Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

ALEXANDER TACK, ANIRBAN MUKHOPADHYAY,
STEFAN ZACHOW

Knee Menisci Segmentation using Convolutional Neural Networks: Data from the Osteoarthritis Initiative[1]

[1]Preprint submitted to Osteoarthritis and Cartilage

ZIB Report 18-15 (March 2018)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30-84185-0
Telefax: +49 30-84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Knee Menisci Segmentation using Convolutional Neural Networks: Data from the Osteoarthritis Initiative

Alexander Tack[a,*], Anirban Mukhopadhyay[b], Stefan Zachow[a]

[a] *Zuse Institute Berlin, Berlin, Germany*
[b] *TU Darmstadt, Darmstadt, Germany*

Abstract

Objective: To present a novel method for automated segmentation of knee menisci from MRIs. To evaluate quantitative meniscal biomarkers for osteoarthritis (OA) estimated thereof.

Method: A segmentation method employing convolutional neural networks in combination with statistical shape models was developed. Accuracy was evaluated on 88 manual segmentations. Meniscal volume, tibial coverage, and meniscal extrusion were computed and tested for differences between groups of OA, joint space narrowing (JSN), and WOMAC pain. Correlation between computed meniscal extrusion and MOAKS experts' readings was evaluated for 600 subjects. Suitability of biomarkers for predicting incident radiographic OA from baseline to 24 months was tested on a group of 552 patients (184 incident OA, 386 controls) by performing conditional logistic regression.

Results: Segmentation accuracy measured as Dice Similarity Coefficient was 83.8% for medial menisci (MM) and 88.9% for lateral menisci (LM) at baseline, and 83.1% and 88.3% at 12-month follow-up. Medial tibial coverage was signifi-

*Corresponding author: Alexander Tack, Department of Visual Data Analysis, Zuse Institute Berlin, Berlin, Germany

Email addresses: tack@zib.de (Alexander Tack), anirban.mukhopadhyay@gris.tu-darmstadt.de (Anirban Mukhopadhyay), zachow@zib.de (Stefan Zachow)

cantly lower for arthritic cases compared to non-arthritic ones. Medial meniscal extrusion was significantly higher for arthritic knees. A moderate correlation between automatically computed medial meniscal extrusion and experts' readings was found (ρ=0.44). Mean medial meniscal extrusion was significantly greater for incident OA cases compared to controls (1.16±0.93 mm vs. 0.83±0.92 mm; $p<0.05$).

Conclusion: Especially for medial menisci an excellent segmentation accuracy was achieved. Our meniscal biomarkers were validated by comparison to experts' readings as well as analysis of differences w.r.t groups of OA, JSN, and WOMAC pain. It was confirmed that medial meniscal extrusion is a predictor for incident OA.

Keywords: Biomarker, incident OA, Cartilage, Deep Learning, Knee MRI, Statistical Shape Models

1. Introduction

Studies show a substantial correlation between meniscal abnormalities and development of radiographic OA [1, 2, 3, 4], development of cartilage loss [5, 6], and progression of cartilage loss [7, 8]. In contrast to conventional semi-quantitative (SQ) MRI readings, employing the Whole-Organ Magnetic Resonance Imaging Score (WORMS) [9], the Boston-Leeds Osteoarthritis Knee Score (BLOKS) [10], or the MRI Osteoarthritis Knee Score (MOAKS) [11], quantitative meniscal measures are needed to better classify the grade of OA and to identify patients, which have a high risk of developing OA. The potential of quantitative meniscal measures as biomarkers was shown for the prediction of incident OA [12], cartilage loss [13], and for the differentiation between arthritic and non-arthritic knees [14].

Quantitative meniscal measures (e.g. meniscal volume and tibial coverage) require a 3D segmentation of the menisci. Manual segmentation is tedious, time-consuming, and labour intensive. Durations of approximately 35 minutes have been reported for a full segmentation of a meniscus [15]. Thus, several semi-automatic and fully automated methods have been developed. Paproki et al. [16] proposed a fully automated method which utilizes an Active Shape Model scheme [17], in which the shape model is deformed using a template matching procedure minimizing the normalised-cross-correlation between intensity distribution profiles and the corresponding templates. Paproki et al. evaluated their segmentations with the help of manual segmentations provided by Imorphics (Manchester, UK). The "Imorphics gold standard" (IGS) data contains manual segmentations of cartilage (patellar cartilage, femoral cartilage, medial and lateral tibial cartilage) and menisci (medial and lateral separately) from sagittal water-excited Double-Echo Steady-State (DESS) MRI sequences of 88 knees, both, baseline and 12-month follow-up. These IGS data is publicly available[1] as part of the Osteoarthritis Initiative (OAI), which is a multi-center, longitudinal, prospective observational study of knee OA. Paproki et al. [16] reported a median Dice Similarity Coefficient (DSC) of 78.3% for medial menisci (MM) and 83.9% for lateral menisci (LM) at baseline, and 75.3% (MM) and 83.0% (LM) at 12-month follow-up. For evaluation of meniscal measures as potential OA biomarkers, they estimated meniscal volume, tibial coverage, and meniscal extrusion based on segmentation masks acquired by their menisci segmentation method as well as via segmentations of knee bone structures and bone-cartilage interface using a different method [18]. They observed significant differences for meniscal volume, tibial coverage, and meniscal extrusion between groups of radiographic OA and joint space narrowing (JSN).

[1] https://oai.epi-ucsf.org/datarelease/iMorphics.asp

Dam et al. [19] presented a method for fully automated segmentation of arthritic menisci and cartilage employing several intensity- and position-based image features in combination with k-nearest neighbors (k-NN) classification. Their evaluation of segmentation results with respect to the IGS data lead to a mean DSC of 76% (MM), 83% (LM), 81.2% for medial tibial cartilage (MTC), and 86.6% for lateral tibial cartilage (LTC).

Recently, Convolutional Neural Networks (CNNs) have shown great potential when applied to musculoskeletal image segmentation tasks. The main benefit of CNNs compared to classical hand-engineered features is, that convolutional image filters are learnt in an optimization process. Thus, they are adapted in an automated fashion for a high-level representation of the training image data in the best possible way. This requires a mask representing the objects of interests, which can be seen in the image. Prasoon et al. [20] presented an approach for tibial cartilage segmentation using three 2D CNNs, which utilize the axial, coronal, and sagittal image planes of the 3D MRI as input images. Liu et al. [22] published a method employing the CNN architecture "SegNet" [21] in combination with 3D simplex deformable modeling. They showed that the SegNet (originally developed for road and indoor scene segmentation tasks) can be successfully trained on medical images and applied to the task of cartilage and bone segmentation of the knee joint.

In this paper we present results of a fully automated segmentation method for knee menisci based on CNNs in combination with Statistical Shape Models (SSMs). We compare the segmentation accuracy of our method to a publicly available implementation of SegNet [21]. As a side note, we also report preliminary results for segmentation of tibial cartilage using CNNs only. Segmentation accuracy was evaluated with the help of the IGS data. Based on our segmentations, an analysis of meniscal volume, tibial coverage, and meniscal extrusion as

suitable biomarkers for OA has been performed. We investigated the relationship between these meniscal measurements and various OA features, i.e. JSN, OARSI OA grade, and WOMAC pain score. We further compared our automated meniscal extrusion quantification to manual SQ scoring using MOAKS by Boston Core Imaging Lab, which is part of the OAI FNIH data[2]. Finally, we investigated whether our meniscal extrusion measures are suitable indicators for incident OA.

2. Materials and Methods

In this section the cohorts, the assessed methods, as well as the evaluation metrics and experiments are described. As shown in Fig. 1, our presented method consists of three steps: I) Sagittal 2D DESS MRI slices were segmented with 2D CNNs. II) The resulting 2D masks were concatenated to a 3D mask for each MRI dataset. SSMs were adjusted in order to remove small isolated regions in the segmentation ("segmentation islands") as well as to reconstruct anatomically plausible menisci and to distinguish between medial and lateral ones. III) 3D image subvolumes were extracted from the MR image with the help of the adjusted SSM from step II. All subvolumes were then segmented by 3D CNNs. Resulting segmentation masks were fused into one 3D mask with the original MRI's dimensions for MM and LM.

The results of this "3-step" approach were compared to a "2-step" approach employing CNNs only – skipping the intermediate SSM adjustment of step II. Results of both approaches were compared to those of a publicly available implementation of SegNet [21]. This "off-the-shelf" SegNet was trained on sagittal 2D MRI slices using the IGS data in a 2-fold cross-validation setting. 2D MRI slices were then segmented and concatenated to a 3D mask.

[2]https://oai.epi-ucsf.org/datarelease/FNIH.asp

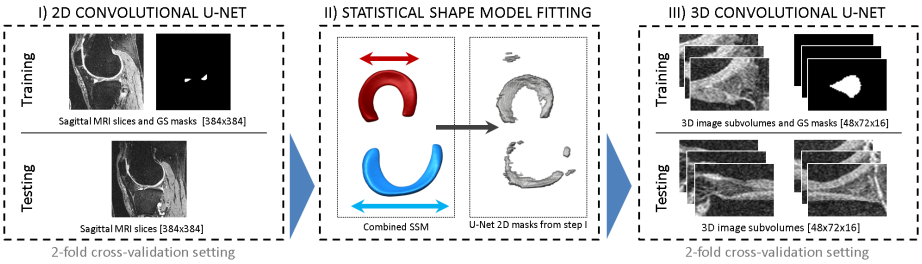

Figure 1: Illustration of the automated segmentation workflow for menisci employing CNNs and SSMs. I) Training (top): Using the gold standard (GS) data provided by iMorphics, 2D U-Nets are trained on sagittal slices for the medial meniscus (MM) and the lateral meniscus (LM). Testing (bottom): For a newly given MRI all sagittal slices are segmented using the 2D U-Nets and the resulting 2D masks are concatenated to a 3D mask. II) Only for testing: The combined SSM is adjusted to the 3D mask removing segmentation islands and yielding plausible MM and LM masks. III) Training: Using the GS data, 3D U-Nets are trained on 3D image subvolumes for MM, and LM, respectively. Testing: Using the masks from step II, 3D image subvolumes are extracted. All subvolumes are segmented by the 3D U-Net. Resulting subvolume masks are fused to one 3D mask for MM and LM, respectively.

2.1. Subjects

In this study sagittal DESS MRI data of the OAI were used. Five different datasets have been analysed. Dataset A was chosen as the IGS data cohort, consisting of 88 subjects for two timepoints, baseline and 12-month follow-up. Hence, Dataset A allows for a direct comparison of our segmentation results to the results of Paproki et al. and Dam et al. Dataset B was chosen as the 600 patients belonging to the baseline timepoint of the OAI FNIH MOAKS SQ reading study[3]. Dataset C was drawn from the OAI as 184 subjects with Kellgren Lawrence Grade (KLG) 0 or 1 at baseline. These knees developed KLG$\geq$2 with joint space narrowing within 24 months. Dataset D was randomly drawn from the OAI as a control group of 184 subjects, which did not develop radiographic OA. The motivation for choosing Dataset E was to assess the influence of the patients' BMIs on our measures and to be comparable to Emmanuel et al. (controls with a mean BMI of 27.6$\pm$0.3 kg/m^2) [12]. Thus, Dataset E was drawn from the OAI database for patients who were not developing radiographic OA

[3]https://oai.epi-ucsf.org/datarelease/FNIH.asp

and whose BMIs were densely distributed around the cohort's mean. Demographic data of these datasets is given in Table I.

[Table I (Demographic data) here]

2.2. Statistical Shape Models of menisci

SSMs were used to improve the segmentations of the 3D CNNs. We developed three different SSMs: One for the lateral meniscus (SSM_{lat}), one for the medial meniscus (SSM_{med}), and a combined one including the two (SSM_{comb}). Triangulated meshes were generated directly from the IGS data. To establish point correspondences between these meshes we used the method developed by Lamecker [23]. All three SSMs were established using Principal Component Analysis. Two datasets were rejected from integration into the medial and the combined SSM, because of missing medial anterior and medial posterior horns (patient-IDs: 9311328 and 9750920).

2.3. Automated segmentation of menisci from MRI data

Menisci were segmented using a combination of CNNs and SSMs. Therefore, two different U-Nets were employed: A 2D U-Net and a 3D U-Net, both heavily inspired by the original U-Net architecture [24]. Both networks were trained in a 2-fold cross-validation setting on the IGS data. The data was split numerically sorted by the patient IDs into the first and second half. We evaluated two

different segmentation pipelines to investigate the influence of the SSM adjustment step: In the "2-step" approach, the 3D image subvolumes were computed according to the 2D U-Net segmentation masks from step I. In the "3-step" approach, the subvolumes were computed according to the SSM segmentation masks from step II.

2.3.1. Step I: 2D U-Net

For an unseen MR image, the first step is the segmentation of all sagittal slices employing a 2D U-Net. The 2D U-Net was trained on sagittal 2D MRI slices containing the MM or LM. The respective architecture is shown in Fig. 2. The network was implemented in Keras [25] with the backend Theano [26] using the stochastic Adam optimizer [27] (learning rate = 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-08}$) to maximize the Dice Similarity Coefficient, $\mathrm{DSC}(\mathcal{X}, \mathcal{Y}) = \frac{2|\mathcal{X} \cap \mathcal{Y}|}{|\mathcal{X}|+|\mathcal{Y}|}$, between the network's 2D mask ($\mathcal{Y}$) and the IGS data ($\mathcal{X}$) [28].

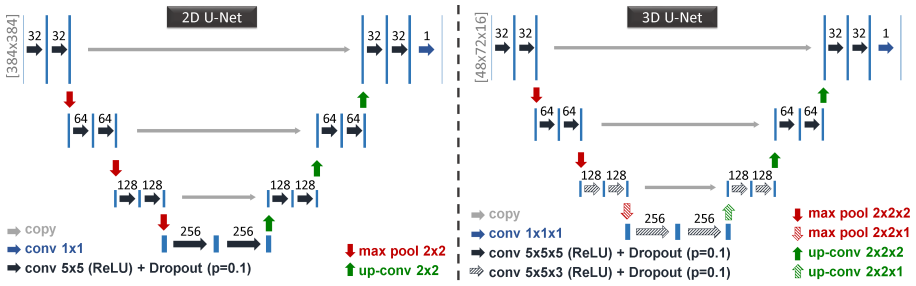

Figure 2: Architecture of the 2D convolutional U-Net (left) and the 3D U-Net (right). The inputs of the networks are sagittal MRI slices (dimension: 384×384) in the 2D case, and image subvolumes (dimension: 48×72×16) in the 3D case. Both U-Nets have four convolutional layers with an increasing number of convolutional filters (32, 64, 128, 256) employing rectified linear units (ReLU) and dropout layers (probability: 10%).

2.3.2. Step II: SSM adjustment

An SSM_{comb} not containing the respective geometry (2-fold cross-validation) was adjusted to the 3D segmentation mask from step I. Both, the transformation and the SSM's shape modes were adjusted iteratively. A simple appearance

model based on binary masks (background = 0, menisci = 1) was applied to drive the deformation of the SSM with respect to the mask's gradient.

2.3.3. Step III: 3D U-Net

A prerequisite for the third step is an inital segmentation mask, either acquired by employing the 2D U-Net (2-step approach), or by adjusting an SSM to the 2D U-Net's results (3-step approach). For each slice of the initial segmentation mask, 3D subvolumes were generated at the center of mass of each connected component. These 3D subvolumes (dimension: 48×72×16, where x-dimension is superior-inferior, y-dimension is anterior-posterior, and z-dimension is lateral-medial) were then segmented by a 3D U-Net (see Fig. 2). The resulting (partially overlapping) segmentation masks of all 3D image subvolumes were combined into one 3D mask using a majority voting scheme. This means, that in this resulting mask only those voxels were considered as meniscal tissue, which had a majority of subregions classifying them as such.
3D U-Nets were trained on 3D subvolumes containing MM and LM, respectively. These subvolumes were computed for training with the help of the IGS masks. The same parameters as in the 2D case were used for optimization.

2.4. Automated segmentation of tibial cartilage from MRI data

In order to calculate the coverage of the tibial cartilage by the menisci, as well as to calculate the meniscal extrusion relative to the tibial plateau, tibial cartilage was also segmented using a 3D U-Net only. The procedure is similar to the approach described for the menisci in 2.3. 3D image subvolumes (64×32×32) were extracted based on the automated tibial bone segmentation (see [29]). The 3D U-Nets were trained in a 2-fold cross-validation setting on the IGS data for

medial and lateral tibial cartilage. The 3D U-Net architecture was kept very similar to the architecture used for segmentation of menisci (cf. Fig. 2) being only adjusted such that the y and z dimensions are not downsampled for the last two layers and by using 5×3×3 convolutional filters instead of 5×5×3.

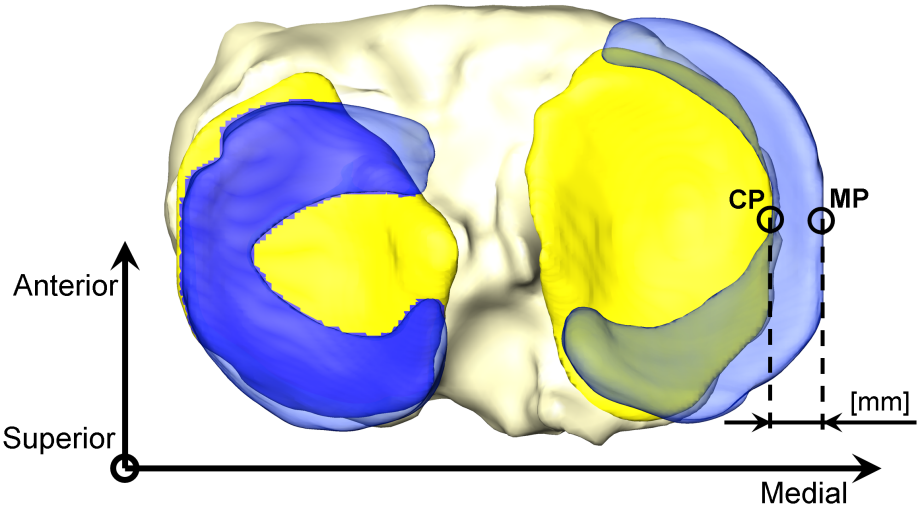

Figure 3: Tibial coverage is estimated as the ratio of tibial cartilage covered and uncovered by the meniscus (left). Meniscal extrusion is estimated as the maximum distance in epicondyle direction between the outermost tibial cartilage point CP and the outermost meniscal point MP (right).

2.5. Quantitative meniscal measurements

Meniscal volume, tibial coverage, and meniscal extrusion were calculated based on the masks of the proposed segmentation methods for menisci and cartilage.

Meniscal volume is directly given by integration of the respectively labeled voxels of the segmentation. Tibial coverage is estimated as the ratio of tibial cartilage covered and uncovered by the meniscus, for the lateral and medial side, respectively (Fig. 3, left). Here, alignment of the bone with the scan-direction was assumed.

Degeneration of menisci is characterized by an altered meniscal morphology as

well as a change in meniscal position in medio-lateral and/or anterio-posterior direction. The reference for calculating the meniscal extrusion is usually the native edge of the tibial plateau excluding any possible osteophytes. As shown for the medial tibia (Fig. 3, right), we defined these medial and lateral edges according to our cartilage masks. Our measurements for meniscal extrusion include both, medial extrusion for medial menisci and lateral extrusion for lateral menisci along the direction of the epicondylar axis (Fig. 3, right).

2.6. Validation

Segmentation accuracy was evaluated with the DSC, the Hausdorff distance (d_H), and the mean average surface distance (d_{MASD}). Let $\mathcal{X}$ be the set of voxels representing the volume of the gold standard mask, and $\mathcal{Y}$ the set of voxels of our 3D segmentation mask, we use

$$\mathcal{N}_x = \underset{y \in \mathcal{Y}}{\arg\min} \|x - y\|_2 \quad \text{and} \quad \mathcal{N}_y = \underset{x \in \mathcal{X}}{\arg\min} \|x - y\|_2 \tag{1}$$

to denote the nearest neighbour of x in $\mathcal{Y}$ and of y in $\mathcal{X}$, respectively. The Hausdorff distance d_H is the greatest of all the distances between the voxels in $\mathcal{X}$ and the respective closest voxel in $\mathcal{Y}$ and vice versa. It is defined as

$$d_H(\mathcal{X}, \mathcal{Y}) = \max \left\{ \max_{x \in \mathcal{X}} \|x - \mathcal{N}_x\|_2, \max_{y \in \mathcal{Y}} \|y - \mathcal{N}_y\|_2 \right\} . \tag{2}$$

The mean average surface distance d_{MASD} computes the average of the distances from the voxels of the gold standard mask to the nearest voxel in our segmentation as well as the distance from the voxels of our mask to the nearest voxel of the gold standard. It is defined as

$$d_{MASD}(\mathcal{X}, \mathcal{Y}) = \frac{1}{\|\mathcal{X}\| + \|\mathcal{Y}\|} \left(\sum_{x \in \mathcal{X}} \|x - \mathcal{N}_x\|_2 + \sum_{y \in \mathcal{Y}} \|y - \mathcal{N}_y\|_2 \right) . \tag{3}$$

Thus, d_{MASD} is considering the two-sided Euclidean distance between the voxels of the gold standard and our mask, as implemented in AmiraZIBEdition [30]. We evaluated the differences between groups of OARSI OA grades (rOA), JSN grades, and WOMAC pain scores for the estimated meniscal measures using unpaired t-tests for Dataset B in MATLAB R2014b (The MathWorks Inc., Natick, Massachusetts, United States). Also, non-parametric Spearman correlation analysis was performed to assess the association between meniscal measures and the respective groups. We further assessed the meniscal extrusion estimated from our segmentation masks compared to MOAKS SQ reading using non-parametric Spearman's correlation coefficients (ρ) in IBM SPSS Statistics 24 (IBM SPSS Statistics, Armonk, NY, USA). To test the hypothesis that our automated measures predict incident radiographic OA, meniscal extrusion was calculated for all subjects of Dataset C, Dataset D, and Dataset E. Conditional logistic regression adjusted for age, sex, and BMI was performed to compare cases of incident OA (Dataset C) with the controls of Dataset D as well as of Dataset E.

To generate a qualitative visualization of the d_{MASD} error comparable to Paproki et al., SSM_{lat} and SSM_{med} were adjusted to our 3D segmentation masks (Appendix A.2). The d_{MASD} was calculated between the adjusted SSMs and the IGS data. Employing the correspondence of the SSMs' vertices, the d_{MASD} was averaged for each SSM vertex over each subject of Dataset A.

For segmentation of the tibial cartilage, the DSC values were calculated with the help of the IGS data, averaged over all 88 subjects of Dataset A.

[Table II (Segmentation accuracy) here]

3. Results

In this section we present results on segmentation accuracy, computational performance of our fully automated approach, as well as suitability of derived measures as biomarkers for OA.

3.1. Segmentation accuracy

For the SegNet, the segmentation accuracy in terms of the DSC was 80.87±3.92% for both menisci, MM and LM, at baseline and 79.92±4.75% at follow-up (Table II). Employing the presented 3-step method the mean DSC increased to 83.84±6.10% (MM) and 88.86±2.39% (LM) at baseline, and 83.14±6.28% (MM) and 88.25±3.08% (LM) at 12-month follow-up. There was no subject with a DSC value $< 55\%$ (Fig. 4). The d_{MASD} was 0.43±0.58 mm (MM) and 0.23±0.08 mm (LM) at baseline and 0.46±0.65 mm (MM) and 0.25±0.09 mm (LM) at 12-month follow-up. The averaged d_{MASD} is displayed in a color coded manner on top of the mean SSM shape in A.2.

Segmentation of the tibial cartilage lead to a mean DSC of 85.13±10.5% for medial tibial cartilage (MTC) and 90.23±4.64% for lateral tibial cartilage (LTC) at baseline, and 85.86±5.03% (MTC) and 90.2±2.64% (LTC) at 12-month follow-up.

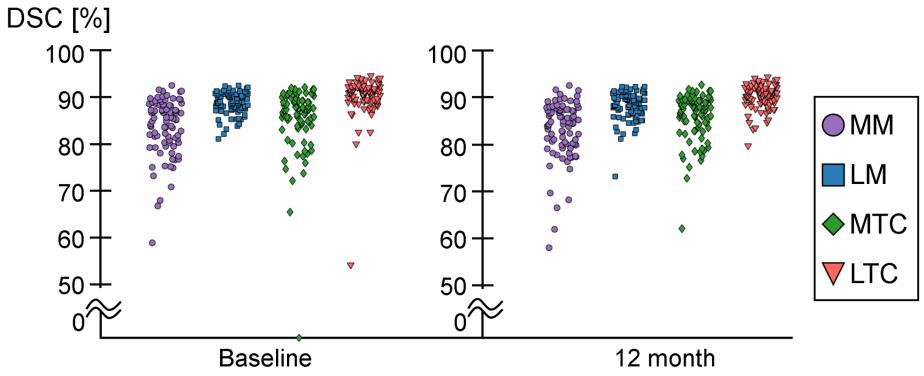

Figure 4: Dataset A: Scatter plot of the resulting DSC values as yielded by the proposed 3-step method for medial menisci (MM), lateral menisci (LM), medial tibial cartilage (MTC), and lateral tibial cartilage (LTC) for the baseline timepoint (left) and the 12-month follow-up visit (right). For one outlier the MTC was not segmented at all (baseline) due to an inaccurate tibial bone segmentation.

3.2. Computational performance

The entire segmentation ran on a PC (GPU: GeForce GTX 1080 Ti; CPU: two Intel Xeon E5-2650 v2) in a fully automated fashion. Bone segmentation [29] took about 3 minutes for one knee. Our method for menisci segmentation took about 1.5 minutes per pair of MM and LM on a personal computer. It is thus very time efficient compared to Paproki et al. (27.2 min). Tibial cartilage segmentation took about one minute per knee.

[Table III (Prediction of incident OA) here]

3.3. Meniscal measures as biomarkers for OA, JSN, and WOMAC pain score

Results for quantitative meniscal measures are given in Appendix A.1. We found no significant differences with respect to meniscal volume between groups of rOA, JSN, and WOMAC grades for both, MM and LM. In a multitude of comparisons medial tibial coverage was significantly lower between groups of higher rOA, JSN, and WOMAC pain scores and groups of lower rOA, JSN, and WOMAC pain scores. In addition, a significantly lower lateral tibial coverage was found comparing lateral JSN 1 with lateral JSN 0 ($p = 0.016$). Greater medial meniscal extrusion was associated with advanced rOA and JSN grades and significant differences were found between groups of rOA and JSN.

We found a moderate Spearman's correlation between our meniscal extrusion measurements and the MOAKS experts' SQ reading for MM ($\rho = 0.439$), but only a weak correlation for LM ($\rho = 0.11$).

Comparing cases of incident OA with control knees of Dataset D, non-significant tendencies of higher meniscal volume and greater medial meniscal extrusion were observed (Table III). For Dataset E significantly greater medial meniscal extrusion ($p = 0.001$), less tibial coverage ($p = 0.039$), as well as less lateral meniscal volume ($p = 0.016$) were found compared to incident OA cases. As computed by conditional logistic regression, the adjusted odds ratio (OR) was 1.51 (95% CI: 1.18,1.94) for medial meniscal extrusion. Mean lateral meniscal extrusion values were negative (relative to the tibial plateau) for both, incident and control cases. Incident OA cases had significantly less lateral meniscal extrusion ($p = 0.002$).

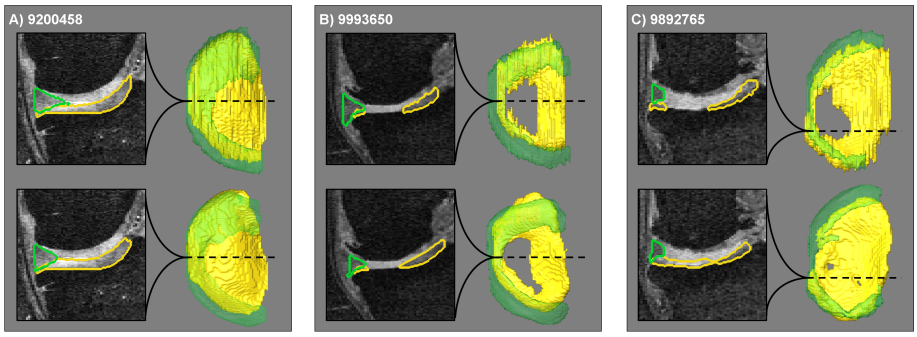

Figure 5: Comparison of IGS segmentations (upper row) to our automated results (lower row) for three subjects. Outlines of meniscal (green) and tibial cartilage segmentations (yellow) are shown on the left side for MRIs zoomed to the medial compartement. On the right side a corresponding rendering of the medial tibial cartilage (yellow) as well as the medial meniscus (green, transparent) are visible (view from axial). Our methods yielded a qualitatively good agreement with the IGS data in A) for a relatively healthy cartilage and in B) for a cartilage containing a full thickness denudation. In C) our method misclassified voxels in the area of a denudation as cartilage.

4. Discussion

Our method for automated segmentation of knee menisci and tibial cartilage allowed us to compute a set of meniscal biomarkers, i.e. meniscal volume, tibial coverage, and meniscal extrusion. Assuming a reliable gold standard segmentation, the accuracy of these measures depends on the ability of the CNNs to learn the appearance of menisci and cartilage, and to apply this knowledge to unseen data. Accurate meniscal measures facilitate the analysis of the association between meniscal pathology and the development or progression of OA.

In terms of segmenation accuracy, the 2D U-Net performed slightly better than the SegNet, motivating our choice of the U-Net for the presented approach (Table II). Compared to the application of only a 2D U-Net, our 2-step approach improved the results. As applied in our 3-step approach, the SSM plausibly corrects the results of the 2D U-Net step, and thus clearly reduced the maximum errors. This allowed for a more precise definition of 3D image subregions for the final 3D U-Net step and clearly improved the accuracy compared to the 2-step approach. The achieved DSC values are higher compared to the results

reported in the literature, especially for MM. A qualitative comparison of our d_{MASD} errors (A.2) with that of Paproki et al. indicates that our accuracy is higher for the medial and lateral body. However, Paproki et al. achieved a qualitatively higher accuracy at the meniscal horns. Our first cartilage segmentation results are promising and seem to be sligthly better than the results reported by Dam et al. [19]. We plan to extend the cartilage segmentation method by femoral and patellar cartilage and to perform further investigations with respect to cartilage measures and OA.

Among the outliers for MM (Fig. 4) we found the patients 9311328 (DSC=58.8%) and 9750920 (DSC=66.8%), which we omitted for SSM generation. For these two cases delineating the meniscus is especially challenging and the DSC value will decrease fast due to its small volume.

We found significant differences between groups of OA grades and medial JSN for medial tibial coverage and medial meniscal extrusion, indicating a high correlation. For the lateral meniscus our only observation was a lower tibial coverage for cases of JSN 1 compared to no JSN. In contrast to Paproki et al. we did not find significant differences with respect to meniscal volume for MM and LM. While Paproki et al. found no differences for groups of WOMAC pain scores, we did find such differences, suggesting that meniscal degeneration might be related to pain, after all.

Spearman's correlation was calculated between automated meniscal extrusion measures and MOAKS SQ readings for 600 subjects. Only a weak correlation was found for LM, and a moderate correlation for MM (ρ=0.439). The correlation of our medial meniscal extrusion values is comparable to the one reported by Bloecker et al. for knees with JSN [15]. They investigated the correlation between meniscal measures and MOAKS SQ analysis (ρ ranging from 0.39 to 0.56 for JSN knees and from 0.33 to 0.72 for no-JSN knees) based on *manual*

segmentations of medial menisci and medial tibial cartilage.

In terms of the prediction of incident OA using meniscal measurements, our results confirm the findings of previous studies, supporting the hypothesis that medial meniscal extrusion is an indicator for incident OA. As recently shown by van der Voet et al. [31] medial meniscal extrusion according to MOAKS scoring was significantly associated with a higher incidence of knee OA in a cohort of middle-aged overweight and obese women (BMI$\geq$27 kg/m^2). Emmanuel et al. chose a cohort with very little standard deviation of the BMI (27.6$\pm$0.3 kg/m^2) and they showed that quantitative measures of medial meniscal extrusion predicts incident OA. These measures were, however, computed based on manual segmentations. We wanted to investigate the differences in our measurements between controls in Dataset D (randomly drawn, BMI 29.6$\pm$4.4 kg/m^2) and controls in Dataset E (selectively drawn, BMI 27.5$\pm$0.7 kg/m^2). Despite adjustment of the conditional logistic regression for the BMI, we found significant outcomes only for Dataset E. To our knowlege, this is the first time the potential of meniscal measurements derived from a fully automated segmentation was shown for the prediction of incident OA and our method will allow for a further assessment of the relationship of the patient's BMI, meniscal extrusion, and incident OA.

The major limitation of our study is that all segmentation methods were optimized for sagittal DESS MRI data. Applying the proposed CNN framework to different MRI sequences requires reliable training data for supervised learning. It would be especially interesting to investigate if coronal weDESS MRIs increase the accuracy for automated meniscal extrusion measurements. Further, the segmentation of intermediate-weighted Turbo Spin Echo (IW TSE) MRIs would enable research of automated meniscal tear detection. A further limitation is cartilage denudations, decreasing the accuracy of the tibial cover-

age as well as the meniscal extrusion measures. As illustrated in Fig. 5, our CNN-based method for segmentation of tibial cartilage is in principle capable to skip full thickness cartilage denudations during classification. However, as our method is fully automated, errors do occur and for some cases there was a mismatch between our results and the Imorphics gold standard segmentation (c.f. Fig. 5c). Moreover, we considered only full thickness cartilage denudations for the computation of tibial coverage and meniscal extrusion. In future work, the influence of adding knowledge of partial denudations into the computation (i.e. cartilage thickness $\leq$ 0.5mm) will be evaluated. Also, with our current definition of meniscal extrusion, areas of cartilage denudations may lead to an overestimation of meniscal extrusion, which is especially important for longitudinal analysis of moderate or severe OA cases. For this reason, a method which detects the tibial plateau independent of the presence or absence of tibial cartilage should be pursued.

CNNs as a technique of deep learning are a powerful tool for detecting anatomical structures in medical image data. The existence of sufficient training data (gold standard segmentations) would enable the automated segmentation of the whole knee, i.e. cartilage, bones, menisci, and ligaments. With the development of increasingly powerful and complex deep learning algorithms the results will certainly approach the quality of an expert's segmentation – with the additional possibility to enforce three-dimensional smoothness of the structures. These segmentations could be a basis for research with respect to automated OA scoring. While OA features could be quantified similarly to MOAKS scoring, additional computed features, which are hard to perceive in 2D (e.g. surface curvatures, intensity distributions) could also be employed. While manual semi-quantitative MRI reading is still the method of choice, these measurements could be computed automatically for large cohorts and might have the potential of a higher

sensitivity with respect to OA progression or the prediction of incident OA.

Computational Reproducibility

The 2D and 3D convolutional U-Net code is available via github.com/AlexanderTack/Menisci-Segmentation and the weights of the trained networks via doi.org/10.12752/4.TMZ.1.0.

Acknowledgement

We would like to thank the reviewers for their valuable comments and suggestions which considerably improved the clarity and quality of this manuscript. The authors gratefully acknowledge the financial support by the German federal ministry of education and research (BMBF) research network on musculoskeletal diseases, grant no. 01EC1408B (Overload/PrevOP).

The OAI 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.

Data provided from the FNIH OA Biomarkers Consortium Project are made possible through grants and direct or in-kind contributions by: AbbVie; Amgen; Arthritis Foundation; Artialis; Bioiberica; BioVendor; DePuy; Flexion Therapeutics; GSK; IBEX; IDS;Merck Serono; Quidel; Rottapharm — Madaus; Sanofi; Stryker; the Pivotal OAI MRI Analyses (POMA) study, NIH HHSN2682010000

21C; and the Osteoarthritis Research Society International.

Author contributions

AT, AM and SZ designed the study. AT and AM implemented the proposed methods. AT collected the data, performed the statistical evaluation and executed the experiments. SZ obtained the funding resources for this project. AT, AM and SZ reviewed the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

[1] Englund M, Guermazi A, Roemer FW, Aliabadi P, Yang M, Lewis CE, *et al.* Meniscal tear in knees without surgery and the development of radiographic osteoarthritis among middle-aged and elderly persons: The multicenter osteoarthritis study. Arthritis Rheum 2009;60(3):831–839

[2] Lohmander LS, Englund PM, Dahl LL, Roos EM. The long-term consequence of anterior cruciate ligament and meniscus injuries. Am J Sports Med 2007;35(10):1756–1769

[3] Badlani JT, Borrero C, Golla S, Harner CD, Irrgang JJ. The effects of meniscus injury on the development of knee osteoarthritis. Am J Sports Med 2013;41(6):1238–1244

[4] Englund M, Roemer FW, Hayashi D, Crema MD, Guermazi A. Meniscus pathology, osteoarthritis and the treatment controversy. Nat Rev Rheumatol 2012;8(7):412–419

[5] Hunter DJ, Zhang YQ, Niu JB, Tu X, Amin S, Clancy M, *et al.* The association of meniscal pathologic changes with cartilage loss in symptomatic knee osteoarthritis. Arthritis Rheum 2006;54(3):795–801

[6] Alizai H, Roemer FW, Hayashi D, Crema MD, Felson DT, Guermazi A. An update on risk factors for cartilage loss in knee osteoarthritis assessed using MRI-based semiquantitative grading methods. Eur Radiol 2015;25(3):883–893

[7] 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

[8] 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

[9] Peterfy CG, Guermazi A, Zaim S, Tirman PFJ, Miaux Y, White D, *et al.* Whole-organ magnetic resonance imaging score (WORMS) of the knee in osteoarthritis. Osteoarthritis Cartilage 2004;12:177–190

[10] Hunter DJ, Lo GH, Gale D, Grainger AJ, Guermazi A, Conaghan PG. The reliability of a new scoring system for knee osteoarthritis MRI and the validity of bone marrow lesion assessment: BLOKS (Boston–Leeds Osteoarthritis Knee Score). Ann. Rheum. Dis. 2008;67:206–211

[11] Hunter DJ, Guermazi A, Lo GH, Grainger AJ, Conaghan PG, Boudreau RM, *et al.* Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score) Osteoarthritis Cartilage 2011;19:990–1002

[12] Emmanuel K, Quinn E, Niu J, Guermazi A, Roemer F, Wirth W, *et al.* Quantitative measures of meniscus extrusion predict incident radiographic knee osteoarthritis - data from the Osteoarthritis Initiative. Osteoarthritis Cartilage 2016;24(2):262–269

[13] Bloecker K, Wirth W, Guermazi A, Hunter DJ, Resch H, Hochreiter J, *et al.* Relationship between medial meniscal extrusion and cartilage loss in specific femorotibial subregions: data from the osteoarthritis initiative. Arthritis Care Res 2015;67(11):1545–1552

[14] Wirth W, Frobell RB, Souza RB, Li X, Wyman BT, Graverand L, *et al.* A three-dimensional quantitative method to measure meniscus shape, position, and signal intensity using MR images: A pilot study and preliminary results in knee osteoarthritis. Magn Reson Med 2010;63(5):1162–1171

[15] Blöcker K, Guermazi A, Wirth W, Kwoh CK, Resch H, Hunter DJ, *et al.* Correlation of semiquantitative vs quantitative MRI meniscus measures in osteoarthritic knees: results from the Osteoarthritis Initiative. Skeletal Radiol 2014;43(2):227–232

[16] Paproki A, Engstrom C, Chandra SS, Neubert A, Fripp J, Crozier S. Automated segmentation and analysis of normal and osteoarthritic knee menisci from magnetic resonance images–data from the Osteoarthritis Initiative. Osteoarthritis Cartilage 2014;22(9):1259–1270

[17] Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models-their training and application. Comp Vis Image Und 1995;61(1):38–59

[18] Fripp J, Crozier S, Warfield SK, Engstrom C, Ourselin S. Automatic segmentation of the bone and extraction of the bone–cartilage interface from magnetic resonance images of the knee. Phys Med Biol 2007;52(6):1617–1631

[19] Dam EB, Lillholm M, Marques J, Nielsen M. Automatic segmentation of high-and low-field knee MRIs using knee image quantification with data from the osteoarthritis initiative. J Med Imaging 2015;2(2):024001–024001

[20] Prasoon A, Petersen K, Igel C, Lauze F, Dam EB, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. MICCAI 2013:246–253

[21] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 2015

[22] Liu F, Zhou Z, Jang H, Samsonov A, Zhao G, Kijowski R. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. Magn Reson Med 2017

[23] Lamecker H. Variational and statistical shape modeling for 3D geometry reconstruction (Dissertation). Freie Universitaet Berlin 2008

[24] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation MICCAI 2015; 234–241

[25] Chollet, François. Keras. `https://github.com/fchollet/keras` 2015

[26] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints 2016;abs/1605.02688:`https://github.com/fchollet/keras` 2015

[27] Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014

[28] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3): 297–302

[29] Seim H, Kainmueller D, Lamecker H, Bindernagel M, Malinowski J, Zachow S. Model-based Auto-Segmentation of Knee Bones and Cartilage in MRI Data. Proc of Medical Image Analysis for the Clinic: A Grand Challenge 2010:215–223

[30] AmiraZIBEdition 2017: `https://amira.zib.de`

[31] van der Voet JA, Runhaar J, van der Plas P, Vroegindeweij D, Oei EH, Bierma-Zeinstra SMA. Baseline meniscal extrusion associated with incident knee osteoarthritis after 30 months in overweight and obese women. Osteoarthritis Cartilage 2017

Table I: Summary of Dataset A (2 timepoints: baseline and 12-month follow-up visit (12m)), Dataset B (baseline), Dataset C (baseline), Dataset D (baseline), and Dataset E (baseline). Listed is the Body Mass Index (BMI), the radiographic OA grade (rOA grade), which is the OARSI OA grade for Dataset A and B and the Kellgren-Lawrence grade for Dataset C, D, and E, medial Joint Space Narrowing (mJSN), lateral Joint Space Narrowing (lJSN), WOMAC pain scores, MOAKS medial meniscal extrusion in medial direction ($\text{Extrusion}_{\text{MM}}$), and MOAKS lateral meniscal extrusion in lateral direction ($\text{Extrusion}_{\text{LL}}$). Extrusion grades are defined by the scoring system MOAKS as Grade 0: < 2mm, Grade 1: $2.0 - 2.9$mm, Grade 2: $3.0 - 4.9$mm, and Grade 3: $>= 5.0$mm.

	Dataset A	**Dataset B**	**Dataset C**	**Dataset D**	**Dataset E**
Number of subjects	88	600	184	184	184
Sex (male;female)	(45,43)	(247,353)	(70,114)	(70,114)	(70,114)
Age [years]	61.24 ± 9.98	61.55 ± 8.88	61.49 ± 8.61	60.61 ± 9.05	61.17 ± 9.49
BMI [kg/m^2]	31.06 ± 4.61	30.72 ± 4.78	29.60 ± 4.43	27.64 ± 4.50	27.52 ± 0.73
rOA grade (0;1;2;3;4)	(0,0,15,56,17)	(15,154,164,242,25)	(34,150,0,0,0)	(34,150,0,0,0)	(35,149,0,0,0)
mJSN score (0;1;2)	(16,55,17)	(280,292,28)	(136,47,1)	(152,32,0)	(154,30,0)
lJSN score (0;1;2)	(74,14,0)	(543,57,0)	(163,21,0)	(164,20,0)	(166,18,0)
WOMAC pain score (0;1-10;11-20)	(0,0,0)	(233,348,19)	(77,100,7)	(94,87,3)	(92,91,1)
MOAKS $\text{Extrusion}_{\text{MM}}$ (0;1;2;3)	n.a.	(202, 176, 164, 56)	n.a.	n.a.	n.a.
MOAKS $\text{Extrusion}_{\text{LL}}$ (0;1;2;3)	n.a.	(577, 7, 15, 1)	n.a.	n.a.	n.a.
Timepoints	Baseline, 12m	Baseline	Baseline	Baseline	Baseline

Table II: Segmentation accuracies for SegNet [23], a combination of 2D U-Net and 3D U-Net, and the proposed 3-step segmentation, employing 2D U-Nets, SSMs, and 3D U-Nets. All methods were run in a 2-fold cross-validation setting. DSC, d_H and d_{MASD} were calculated for the resulting volumetric segmentation masks compared to the ground standard segmentations for two timepoints, baseline and 12 month follow-up. Mean values ± standard deviation are supplied for the first 44 subjects (first 44), last 44 subjects (last 44), and averaged over all 88 subjects (all 88).

Baseline	DSC [%]	d_H [mm]	d_{MASD} [mm]	**12 months**	DSC [%]	d_H [mm]	d_{MASD} [mm]
2D SegNet				**2D SegNet**			
MM & LM				*MM & LM*			
- first 44	80.43±4.04	8.04±5.60	0.43±0.17	- first 44	79.57±4.91	7.90±6.05	0.44±0.16
- last 44	81.30±3.78	10.15±7.92	0.44±0.14	- last 44	80.27±4.62	15.94±18.84	0.48±0.20
- all 88	**80.87±3.92**	**9.09±6.90**	**0.44±0.16**	- all 88	**79.92±4.75**	**11.92±14.48**	**0.46±0.18**
2D U-Net + 3D U-Net				**2D U-Net + 3D U-Net**			
I) 2D U-Net				**I) 2D U-Net**			
MM & LM				*MM & LM*			
- first 44	81.39±4.68	15.02±13.09	0.46±0.21	- first 44	80.87±5.09	14.11±13.02	0.46±0.15
- last 44	84.28±3.75	9.03±11.26	0.36±0.13	- last 44	83.27±4.28	10.74±10.94	0.40±0.20
- all 88	**82.84±4.46**	**12.03±12.51**	**0.41±0.18**	- all 88	**82.07±4.83**	**12.42±12.08**	**0.43±0.18**
II) 3D U-Net				**II) 3D U-Net**			
MM				*MM*			
- first 44	75.95±11.33	10.78±11.38	1.45±2.40	- first 44	72.53±16.97	11.64±12.65	2.36±6.19
- last 44	78.68±14.39	8.17±10.67	1.82±7.88	- last 44	78.81±10.21	8.52±9.48	1.05±2.49
- all 88	**77.32±12.95**	**9.48±11.05**	**1.63±5.79**	- all 88	**75.67±14.28**	**10.08±11.23**	**1.71±4.73**
LM				*LM*			
- first 44	83.10±4.34	5.16±2.32	0.40±0.21	- first 44	82.92±5.22	5.61±3.42	0.43±0.29
- last 44	85.88±3.69	6.14±5.29	0.39±0.32	- last 44	86.44±3.41	4.37±2.12	0.30±0.11
- all 88	**84.49±4.24**	**5.65±4.09**	**0.40±0.27**	- all 88	**84.68±4.73**	**4.99±2.90**	**0.36±0.23**
2D U-Net + SSM + 3D U-Net				**2D U-Net + SSM + 3D U-Net**			
I) 2D U-Net				**I) 2D U-Net**			
MM & LM				*MM & LM*			
- first 44	81.39±4.68	15.02±13.09	0.46±0.21	- first 44	80.87±5.09	14.11±13.02	0.46±0.15
- last 44	84.28±3.75	9.03±11.26	0.36±0.13	- last 44	83.27±4.28	10.74±10.94	0.40±0.20
- all 88	**82.84±4.46**	**12.03±12.51**	**0.41±0.18**	- all 88	**82.07±4.83**	**12.42±12.08**	**0.43±0.18**
II) SSM fitting				**II) SSM fitting**			
MM				*MM*			
- first 44	73.32±8.64	7.12±5.14	0.82±1.35	- first 44	72.67±8.65	6.84±5.19	0.84±1.37
- last 44	73.79±7.36	6.37±4.04	0.70±0.58	- last 44	71.08±8.90	7.51±5.15	0.83±0.75
- all 88	**73.55±7.98**	**6.75±4.61**	**0.76±1.04**	- all 88	**71.87±8.76**	**7.18±5.15**	**0.83±1.10**
LM				*LM*			
- first 44	73.59±10.87	5.57±1.98	0.67±0.47	- first 44	74.30±8.56	5.84±2.00	0.64±0.38
- last 44	79.15±5.03	5.69±2.46	0.52±0.23	- last 44	78.03±5.99	6.00±2.63	0.56±0.25
- all 88	**76.37±8.87**	**5.63±2.22**	**0.60±0.38**	- all 88	**76.16±7.58**	**5.92±2.32**	**0.60±0.32**
III) 3D U-Net				**III) 3D U-Net**			
MM				*MM*			
- first 44	83.82±6.35	5.41±5.01	0.44±0.73	- first 44	83.56±6.04	4.80±1.88	0.36±0.19
- last 44	83.85±5.91	5.82±4.01	0.41±0.38	- last 44	82.72±6.55	6.78±6.88	0.56±0.90
- all 88	**83.84±6.10**	**5.61±4.51**	**0.43±0.58**	- all 88	**83.14±6.28**	**5.79±5.11**	**0.46±0.65**
LM				*LM*			
- first 44	88.28±2.58	3.79±1.22	0.24±0.07	- first 44	87.70±3.42	4.31±1.73	0.26±0.10
- last 44	89.44±2.06	3.89±2.16	0.23±0.09	- last 44	88.81±2.62	3.84±1.54	0.24±0.09
- all 88	**88.86±2.39**	**3.84±1.74**	**0.23±0.08**	- all 88	**88.25±3.08**	**4.08±1.64**	**0.25±0.09**

Table III: Meniscal measurements (means and std. dev.) for incident OA cases (Dataset C) compared to non-incident OA controls (Dataset D as well as Dataset E).

	Dataset C Mean ± SD	**Dataset D** Mean ± SD	**Diff. to C** %	**Adjusted OR*** 95% CI	**p-value**	**Dataset E** Mean ± SD	**Diff. to C** %	**Adjusted OR*** 95% CI	**p-value**
Medial meniscus									
Meniscal volume [mm^3]	2750 ± 713	2636 ± 737	4.3	1.34 (0.99, 1.81)	0.054	2676 ± 827	2.8	1.15 (0.85, 1.57)	0.361
Tibial coverage [%]	52.0 ± 5.4	52.1 ± 4.7	-0.2	1.01 (0.82, 1.26)	0.896	53.2 ± 5.7	-2.3	0.80 (0.64, 1.00)	**0.047**
Meniscal extrusion [mm]	1.16 ± 0.93	1.00 ± 0.89	16.0	1.21 (0.97, 1.50)	0.093	0.83 ± 0.92	39.8	1.51 (1.18, 1.94)	**0.001**
Lateral meniscus									
Meniscal volume [mm^3]	2649 ± 623	2554 ± 611	3.7	1.29 (0.97, 1.72)	0.080	2535 ± 593	4.5	1.46 (1.07, 1.98)	**0.016**
Tibial coverage [%]	61.5 ± 5.5	61.0 ± 5.1	0.8	1.13 (0.91, 1.40)	0.264	61.8 ± 5.6	-0.5	1.00 (0.80, 1.24)	0.978
Meniscal extrusion [mm]	-0.97 ± 0.89	-0.92 ± 0.87	-5.4	0.93 (0.76, 1.15)	0.528	-0.70 ± 0.76	-38.6	0.68 (0.54, 0.87)	**0.002**

*Adjusted Odds Ratio per 1 standard deviation increase (conditional logistic regression adjusted for age, sex, and BMI)

Appendix A

A.1 Quantitative measures for the 3-step approach

A. 1: 3-step approach of 2D U-Net, SSM, 3D U-Net. Meniscal volume V [mm^3], tibial coverage TC [%], and meniscal extrusion ME [mm] estimated from our medial menisci (MM) and lateral menisci (LM) masks for Dataset B. Differences between groups of radiographic OA rOA (no: 0,1; moderate: 2; advanced: 3,4), joint space narrowing JSN (0, 1, 2) and WOMAC pain scores (0, [1-10], [11-20]) were evaluated using two-sample t-tests (significance level: 0.05). Non-parametric Spearman correlation analysis was used to assess the association between meniscal measures and the respective groups.

Radiographic OA

				unpaired t-tests			Spearman's rho	
OA	no	moderate	advanced	no vs. moderate	no vs. adv.	moderate vs. adv.	ρ	p
MM								
V	2704.65±767.93	2665.14±763.45	2715.66±735.87	0.638	0.881	0.496	0.03	0.539
TC	49.6±7.5	51.2±6.4	44.4±8.4	**0.039**	**<0.0001**	**<0.0001**	-0.32	**<0.0001**
ME	1.2±0.9	1.1±0.8	1.6±1.0	0.195	**<0.0001**	**<0.0001**	0.21	**<0.0001**
LM								
V	2634.42±632.31	2611.67±593.08	2709.56±636.96	0.735	0.229	0.113	0.06	0.126
TC	59.0±4.9	59.6±6.1	59.0±5.5	0.349	0.956	0.298	-0.02	0.631
ME	0.0±0.2	0.0±0.2	0.0±0.2	0.347	0.780	0.468	-0.03	0.525

Joint Space Narrowing

Grade	0	1	2	0 vs. 1	0 vs. 2	1 vs. 2	ρ	p
MM								
V	2695.30±769.89	2703.38±741.24	2685.08±696.85	0.898	0.946	0.900	0.03	0.519
TC	51.5±6.0	44.9±8.5	39.3±7.1	**<0.0001**	**<0.0001**	**0.001**	-0.46	**<0.0001**
ME	1.0±0.9	1.5±1.0	2.1±0.9	**<0.0001**	**<0.0001**	**0.004**	0.32	**<0.0001**
LM								
V	2655.83±623.55	2716.98±635.83	–	0.482	–	–	0.04	0.391
TC	59.3±5.5	57.5±5.3	–	**0.016**	–	–	-0.08	**0.040**
ME	0.0±0.2	0.0±0.1	–	0.794	–	–	-0.03	0.505

WOMAC pain score

Score	0	1-10	11-20	0 vs. 1-10	0 vs. 11-20	1-10 vs. 11-20	ρ	p
MM								
V	2770.62±740.70	2660.48±760.69	2518.41±665.89	0.084	0.152	0.426	-0.08	0.055
TC	48.8±7.0	47.1±9.0	45.1±5.6	**0.019**	**0.028**	0.342	-0.09	**0.027**
ME	1.3±1.0	1.3±1.0	1.7±0.9	0.404	0.059	0.099	0.07	0.103
LM								
V	2675.99±604.23	2652.37±640.22	2655.25±603.84	0.656	0.886	0.985	-0.02	0.612
TC	59.1±5.3	59.2±5.7	59.1±4.3	0.912	0.980	0.988	0.01	0.817
ME	0.0±0.2	0.0±0.2	0.0±0.0	0.527	0.455	0.482	-0.02	0.711

A.2: Visualisation of the averaged d_{MASD}

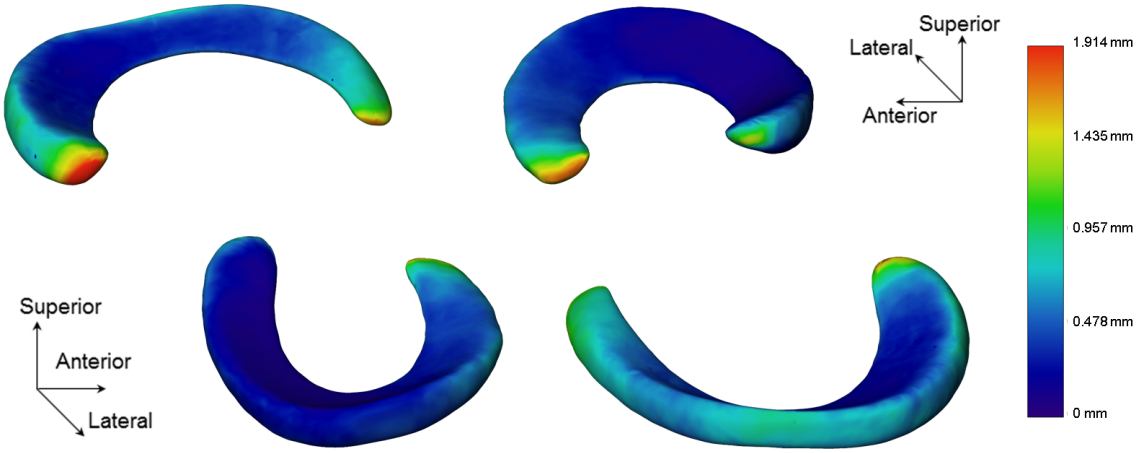

Figure 1: Mean average surface distance d_{MASD} between meshes generated from our automated menisci masks for the 3-step method (2D CNN, SSM, and 3D CNN) and meshes generated from the IGS data averaged over all 88 subjects. Our meshes were generated by adjustment of SSM_{lat} and SSM_{med} to our automated segmentation masks in order to achieve point correspondence. For comparability, we chose the same menisci depiction and colormap as Paproki et al. [16].