features show highest potential for a holistic assessment of KOA by classification of KLG and JSN as well as by predicting a possible occurrence of incident KOA or the need for a total knee replacement (TKR) within a time frame of up to five years. With this work we aim to set a basis for future developments within KOA research and diagnosis by supplementing the OAI database with the segmented structures (https://pubdata.zib.de).

Materials and methods

Study population

This study is based on 3D sagittal Double Echo Steady-State (DESS) MRI data acquired by the OAI using Siemens Trio 3.0 Tesla scanners [19]. A total of 48,073 datasets are available (see Table 1), which split up into 7 time points: baseline visit (v00, MRI data of 9,494 knees), 1 year follow-up (v12, 8,187 scans), 2 year follow-up (v24, 7,534 scans), 3 year follow-up (v36, 5,604 scans), 4 year follow-up (v48, 6,743 scans), 6 year follow-up (v72, 5,508 scans), and 8 year follow-up (v96, 5,003 scans).

All datasets of the retrospective OAI database study are fully anonymized. Ethics approval for the OAI database was obtained by the OAI coordinating center and by each OAI clinical

Table 1. Demographics of data used in this study summarized for all considered time points of the OAI database.

Visit	# Images	Side (left, right)	Sex (male, female)	Age [years]	BMI [kg/m^2]	KLG (0, 1, 2, 3, 4, NA)	mJSN (0, 1, 2, 3, NA)	lJSN (0, 1, 2, 3, NA)
v00	9345	4639, 4706	3847, 5498	61.09 ± 9.18	28.59 ± 4.83	3404, 1565, 2329, 1203, 284, 560	5699, 1909, 978, 199, 560	8089, 361, 248, 87, 560
v12	8025	4006, 4019	3347, 4678	62.13 ± 9.13	28.44 ± 4.79	2954, 1371, 2103, 1130, 328, 139	5054, 1699, 905, 228, 139	7232, 306, 247, 101, 139
v24	7338	3660, 3678	3114, 4224	62.93 ± 9.09	28.39 ± 4.84	2681, 1246, 1914, 1054, 331, 112	4611, 1545, 835, 234, 113	6611, 279, 238, 97, 113
v36	5500	2033, 3467	2356, 3144	63.64 ± 9.06	28.38 ± 4.80	1973, 910, 1429, 835, 260, 93	3414, 1149, 658, 186, 93	4932, 207, 192, 76, 93
v48	6616	3276, 3340	2819, 3797	64.68 ± 9.06	28.45 ± 4.88	2357, 1072, 1690, 924, 332, 241	4094, 1313, 732, 236, 241	5824, 240, 214, 97, 241
v72	5413	2692, 2721	2317, 3096	66.07 ± 8.82	28.25 ± 4.94	1692, 886, 421, 176, 23, 2215	2544, 499, 139, 16, 2215	3034, 113, 43, 8, 2215
v96	4759	2305, 2454	2055, 2704	67.56 ± 8.64	28.38 ± 5.02	1595, 807, 407, 208, 38, 1704	2406, 455, 166, 28, 1704	2887, 115, 44, 10, 1703

NA: 'Not available'; no measurement was performed within the OAI study for these knees.

https://doi.org/10.1371/journal.pone.0258855.t001

site. All patients provided written informed consent for participation in the OAI and to have their data from their medical records used in research.

Automatic segmentation of MRI data

The anatomical structures most affected by KOA are the bones, cartilages and menisci of the knee. To compute features that might be suitable for being used as biomarkers for any or each of these structures individually we first employ methods of image segmentation to specify the anatomical ROIs (see Fig 1). We utilize the method of Ambellan et al. 2019 [20], to segment the distal femoral bone (FB) and the proximal tibial bone (TB) as well as the femoral and tibial cartilage (FC, TC). In addition, the method of Tack et al. 2018 [21], is utilized to segment the medial and lateral meniscus (mM, lM).

Our fully automatic segmentation method has a run time of approx. 10 minutes per MRI dataset on a common workstation. This means the segmentation would almost take a year for all 48,073 datasets on a single machine without any parallelization. To carry out this massive segmentation effort, we employed a combination of high performance computing (HLRN https://www.hlrn.de/supercomputer-e/hlrn-iv-system/?lang=en) and GPU-based application

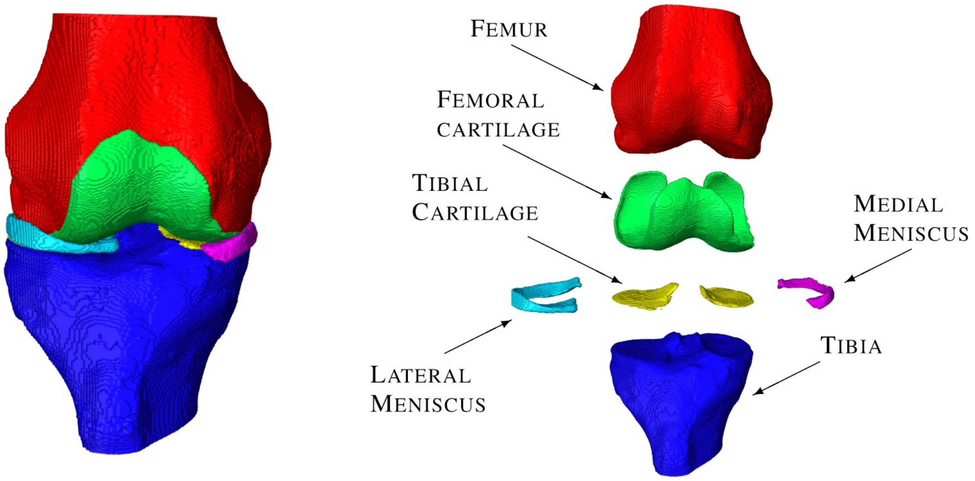

Fig 1. Anatomical regions of interest resulting from the automated segmentation methods. Our automated segmentation methods yield segmentation masks for the femoral bone, tibial bone, femoral and tibial cartilage, and both menisci.

https://doi.org/10.1371/journal.pone.0258855.g001

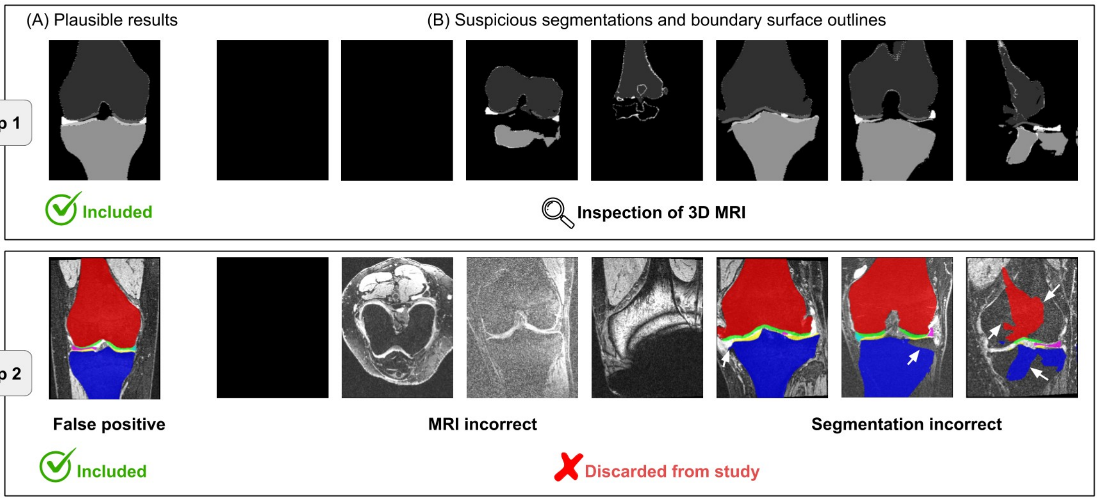

Fig 2. Quality assurance of automated segmentations. Step 1: Segmentation results (outlines) are shown for coronal slices of the MRI. Plausible results (A) are included in this study. For suspicious results (B) the complete 3D MRI is inspected (Step 2). Examples for discarded cases are (from left to right): MRI can not be loaded, wrong MRI orientation, doubtful MRI appearance, MRI artifacts, incorrect menisci segmentation, incorrect bone segmentation, incorrect overall segmentation (incorrect segmentations indicated by white arrows). 1,850 cases which appeared suspicious in Step 1 were identified as false positives in Step 2 and remained included in this study.

https://doi.org/10.1371/journal.pone.0258855.g002

of machine learning (NVIDIA DGX-1 system (NVIDIA Corporation, California, United States), provided by the Max Delbrück Center for Molecular Medicine in Berlin, Germany (https://www.mdc-berlin.de/), which allowed us to segment 48,073 MRI datasets in about 5 weeks.

Data: Inclusion and exclusion criteria

The employed methods allowed us to segment several thousand datasets in a fully automated fashion. Our motivation is to include only datasets having a plausible segmentation in the analysis of features to test their possible suitability as KOA biomarkers. To ensure a sufficient quality for such an analysis, we visually inspected each segmented dataset—a process that took roughly 90 person hours. We empirically found coronal slice number 190 (± 5) to typically show all structures of interest since the field of view of the MRI data from the OAI database is well standardized (Fig 2). For an efficient verification of the segmentation results, we first limited ourselves to inspect the aforementioned 11 slices, in which all relevant structures can be seen. Only if there were inconsistencies in the expected segmentation, the whole MRI data set was checked. If any anatomical structure has not been properly segmented or any part of the segmentation looked suspicious in the respective slices, the complete 3D segmentation was again inspected (Fig 2). Out of 48,073 scans 2,927 (6.1%) were selected for additional 3D inspection and 1,077 datasets (2.2%) were excluded from this study due to bad image quality or segmentation failures (S1 Fig). In total 97.8% of all MR images were successfully segmented and included in this study resulting in 46,966 datasets.

Computation of features

Our segmentation yields a disjoint set of labeled voxels for all relevant structures that have been identified in a given MR image. To compute features for KOA assessment of anatomical structures, we use different methods to describe their representation. The amount of voxels, for instance, is proportional to both, the MRI resolution as well as the size of the respective structure. Hence, features related to the volume of an anatomical structure can be directly computed from the segmentations. For more sophisticated features describing the shape of a structure the complexity of the data is reduced by considering only the boundaries of the segmented regions. To achieve this, we represent the respective bounding surfaces as triangulated meshes, which well approximate complexly shaped objects [22]. Based on the surface representations, we compute the surface areas and we statistically analyze the variation in shape for a population of surfaces. In our study, we divide all considered features into two groups: (A) Measurements of volumes, areas and distances which are in the following called "MEAS" features, and (B) Features based on a low-dimensional shape encoding "LDSE" (Fig 3). In detail, the following features are investigated:

MEAS features. *Volume*. Swelling of knee soft tissues can lead to an increase of volume whereas severe degeneration may lead to a complete loss. Hence, six MEAS features are chosen as the volumes V of mM, lM, FC, total TC, medial TC (mTC), and lateral TC (lTC), which are all computed directly from the labeled voxels. Since volumes of anatomical structures of the

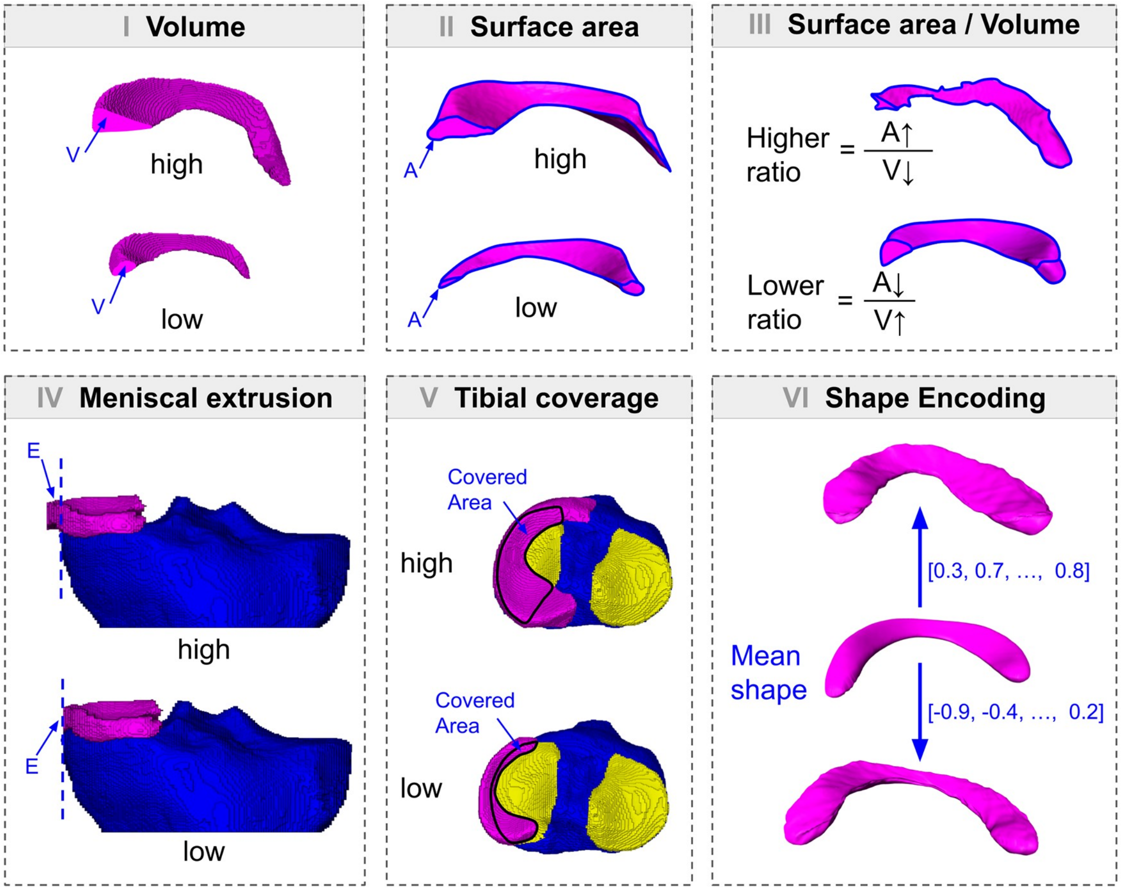

Fig 3. Computation of KOA features. KOA features are exemplarily shown for the medial meniscus (purple), tibial bone (blue) and tibial cartilage (yellow). MEAS features are computed based on segmented voxels and surfaces generated thereof. For MEAS, (I) volumes are computed for femoral cartilage, medial/lateral/total tibial cartilage, and both menisci, (II) surface areas are computed for both menisci, (III) ratio of volume and surface area is computed for both menisci, and (IV) meniscal extrusion is computed for both menisci. (V) Tibial coverage is computed for both tibial plateaus. (VI) LDSE features are computed for femoral bone, tibial bone, and medial/lateral meniscus representing individual shapes relative to the mean shape of the analyzed population.

https://doi.org/10.1371/journal.pone.0258855.g003

knee are known to be correlated with the body size of the subjects [23], all volumes are normalized by the subject's height.

Surface area and ratio of surface area to volume. Especially for the menisci the volume may remain unchanged even if the shape changes, i.e., arthritic menisci may become flat while preserving their volume. To analyze this kind of variation, we selected additional four MEAS features as the surface area A for mM and lM, as well as the ratio of surface area to volume $R_{A,V}$ for mM and lM, respectively.

Meniscal extrusion and tibial coverage. In addition to changes of volume and shape, anatomical structures (such as the menisci) may also show changes of relative positions to other structures within the knee joint. Meniscal extrusion, for example, is highly correlated with tibial coverage, i.e., the more the meniscus is located outside of the joint, the less tibial cartilage is covered [24]. Thus, additional four MEAS features are chosen as the extrusions of mM and lM in medio-lateral direction measured in *mm* as well as coverages of mTC and lTC by the menisci measured as a percentage (see Tack et al. 2018) [21].

LDSE features. Arthritic bones commonly develop osteophytes as well as deformations of the articular surfaces. The shape of the menisci might also change in a more complex manner (e.g. local deformations of the surface), which can not be fully captured by the simple surface area and volume features as contained in MEAS. Therefore, in addition to MEAS, we compute geometrical features assessing the shape of distal femurs, proximal tibias, and the menisci based on a statistical LDSE for all 46,996 datasets, which are divided into seven time points of the OAI database. Using methods from Riemannian shape statistics [25, 26], the mean shape of each anatomy is computed for all subjects segmented in every time point. This mean shape is represented by a common parametrization (i.e., a triangulation).

The variation of all training shapes to the mean is analyzed employing Principal Geodesic Analysis [27]. This analysis yields feature vectors that form a compact encoding for every input shape. In our study the length of the feature vector is proportional to the number of subjects analyzed (see S1 Table) and independent of the geometry's spatial sampling (i.e., the number of sample points of the parametrization).

The feature vectors are ordered by the magnitude of explained variation in shape [28]. We decided to consider only the 300 most significant features per anatomical structure for our analysis since additional features would contribute only with very little variation in shape [29]. The LDSE features in the following investigation are denoted with LDSE-FB for femoral bone, LDSE-TB for tibial bone, LDSE-mM for medial meniscus, and LDSE-lM for lateral meniscus. Additionally, we analyzed combined encodings LDSE-COMB consisting of the first 75 features of each anatomical structure, again resulting in 300 features. An overview of MEAS and LDSE features is shown in S2 Table.

Suitability of the features as potential KOA biomarkers

We investigate the potential MEAS and LDSE features (i) to *classify* the current disease state and (ii) to *predict* a disease state which might develop in the future (Table 2). Annotations from image reading studies of the OAI database are used as the gold standard. The entire procedure from the determination of features, their appropriate selection, up to the assessment of their suitability as KOA biomarkers is depicted in Fig 4.

Evaluation of features for *classification* purposes. We employed features for *classification* of non-arthritic subjects vs. early KOA and for *classification* of non-arthritic vs. subjects with severe KOA. Non-arthritic subjects were defined as KLG ≤ 1, early KOA as KLG 2, and severe KOA as KLG ≥ 3. In addition, we employed features for the *classification* of medial JSN (mJSN) and lateral JSN (lJSN) (Table 2).

Table 2. Analyses performed to investigate the suitability of features to describe different aspects of KOA. Features are employed to classify between KLGs, grades of joint space narrowing (JSN), as well as to predict incident KOA and total knee replacement (TKR).

Classification		Prediction	
KLG	**JSN**	**incident KOA**	**TKR**
0 vs. 1 vs. 2 vs. 3 vs. 4	mJSN: 0 vs. 1 vs. 2 vs. 3	within 1 year	within 1 year
[0;1] vs. 2 vs. [3;4]	lJSN: 0 vs. 1 vs. 2 vs. 3	within 2 years	within 2 years
[0;1] vs. [2;3;4]		within 3 years	within 3 years
0 vs. 2		within 4 years	within 4 years
0 vs. 4		within 5 years	within 5 years

https://doi.org/10.1371/journal.pone.0258855.t002

Evaluation of features for *prediction* purposes. We evaluated if the analyzed features can *predict* which subjects may develop incident KOA (KLG ≥ 2, with joint space narrowing) as well as which subjects may receive a TKR. Both analyses considered a time frame of up to five years (Table 2).

Supervised learning via support vector machines

We trained fast and efficient linear support vector machines (SVMs) [30] to classify as well as to predict the aspects of KOA described in the previous section. Standard SVM parameters were chosen (scikit learn: LinearSVC) as well as a squared L_2 penalty [30]. The SVM was utilized in a Monte-Carlo cross-validation framework: For each investigation, data were randomly split into 90% for training and 10% for testing. The data for training and testing were drawn in a balanced fashion limiting all classes to the one with the least subjects. To obtain results which are representative for the complete cohort, this procedure was repeated 1,000 times. Each respective feature x was min-max normalized over all training datasets: $z = \frac{x - \min(x)}{x_{\max} - x_{\min}}$.

For each classification setting those cases were discarded that have a missing label (e.g. missing KLG information) or any missing feature.

Balanced accuracy BA $= \frac{\text{Sensitivity}+\text{Specificity}}{2}$ was used to evaluate the classification of KLG and JSN grades since it can be computed for both, binary and multi-class classification approaches. Additionally, the agreement of our classifications and predictions to the gold standard annotations was assessed employing weighted kappa [31]. The quality of the employed features was evaluated for prediction of incident KOA as well as for prediction of TKR in terms of sensitivity and specificity using the area under the Receiver Operating Characteristic curve (AUC) [32].

Logistic regression to assess feature importance for prediction of knee arthroplasty

TKR poses an important outcome. For this reason, we perform logistic regression employing MEAS features as independent variables for the prediction of knee arthroplasty. This evaluation is performed for a time frame of one year. The odds ratios (ORs) are provided for each MEAS feature to assess the influence of the respective feature for TKR prediction.

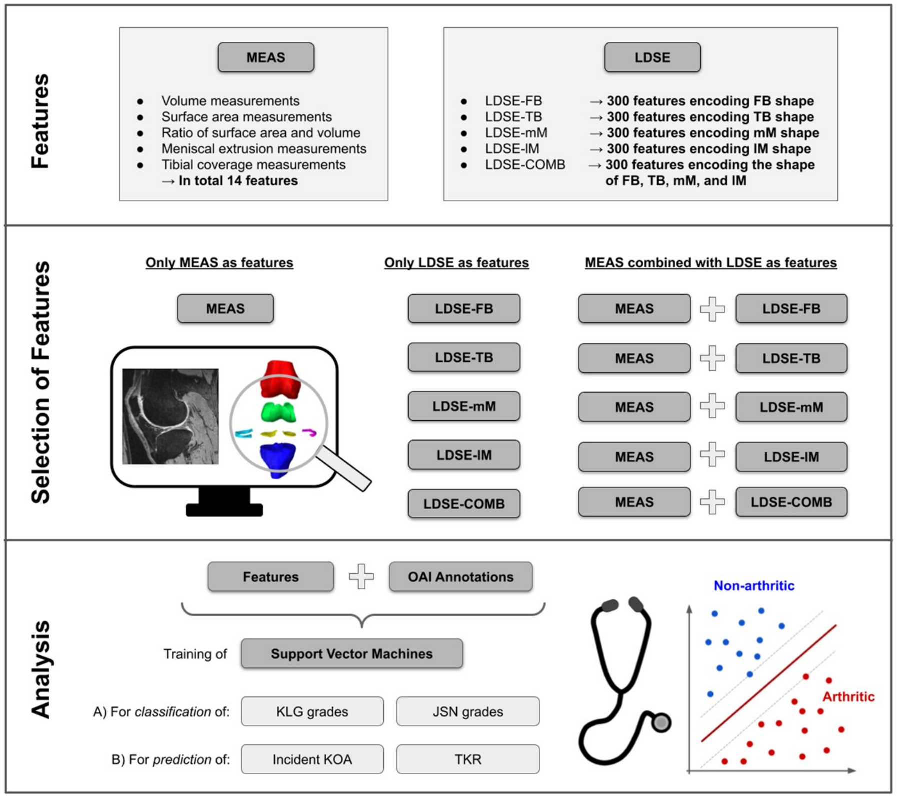
Features
MEAS
Volume measurements
Surface area measurements
Ratio of surface area and volume
Meniscal extrusion measurements
Tibial coverage measurements
→ In total 14 features
LDSE
LDSE-FB → 300 features encoding FB shape
LDSE-TB → 300 features encoding TB shape
LDSE-mM → 300 features encoding mM shape
LDSE-lM → 300 features encoding lM shape
LDSE-COMB → 300 features encoding the shape of FB, TB, mM, and lM
Selection of Features
Only MEAS as features
Only LDSE as features
MEAS combined with LDSE as features
MEAS
LDSE-FB
LDSE-TB
LDSE-mM
LDSE-lM
LDSE-COMB
MEAS + LDSE-FB
MEAS + LDSE-TB
MEAS + LDSE-mM
MEAS + LDSE-lM
MEAS + LDSE-COMB
Analysis
Features + OAI Annotations
Training of Support Vector Machines
A) For classification of: KLG grades JSN grades
B) For prediction of: Incident KOA TKR
Non-arthritic
Arthritic

Fig 4. Selection of KOA features and performed analysis. Top: Computation and determination of suitable features. Middle: Selection of features for further analyses. Chosen are either all features of MEAS, or LDSE-FB/LDSE-TB/LDSE-mM/LDSE-lM features encoding *one* anatomy, LDSE-COMB features encoding *all* anatomies, or MEAS combined with one kind of LDSE features. Bottom: Quantitative analysis of the suitability of selected features as KOA biomarkers by training of support vector machines for classification and prediction purposes.

https://doi.org/10.1371/journal.pone.0258855.g004

Results

Evaluation of features for classification purposes

All results for classification of different grades of KLG (see Table 3) as well as JSN (see Table 4) are averaged over all time points of the OAI study. Detailed information on the results per single time point can be found in the supplementary material.

Classification of Kellgren-Lawrence grades. We conducted 4 classification studies in which different grades of KLG were differentiated.

1. Non-arthritic knees were classified against severely arthritic ones, i.e., KLG 0 vs. KLG 4. The resulting BA ranged from 88% ± 5% using LDSE-lM features up to 99% ± 1% using LDSE-COMB features *plus* MEAS features.

2. Since KLG 2 is regarded as the first grade at which safe indicators are visible to reliably confirm KOA, whereas KLG 0 (no abnormalities) is without any doubt considered as a non-arthritic knee, non-arthritic knees were classified against early KOA, i.e., KLG 0 vs. KLG 2. Here, MEAS yielded the lowest BA (67% ± 3%). LDSE-COMB features *plus* MEAS features performed best, resulting in a BA of 84% ± 2%.

3. The classification of non-arthritic knees (KLG [0;1]) against diseased ones (KLG [2;3;4]) confirmed our previous finding: MEAS features yielded the lowest BA (71% ± 2%) and

Table 3. Classification of KLG (average over all time points weighted by number of cases).

			Balanced Accuracy		
	5-class	**3-class**	**binary classification**		
Features	0 vs 1 vs 2 vs 3 vs 4	[0;1] vs 2 vs [3;4]	[0;1] vs [2;3;4]	0 vs 2	0 vs 4
	N = 41,932	N = 41,932	N = 41,932	N = 26,949	N = 18,252
MEAS	0.44 ± 0.04	0.60 ± 0.02	0.71 ± 0.02	0.67 ± 0.03	0.94 ± 0.03
LDSE-FB	0.41 ± 0.04	0.64 ± 0.02	0.81 ± 0.02	0.79 ± 0.02	0.95 ± 0.03
LDSE-FB + MEAS	0.45 ± 0.04	0.68 ± 0.02	0.82 ± 0.02	0.80 ± 0.02	0.97 ± 0.02
LDSE-TB	0.40 ± 0.04	0.62 ± 0.03	0.79 ± 0.02	0.77 ± 0.02	0.92 ± 0.04
LDSE-TB + MEAS	0.43 ± 0.04	0.67 ± 0.02	0.80 ± 0.02	0.78 ± 0.02	0.96 ± 0.03
LDSE-mM	0.42 ± 0.04	0.63 ± 0.02	0.77 ± 0.02	0.75 ± 0.02	0.91 ± 0.04
LDSE-mM + MEAS	0.47 ± 0.04	0.67 ± 0.02	0.78 ± 0.02	0.76 ± 0.02	0.97 ± 0.02
LDSE-lM	0.37 ± 0.04	0.56 ± 0.02	0.72 ± 0.02	0.69 ± 0.02	0.88 ± 0.05
LDSE-lM + MEAS	0.44 ± 0.04	0.64 ± 0.02	0.76 ± 0.02	0.73 ± 0.03	0.96 ± 0.03
LDSE-COMB	0.50 ± 0.04	0.72 ± 0.02	0.84 ± 0.01	0.84 ± 0.02	0.99 ± 0.02
LDSE-COMB + MEAS	0.52 ± 0.04	0.73 ± 0.02	0.84 ± 0.01	0.84 ± 0.02	0.99 ± 0.01
			Weighted Kappa		
	5-class	**3-class**	**binary classification**		
Features	0 vs 1 vs 2 vs 3 vs 4	[0;1] vs 2 vs [3;4]	[0;1] vs [2;3;4]	0 vs 2	0 vs 4
	N = 41,932	N = 41,932	N = 41,932	N = 26,949	N = 18,252
MEAS	0.68 ± 0.05	0.58 ± 0.04	0.41 ± 0.04	0.34 ± 0.05	0.88 ± 0.07
LDSE-FB	0.68 ± 0.05	0.65 ± 0.03	0.62 ± 0.03	0.58 ± 0.04	0.90 ± 0.06
LDSE-FB + MEAS	0.71 ± 0.04	0.70 ± 0.03	0.65 ± 0.03	0.60 ± 0.04	0.94 ± 0.05
LDSE-TB	0.63 ± 0.06	0.61 ± 0.04	0.58 ± 0.03	0.53 ± 0.05	0.85 ± 0.07
LDSE-TB + MEAS	0.68 ± 0.05	0.68 ± 0.04	0.61 ± 0.03	0.56 ± 0.05	0.92 ± 0.05
LDSE-mM	0.64 ± 0.05	0.63 ± 0.04	0.54 ± 0.03	0.50 ± 0.05	0.83 ± 0.08
LDSE-mM + MEAS	0.73 ± 0.05	0.68 ± 0.03	0.56 ± 0.04	0.51 ± 0.04	0.94 ± 0.05
LDSE-lM	0.56 ± 0.06	0.50 ± 0.04	0.44 ± 0.04	0.39 ± 0.05	0.76 ± 0.09
LDSE-lM + MEAS	0.69 ± 0.05	0.64 ± 0.03	0.52 ± 0.03	0.45 ± 0.05	0.91 ± 0.06
LDSE-COMB	0.78 ± 0.04	0.75 ± 0.03	0.68 ± 0.03	0.68 ± 0.04	0.97 ± 0.03
LDSE-COMB + MEAS	0.78 ± 0.04	0.75 ± 0.03	0.69 ± 0.03	0.68 ± 0.04	0.98 ± 0.03

https://doi.org/10.1371/journal.pone.0258855.t003

Table 4. Balanced accuracy and weighted kappa for classification of joint space narrowing (JSN) for medial and lateral femorotibial compartment averaged over all time points weighted by number of cases per time point.

Balanced Accuracy			**Weighted Kappa**		
	Medial	**Lateral**		**Medial**	**Lateral**
Features	**N = 41,932**	**N = 41,932**	**Features**	**N = 41,932**	**N = 41,932**
MEAS	0.54 ± 0.07	0.57 ± 0.11	MEAS	0.72 ± 0.08	0.72 ± 0.14
LDSE-FB	0.43 ± 0.08	0.40 ± 0.11	LDSE-FB	0.50 ± 0.13	0.42 ± 0.20
LDSE-FB + MEAS	0.48 ± 0.08	0.46 ± 0.12	LDSE-FB + MEAS	0.60 ± 0.12	0.57 ± 0.18
LDSE-TB	0.44 ± 0.08	0.40 ± 0.12	LDSE-TB	0.50 ± 0.13	0.42 ± 0.22
LDSE-TB + MEAS	0.49 ± 0.08	0.46 ± 0.12	LDSE-TB + MEAS	0.61 ± 0.12	0.58 ± 0.17
LDSE-mM	0.52 ± 0.08	0.33 ± 0.11	LDSE-mM	0.67 ± 0.11	0.24 ± 0.24
LDSE-mM + MEAS	0.56 ± 0.08	0.46 ± 0.12	LDSE-mM + MEAS	0.72 ± 0.08	0.53 ± 0.20
LDSE-lM	0.40 ± 0.08	0.49 ± 0.10	LDSE-lM	0.42 ± 0.14	0.57 ± 0.14
LDSE-lM + MEAS	0.48 ± 0.08	0.51 ± 0.11	LDSE-lM + MEAS	0.62 ± 0.12	0.59 ± 0.15
LDSE-COMB	0.56 ± 0.08	0.50 ± 0.12	LDSE-COMB	0.71 ± 0.10	0.64 ± 0.16
LDSE-COMB + MEAS	0.57 ± 0.09	0.52 ± 0.12	LDSE-COMB + MEAS	0.73 ± 0.09	0.67 ± 0.15

https://doi.org/10.1371/journal.pone.0258855.t004

LDSE-COMB features *plus* MEAS features performed best resulting in a BA of 84% ± 1%. For classification of KLG [0;1] vs. KLG 2 vs. KLG [3;4] in a 3-class setting the BA slightly decreased to 60% ± 2% for MEAS features and 73% ± 2% for LDSE-COMB features *plus* MEAS features.

4. Differentiation between all 5 grades of KOA is the most challenging task. The weakest differentiation between these five grades was possible using LDSE-lM features (37% ± 4%). Best results were achieved using LDSE-COMB (50% ± 4%) which was further improved using LDSE-COMB features *plus* MEAS features (52% ± 4%).

In all classification studies, LDSE-FB features performed better than LDSE-TB features in terms of BA with a margin of at least 1 and up to 3 percent. Also, the LDSE-mM features always performed better than the LDSE-lM features with a margin of at least 3 and up to 7 percent. LDSE-FB features usually scored better or at least similar good as LDSE-mM features. MEAS features never scored better than the LDSE-COMB features, however, a combination of both features mostly improved the classification results in the 3-class and the 5-class setting. Employing LDSE-COMB features *plus* MEAS features, a moderate agreement was found as measured with the weighted kappa coefficient of 0.75 ± 0.03 and 0.78 ± 0.04 for the 3-class and the 5-class analysis, respectively.

Classification of joint space narrowing. We investigated the potential of the features for differentiation of different grades of mJSN and lJSN, respectively.

1. For a classification of mJSN the LDSE-lM features yielded the lowest accuracy (BA of 40% ± 8%). Utilizing LDSE-FB resulted in similar results as LDSE-TB (BA of 43% ± 8% and 44% ± 8%). Among the LDSE features, LDSE-mM yielded best results (BA of 52% ± 8%), however, not as good as LDSE-COMB (BA of 56% ± 8%). MEAS features scored slightly lower than LDSE-COMB (BA of 54% ± 7%). LDSE-COMB *plus* MEAS led to the best results, i.e., a BA of 57% ± 9%.

2. For a classification of lJSN the LDSE-mM features yielded the lowest accuracy (BA of 33% ± 11%). LDSE-FB performed comparable to LDSE-TB (BA of 40% ± 11% and 40% ± 12%, respectively). Among the individual LDSE features, LDSE-lM yielded best results (BA of 49% ± 10%), which was close to LDSE-COMB (BA of 50% ± 12%). The combination with MEAS features improved the results for all individual LDSE features as well as for LDSE--COMB. However, MEAS features alone yielded the best results for classification of lJSN (57% ± 11%).

In terms of kappa statistics the best results were achieved using LDSE-COMB *plus* MEAS for mJSN and MEAS for lJSN yielding moderate agreement in both cases (average weighted kappa of 0.73 ± 0.09 and 0.72 ± 0.14, respectively).

Evaluation of features for prediction purposes

Results for a prediction of incident KOA and TKR within different periods of time (see Table 5) are summarized in this section.

Prediction of incident knee osteoarthritis. Utilization of individual LDSE features for prediction of incident KOA yielded AUCs in a range from 0.52 to 0.56. Adding MEAS features to the individual LDSE features either led to an increase of the AUC or at least to a similar value, ranging between 0.53 to 0.57.

Both, MEAS features and LDSE-COMB features performed better than the individual features alone. Over all investigated periods of time the prediction of KOA within one year was best. For prediction within one year, the AUC was 0.60 ± 0.11 using LDSE-COMB features *or*

Table 5. Prediction of incident KOA (KLG >= 2 with JSN) and TKR surgery within *a* years. In both investigations the AUC is evaluated utilizing the respective true positive rate as well as the false positive rate.

	Incident KOA—AUC				
	a = 1 year	*a* = 2 years	*a* = 3 years	*a* = 4 years	*a* = 5 years
	#No_INC = 26,302	#No_INC = 25,890	#No_INC = 25,647	#No_INC = 25,395	#No_INC = 25,225
Features	#INC = 372	#INC = 784	#INC = 1,027	#INC = 1,279	#INC = 1,449
MEAS	0.60 ± 0.11	0.60 ± 0.09	0.59 ± 0.08	0.59 ± 0.07	0.59 ± 0.07
LDSE-FB	0.55 ± 0.12	0.56 ± 0.09	0.56 ± 0.08	0.56 ± 0.07	0.56 ± 0.07
LDSE-FB + MEAS	0.55 ± 0.12	0.57 ± 0.09	0.57 ± 0.08	0.57 ± 0.07	0.57 ± 0.07
LDSE-TB	0.54 ± 0.11	0.52 ± 0.10	0.53 ± 0.08	0.54 ± 0.07	0.53 ± 0.07
LDSE-TB + MEAS	0.56 ± 0.12	0.55 ± 0.10	0.54 ± 0.08	0.55 ± 0.07	0.55 ± 0.07
LDSE-mM	0.55 ± 0.11	0.56 ± 0.09	0.56 ± 0.08	0.56 ± 0.07	0.56 ± 0.07
LDSE-mM + MEAS	0.57 ± 0.11	0.56 ± 0.09	0.56 ± 0.09	0.56 ± 0.07	0.57 ± 0.07
LDSE-lM	0.54 ± 0.10	0.55 ± 0.09	0.55 ± 0.08	0.55 ± 0.07	0.55 ± 0.07
LDSE-lM + MEAS	0.55 ± 0.11	0.56 ± 0.09	0.56 ± 0.09	0.57 ± 0.08	0.57 ± 0.07
LDSE-COMB	0.59 ± 0.11	0.59 ± 0.09	0.59 ± 0.08	0.59 ± 0.07	0.59 ± 0.07
LDSE-COMB + MEAS	0.61 ± 0.12	0.59 ± 0.09	0.60 ± 0.08	0.59 ± 0.07	0.59 ± 0.07
	TKR—AUC				
	a = 1 year	*a* = 2 years	*a* = 3 years	*a* = 4 years	*a* = 5 years
	#No_TKR = 46,575	#No_TKR = 46,281	#No_TKR = 45,992	#No_TKR = 45,726	#No_TKR = 45,481
Features	#TKR = 421	#TKR = 715	#TKR = 1,004	#TKR = 1,270	#TKR = 1,515
MEAS	0.74 ± 0.12	0.73 ± 0.10	0.73 ± 0.08	0.72 ± 0.07	0.72 ± 0.06
LDSE-FB	0.69 ± 0.13	0.71 ± 0.10	0.71 ± 0.08	0.70 ± 0.07	0.70 ± 0.07
LDSE-FB + MEAS	0.72 ± 0.13	0.73 ± 0.09	0.72 ± 0.08	0.72 ± 0.07	0.72 ± 0.06
LDSE-TB	0.65 ± 0.14	0.67 ± 0.10	0.68 ± 0.08	0.67 ± 0.08	0.67 ± 0.07
LDSE-TB + MEAS	0.68 ± 0.13	0.69 ± 0.10	0.70 ± 0.08	0.70 ± 0.07	0.69 ± 0.06
LDSE-mM	0.65 ± 0.14	0.69 ± 0.10	0.69 ± 0.08	0.70 ± 0.08	0.70 ± 0.06
LDSE-mM + MEAS	0.70 ± 0.13	0.72 ± 0.10	0.72 ± 0.08	0.72 ± 0.07	0.73 ± 0.06
LDSE-lM	0.67 ± 0.13	0.68 ± 0.11	0.69 ± 0.09	0.69 ± 0.08	0.69 ± 0.07
LDSE-lM + MEAS	0.70 ± 0.13	0.71 ± 0.10	0.73 ± 0.08	0.73 ± 0.07	0.72 ± 0.06
LDSE-COMB	0.74 ± 0.12	0.74 ± 0.09	0.74 ± 0.08	0.74 ± 0.07	0.73 ± 0.07
LDSE-COMB + MEAS	0.76 ± 0.12	0.75 ± 0.10	0.75 ± 0.08	0.75 ± 0.07	0.74 ± 0.06

https://doi.org/10.1371/journal.pone.0258855.t005

MEAS features, and 0.61 ± 0.12 using LDSE-COMB features *plus* MEAS features. However, a prediction of incident KOA within 5 years also led to reasonable results, i.e., an AUC of 0.59 ± 0.07 when utilizing LDSE-COMB features *plus* MEAS features.

Across all considered time periods, all individual LDSE features consistently yielded a lower AUC than the MEAS features and the LDSE-COMB features, respectively. A combination of LDSE-COMB and MEAS features, however, either improved the results or at least yielded to equal results for every period of time that has been analysed.

Prediction of total knee replacement. LDSE features led to a prediction of TKR with AUCs ranging from 0.65 to 0.71. MEAS features as well as LDSE-COMB features yielded slightly higher AUC values ranging from 0.72 to 0.74 and from 0.73 to 0.74, respectively. In tendency, the best results were achieved for prediction of TKR within one year. The AUC was 0.74 ± 0.12 using LDSE-COMB features, 0.74 ± 0.12 using MEAS features, and 0.76 ± 0.12 using LDSE-COMB features *plus* MEAS features. However, the prediction of receiving TKR within the next 5 years also led to comparable results, i.e., an AUC of 0.74 ± 0.06 when utilizing LDSE-COMB features *plus* MEAS features.

Table 6. TKR prediction via logistic regression using MEAS features. The odds ratios are provided for each feature.

MEAS feature	Odds ratio	95% confidence interval
TC_{mTC}	7.04	6.70 to 7.42
E_{mM}	2.97	2.79 to 3.15
TC_{lTC}	2.64	2.52 to 2.77
A_{mM}	2.16	2.05 to 2.26
V_{FC}	2.09	2.02 to 2.18
V_{lTC}	2.08	1.98 to 2.18
V_{TC}	1.74	1.68 to 1.79
$R^{mM}_{A,V}$	1.64	1.57 to 1.71
E_{lM}	1.60	1.54 to 1.66
A_{lM}	1.46	1.39 to 1.50
$R^{lM}_{A,V}$	1.45	1.39 to 1.50
V_{mTC}	1.23	1.20 to 1.26
V_{lM}	1.19	1.18 to 1.19
V_{mM}	1.13	1.11 to 1.15

https://doi.org/10.1371/journal.pone.0258855.t006

Logistic regression for prediction of total knee replacement. The highest odds ratios were achieved by features related to the MM, such as medial tibial coverage (OR: 7.04, 95% confidence interval (CI): 6.70–7.42) and medial meniscal extrusion (OR: 2.97, 95% CI: 2.79–3.15) as summarized in Table 6. Additionally, also lateral tibial coverage yielded one of the highest ORs (OR: 2.64, 95% CI: 2.52–2.77). The lowest ORs were achieved for MEAS features related to medial tibial cartilage volume (OR: 1.23, 95% CI: 1.20–1.26), LM volume (OR: 1.19, 95% CI: 1.18–1.19), and MM volume (OR: 1.13, 95% CI: 1.11–1.15).

Discussion

Based on our segmentations we analyzed the potential of a set of features for FB, TB, mM, and lM as biomarkers for classification of KLG and JSN and for prediction incident KOA and TKR. Compared to the related work, our classification of KOA aspects yielded similar BA as the method of Nasser et al. [33] for a determination of early KOA (classification of KLG 0 vs. KLG 2—theirs: 82.52%, ours: 84%). However, our kappa was lower than the ones of Tiulpin et al. [34] (5-class classification KLG—theirs: 0.83, ours: 0.78) and Ngyuen et al. [35] (5-class classification KLG—theirs: 0.88, ours: 0.78). For prediction of TKR, a slightly higher AUC was achieved than Eckstein et al. [36] (prediction of TKR—theirs: 0.66, ours: 0.74), but a lower AUC than Tolpadi et al. [37] (prediction of TKR—theirs: 0.83, ours: 0.74). Judging the results achieved by the features investigated in this study, one should consider that (i) the features were computed solely based on our segmentations, i.e., without taking any gray-value intensity information into account, and (ii) the features were computed from MRI data which are (in contrast to X-Ray) not acquired with the knee being in a load-bearing condition. The evaluation of KLG and JSN from X-Rays is highly depending on the rater [38] which can result in kappa statistics between 0.56 and 0.67 [39, 40]. MRI offers additional potential to investigate 3-D features, e.g. of the anatomies' shapes, and should thus be preferred for investigation of novel biomarkers.

Considering the potential of shape encodings of the individual anatomies as biomarkers, the tibial bone showed in tendency the lowest descriptiveness for all analyzed aspects of KOA. This confirms the findings of Neogi et al. [41] who also found that shape variations of the tibial bone are either hard to capture with a low-dimensional encoding or are not as prominent and clear as features of the femoral bone. The representation of the femoral bone showed best results among all individual anatomies for differentiation of KLG. LDSE-mM performed best among the individual structures' representations for a classification of medial JSN and LDSE-lM for a classification of lateral JSN confirming the importance of the menisci in the context of JSN [42]. In summary, the combined encoding LDSE-COMB consisting of features of femoral bone, tibial bone and meniscus yielded better results than any individual encoding of the structures of interest. This highlights the need for a holistic assessment of the entire knee. In future, further analysis of the shape of bones and menisci should be performed, i.e., analyzing the shape of sub-groups of the population in comparison to the mean shape [10]. Such an analysis may form the basis to detect clusters of similar shapes and to identify patterns within the progression of the disease [43].

In all cases (except the classification of lateral JSN) we observed that the results of LDSE--COMB can be further improved by adding MEAS features. Especially the position of the menisci can not be directly explained by a shape encoding and thus the importance of this feature is confirmed. For prediction of TKR, the potential of the MEAS features was further shown. The position of the menisci reflected in the tibial coverage and meniscal extrusion yielded the highest ORs. Volume measurements were less important with respect to TKR prediction.

As far as it concerns the detection of patterns within medical image data, we have great conviction that deep learning is the most promising approach. However, CNNs are relatively agnostic to shape information and focus mainly on the texture of images [44]. Moreover, the decision process of CNNs can be influenced by subtle changes in image information which can be hardly noticed by humans [45]. Thus, the application of deep learning for diagnostics in clinical practice remains a challenge, since results need to be explainable [46]. Furthermore, to demonstrate generalizability, deep-learning based algorithms have to be evaluated in prospective studies using large datasets acquired at different institutions with different imaging parameters and different imaging hardware [11]. Deep learning frameworks often already fuse image data with meta information using concepts e.g. of input level information fusion to add additional knowledge and improve the classification accuracy [37, 47, 48]. The features we investigated could be merged into deep learning frameworks similar as in [37, 47, 48] to provide additional insights wrt. the shape of an anatomical structure. Moreover, in contrast to pure machine learning-based approaches, the features we evaluated in this study are explainable: The semantic segmentations which are the basis of our features can be inspected in 3-D. Additionally, the measurements MEAS can be visualized in the image data and the mean shapes required for LDSE features as well as the modes of variation can be visually explored [28].

With this publication we will make our segmentations publicly available (https://pubdata.zib.de). By using our results, developments in image-based KOA biomarkers for different structures of the knee can be fostered. Our segmentation masks have been created in a fully automated manner. Quality assurance has been employed in order to omit suspicious segmentations. However, there might be still some potential for improvement e.g. in regions of high morphological variability or inhomogeneous image appearance which could be corrected manually or by novel, automated methods. Other researchers are invited to use the provided segmentations as well as to report errors or correct them and transfer the corrected data back in order to update the collection of segmentations. We envisage a platform that enables other researchers to utilize all data contained in the OAI database and to make it easily accessible for future research.

Supporting information

S1 Fig. Flow chart of quality assurance and of the data selection process for data inclusion.
(PDF)

S1 Table. Number of shapes which are utilized for the computation of LDSE features. Our method for automated segmentation yielded triangulated meshes for 46,996 MRI datasets contained in the OAI database. Each time point of the OAI database is analyzed independently to avoid inter-subject correlations between the shapes.
(PDF)

S2 Table. Summary of MEAS and LDSE features. MEAS contains computations of the volume (V_i), surface area (A_i), the ratio of volume and surface area ($R^i_{A,V}$), meniscal extrusion E_i, and tibial coverage TC_i for the respective anatomy i. LDSE contains features encoding the geometry of the distal femoral bone (LDSE-FB), proximal tibial bone (LDSE-TB), medial meniscus (LDSE-mM), lateral meniscus (LDSE-lM), or a combined representation of all 4 anatomies (LDSE-COMB).
(PDF)

S3 Table. Classification of KLG: v00-v24.
(PDF)

S4 Table. Classification of KLG: v36-v72.
(PDF)

S5 Table. Classification of KLG: v96.
(PDF)

S6 Table. Classification of JSN: v00-v36.
(PDF)

S7 Table. Classification of JSN: v48-v96.
(PDF)

Acknowledgments

We gratefully thank the Max Delbrück Center for Molecular Medicine, especially Alf Wachsmann and Martin Siegert, for providing us access to the NVIDIDA DGX-1 system during this project.

The Osteoarthritis Initiative is a public-private partnership comprised of 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 National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

Author Contributions

Conceptualization: Alexander Tack, Felix Ambellan, Stefan Zachow.

Data curation: Alexander Tack, Felix Ambellan.

Formal analysis: Alexander Tack, Felix Ambellan.

Funding acquisition: Stefan Zachow.

Investigation: Alexander Tack, Stefan Zachow.

Methodology: Alexander Tack, Felix Ambellan, Stefan Zachow.

Project administration: Stefan Zachow.

Resources: Stefan Zachow.

Software: Alexander Tack, Stefan Zachow.

Supervision: Stefan Zachow.

Validation: Alexander Tack.

Visualization: Alexander Tack, Stefan Zachow.

Writing – original draft: Alexander Tack, Felix Ambellan, Stefan Zachow.

References

1. Kellgren J, Lawrence J. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957; 16(4):494. https://doi.org/10.1136/ard.16.4.494 PMID: 13498604
2. Segal NA, Frick E, Duryea J, Nevitt MC, Niu J, Torner JC, et al. Comparison of tibiofemoral joint space width measurements from standing CT and fixed flexion radiography. J Orthop Res. 2017; 35(7):1388–1395. https://doi.org/10.1002/jor.23387 PMID: 27504863
3. Reyneke CJF, Lüthi M, Burdin V, Douglas TS, Vetter T, Mutsvangwa TE. Review of 2-d/3-d reconstruction using statistical shape and intensity models and x-ray image synthesis: Toward a unified framework. IEEE Rev Biomed Eng. 2018; 12:269–286. https://doi.org/10.1109/RBME.2018.2876450 PMID: 30334808
4. Guermazi A, Roemer FW, Haugen IK, Crema MD, Hayashi D. MRI-based semiquantitative scoring of joint pathology in osteoarthritis. Nat Rev Rheumatol. 2013; 9(4):236. https://doi.org/10.1038/nrrheum.2012.223 PMID: 23321609
5. Roemer FW, Hunter DJ, Crema MD, Kwoh CK, Ochoa-Albiztegui E, Guermazi A. An illustrative overview of semi-quantitative MRI scoring of knee osteoarthritis: lessons learned from longitudinal observational studies. Osteoarthr Cartil. 2016; 24(2):274–289. https://doi.org/10.1016/j.joca.2015.08.011 PMID: 26318656
6. Eckstein F, Maschek S, Wirth W, Hudelmaier M, Hitzl W, Wyman B, et al. One year change of knee cartilage morphology in the first release of participants from the Osteoarthritis Initiative progression subcohort: association with sex, body mass index, symptoms and radiographic osteoarthritis status. Ann Rheum Dis. 2009; 68(5):674–679. https://doi.org/10.1136/ard.2008.089904 PMID: 18519425
7. Eckstein F, Charles HC, Buck RJ, Kraus VB, Remmers AE, Hudelmaier M, et al. Accuracy and precision of quantitative assessment of cartilage morphology by magnetic resonance imaging at 3.0 T. Arthritis Rheum. 2005; 52(10):3132–3136. https://doi.org/10.1002/art.21348 PMID: 16200592
8. Eckstein F, Ateshian G, Burgkart R, Burstein D, Cicuttini F, Dardzinski B, et al. Proposal for a nomenclature for magnetic resonance imaging based measures of articular cartilage in osteoarthritis. Osteoarthr Cartil. 2006; 14(10):974–983. https://doi.org/10.1016/j.joca.2006.03.005 PMID: 16730462
9. Sharma L, Eckstein F, Song J, Guermazi A, Prasad P, Kapoor D, et al. Relationship of meniscal damage, meniscal extrusion, malalignment, and joint laxity to subsequent cartilage loss in osteoarthritic knees. Arthritis Rheum. 2008; 58(6):1716–1726. https://doi.org/10.1002/art.23462 PMID: 18512777
10. Bredbenner TL, Eliason TD, Potter RS, Mason RL, Havill LM, Nicolella DP. Statistical shape modeling describes variation in tibia and femur surface geometry between Control and Incidence groups from the osteoarthritis initiative database. J Biomech. 2010; 43(9):1780–1786. https://doi.org/10.1016/j.jbiomech.2010.02.015 PMID: 20227696
11. Kijowski R, Liu F, Caliva F, Pedoia V. Deep learning for lesion detection, progression, and prediction of musculoskeletal disease. J Magn Reson Imaging. 2020; 52:1607–1619. https://doi.org/10.1002/jmri.27001 PMID: 31763739
12. Kokkotis C, Moustakidis S, Papageorgiou E, Giakas G, Tsaopoulos D. Machine Learning in Knee Osteoarthritis: A Review. Osteoarthr Cartil Open. 2020; 2(3):1–13. https://doi.org/10.1016/j.ocarto.2020.100069

13. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med. 2018; 15 (11):1–19. https://doi.org/10.1371/journal.pmed.1002699 PMID: 30481176

14. Liu F, Zhou Z, Samsonov A, Blankenbaker D, Larison W, Kanarek A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. Radiology. 2018; 289(1):160–169. https://doi.org/10.1148/radiol.2018172986 PMID: 30063195

15. Fritz B, Marbach G, Civardi F, Fucentese SF, Pfirrmann CW. Deep convolutional neural network-based detection of meniscus tears: comparison with radiologists and surgery as standard of reference. Skelet Radiol. 2020; 49:1207–1217. https://doi.org/10.1007/s00256-020-03458-0

16. Norman B, Pedoia V, Majumdar S. Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. Radiology. 2018; 288(1):177–185. https://doi.org/10.1148/radiol.2018172322 PMID: 29584598

17. Pedoia V, Norman B, Mehany SN, Bucknor MD, Link TM, Majumdar S. 3D convolutional neural networks for detection and severity staging of meniscus and PFJ cartilage morphological degenerative changes in osteoarthritis and anterior cruciate ligament subjects. J Magn Reson Imaging. 2019; 49 (2):400–410. https://doi.org/10.1002/jmri.26246 PMID: 30306701

18. Neyshabur B, Bhojanapalli S, McAllester D, Srebro N. Exploring generalization in deep learning. In: Adv Neural Inf Process Syst; 2017. p. 5947–5956.

19. Peterfy C, Gold G, Eckstein F, Cicuttini F, Dardzinski B, Stevens R. MRI protocols for whole-organ assessment of the knee in osteoarthritis. Osteoarthr Cartil. 2006; 14:95–111. https://doi.org/10.1016/j.joca.2006.02.029 PMID: 16750915

20. Ambellan F, Tack A, Ehlke M, Zachow S. Automated Segmentation of Knee Bone and Cartilage combining Statistical Shape Knowledge and Convolutional Neural Networks: Data from the Osteoarthritis Initiative. Med Image Anal. 2019; 52(2):109–118. https://doi.org/10.1016/j.media.2018.11.009 PMID: 30529224

21. Tack A, Mukhopadhyay A, Zachow S. Knee Menisci Segmentation using Convolutional Neural Networks: Data from the Osteoarthritis Initiative. Osteoarthr Cartil. 2018; 26(5):680–688. https://doi.org/10.1016/j.joca.2018.02.907 PMID: 29526784

22. Zachow S, Zilske M, Hege HC. 3D reconstruction of individual anatomy from medical image data: Segmentation and geometry processing. Takustr. 7, 14195 Berlin: ZIB; 2007. 07–41.

23. Ding C, Cicuttini F, Scott F, Glisson M, Jones G. Sex differences in knee cartilage volume in adults: role of body and bone size, age and physical activity. Rheumatology. 2003; 42(11):1317–1323. https://doi.org/10.1093/rheumatology/keg374 PMID: 12810930

24. Berthiaume MJ, Raynauld JP, Martel-Pelletier J, Labonté F, Beaudoin G, Bloch DA, et al. Meniscal tear and extrusion are strongly associated with progression of symptomatic knee osteoarthritis as assessed by quantitative magnetic resonance imaging. Ann Rheum Dis. 2005; 64(4):556–563. https://doi.org/10.1136/ard.2004.023796 PMID: 15374855

25. Ambellan F, Zachow S, von Tycowicz C. A surface-theoretic approach for statistical shape modeling. In: Med Image Comput Comput Assist Interv. Springer; 2019. p. 21–29.

26. Ambellan F, Zachow S, von Tycowicz C. Rigid motion invariant statistical shape modeling based on discrete fundamental forms. Med Image Anal. 2021; 73:102178 https://doi.org/10.1016/j.media.2021.102178 PMID: 34343840

27. Fletcher PT, Lu C, Pizer SM, Joshi S. Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans Med Imaging. 2004; 23(8):995–1005. https://doi.org/10.1109/TMI.2004.831793 PMID: 15338733

28. Ambellan F, Lamecker H, von Tycowicz C, Zachow S. Statistical shape models: understanding and mastering variation in anatomy. In: Biomedical Visualisation. Springer; 2019. p. 67–84.

29. von Tycowicz C, Ambellan F, Mukhopadhyay A, Zachow S. An efficient Riemannian statistical shape model using differential coordinates. Med Image Anal. 2018; 43:1–9. https://doi.org/10.1016/j.media.2017.09.004 PMID: 28961450

30. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. J Mach Learn Res. 2008; 9(Aug):1871–1874.

31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. https://doi.org/10.2307/2529310 PMID: 843571

32. Akobeng AK. Understanding diagnostic tests 3: receiver operating characteristic curves. Acta Paediatr. 2007; 96(5):644–647. https://doi.org/10.1111/j.1651-2227.2006.00178.x PMID: 17376185

33. Nasser Y, Jennane R, Chetouani A, Lespessailles E, El Hassouni M. Discriminative Regularized Auto-Encoder for Early Detection of Knee OsteoArthritis: Data from the Osteoarthritis Initiative. IEEE Trans Med Imaging. 2020; 39(9):2976–2984. https://doi.org/10.1109/TMI.2020.2985861 PMID: 32286962

34. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. Sci Rep. 2018; 8(1):1727. https://doi.org/10.1038/s41598-018-20132-7 PMID: 29379060

35. Nguyen HH, Saarakkala S, Blaschko MB, Tiulpin A. Semixup: In-and Out-of-Manifold Regularization for Deep Semi-Supervised Knee Osteoarthritis Severity Grading From Plain Radiographs. IEEE Trans Med Imaging. 2020; 39(12):4346–4356. https://doi.org/10.1109/TMI.2020.3017007 PMID: 32804644

36. Eckstein F, Kwoh CK, Boudreau RM, Wang Z, Hannon MJ, Cotofana S, et al. Quantitative MRI measures of cartilage predict knee replacement: a case–control study from the Osteoarthritis Initiative. Ann Rheum Dis. 2013; 72(5):707–714.

37. Tolpadi AA, Lee JJ, Pedoia V, Majumdar S. Deep learning predicts total knee replacement from magnetic resonance images. Sci Rep. 2020; 10(1):1–12. https://doi.org/10.1038/s41598-020-63395-9 PMID: 32286452

38. Sun Y, Günther K, Brenner H. Reliability of radiographic grading of osteoarthritis of the hip and knee. Scand J Rheumatol. 1997; 26(3):155–165. https://doi.org/10.3109/03009749709065675 PMID: 9225869

39. Gossec L, Jordan J, Mazzuca S, Lam MA, Suarez-Almazor M, Renner J, et al. Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the OARSI–OMERACT task force. Osteoarthr Cartil. 2008; 16(7):742–748.

40. Culvenor AG, Engen CN, Øiestad BE, Engebretsen L, Risberg MA. Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Knee Surgery, Sports Traumatology, Arthroscopy. 2015; 23(12):3532–3539. https://doi.org/10.1007/s00167-014-3205-0 PMID: 25079135

41. Neogi T, Bowes MA, Niu J, De Souza KM, Vincent GR, Goggins J, 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. 2013; 65(8):2048–2058. https://doi.org/10.1002/art.37987 PMID: 23650083

42. Gale D, Chaisson C, Totterman S, Schwartz R, Gale M, Felson D. Meniscal subluxation: association with osteoarthritis and joint space narrowing. Osteoarthr Cartil. 1999; 7(6):526–532. https://doi.org/10.1053/joca.1999.0256 PMID: 10558850

43. Bowes MA, Kacena K, Alabas OA, Brett AD, Dube B, Bodick N, et al. Machine-learning, MRI bone shape and important clinical outcomes in osteoarthritis: data from the Osteoarthritis Initiative. Ann Rheum Dis. 2020; 80(4):502–508. https://doi.org/10.1136/annrheumdis-2020-217160 PMID: 33188042

44. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231. 2018.

45. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc IEEE Symp Secur Priv. IEEE; 2017. p. 39–57.

46. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinformatics. 2018; 19(6):1236–1246. https://doi.org/10.1093/bib/bbx044 PMID: 28481991

47. Tiulpin A, Klein S, Bierma-Zeinstra SM, Thevenot J, Rahtu E, van Meurs J, et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci Rep. 2019; 9(1):1–11. https://doi.org/10.1038/s41598-019-56527-3 PMID: 31882803

48. Nguyen HH, Saarakkala S, Blaschko MB, Tiulpin A. DeepProg: A Transformer-based Framework for Predicting Disease Prognosis. arXiv preprint arXiv:2104.03642. 2021.