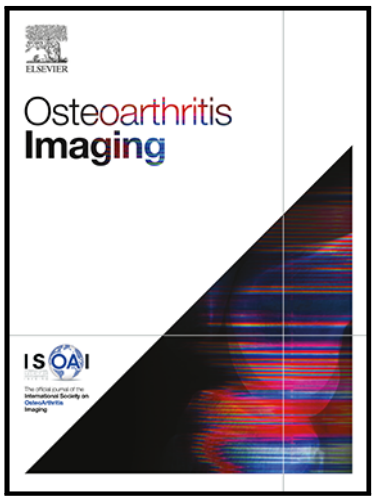
Journal Pre-proof




# Segmentation of Knee MRI Data with Convolutional Neural Networks for Semi-Automated Three-Dimensional Surface-Based Analysis of Cartilage Morphology and Composition

Dimitri A. Kessler , James W. MacKay , Stephen M. McDonnell , Robert L. Janiczek , Martin J. Graves , Joshua D. Kaggie , Fiona J. Gilbert

PII: S2772-6541(22)00004-6
DOI: https://doi.org/10.1016/j.ostima.2022.100010
Reference: OSTIMA 100010

To appear in: *Osteoarthritis Imaging*

Received date: 21 July 2021
Revised date: 12 March 2022
Accepted date: 15 March 2022

Please cite this article as: Dimitri A. Kessler , James W. MacKay , Stephen M. McDonnell , Robert L. Janiczek , Martin J. Graves , Joshua D. Kaggie , Fiona J. Gilbert , Segmentation of Knee MRI Data with Convolutional Neural Networks for Semi-Automated Three-Dimensional Surface-Based Analysis of Cartilage Morphology and Composition, *Osteoarthritis Imaging* (2022), doi: https://doi.org/10.1016/j.ostima.2022.100010

# Segmentation of Knee MRI Data with Convolutional Neural Networks for Semi-Automated Three-Dimensional Surface-Based Analysis of Cartilage Morphology and Composition

Dimitri A. Kessler [a‡], James W. MacKay [a,b], Stephen M. McDonnell [c], Robert L. Janiczek [d], Martin J. Graves [e], Joshua D. Kaggie [a*] and Fiona J. Gilbert [a*]

[a] *Department of Radiology, University of Cambridge, United Kingdom.*

[b] *Norwich Medical School, University of East Anglia, Norwich, United Kingdom.*

[c] *Division of Trauma and Orthopaedic Surgery, University of Cambridge, Cambridge, United Kingdom.*

[d] *Digital Measurement Strategy, Translational Sciences, Immunology, The Janssen Pharmaceutical Companies of Johnson & Johnson, Philadelphia, United States.*

[e] *Cambridge University Hospitals NHS Foundation Trust, Addenbrooke's Hospital, Cambridge, United Kingdom.*

*JDK and FJG share senior authorship to this work.

‡**Corresponding Author**: Dimitri A Kessler, Department of Radiology, Box 218, Cambridge Biomedical Campus, Cambridge, United Kingdom, CB2 0QQ

Tel: +44-7711266011, E-Mail: dak50@cam.ac.uk

Submitted to **Cartilage Imaging** as a ***full-length original research***



**Word Count**

Abstract: 250

Main body: ~3400

# Segmentation of Knee MRI Data with Convolutional Neural Networks for Semi-Automated Three-Dimensional Surface-Based Analysis of Cartilage Morphology and Composition


## Abstract

**Objective:** To assess automatic segmentations for surface-based analysis of cartilage morphology and composition on knee magnetic resonance (MR) images.

**Methods:** 2D and 3D U-Nets were trained on double echo steady state (DESS) images from the publicly available Osteoarthritis Initiative (OAI) dataset with femoral and tibial bone and cartilage segmentations provided by the Zuse Institute Berlin (ZIB). The U-Nets were used to perform automatic segmentation of femoral and tibial bone-cartilage structures (bone and cartilage segmentations combined into one structure) from the DESS images. $T_2$-weighted images from the OAI dataset were registered to the DESS images and used for $T_2$ map calculation. Using the 3D cartilage surface mapping (3D-CaSM) method, surface-based analysis of cartilage morphology (thickness) and composition ($T_2$) was performed using both manual and network-generated segmentations from OAI ZIB testing images. Bland-Altman analyses were performed to evaluate the accuracy of the extracted cartilage thickness and $T_2$ measurements from both U-Nets compared to manual segmentations.

**Results:** Bland-Altman analysis showed a mean bias [95% limits of agreement] for femoral and tibial cartilage thickness measurements ranging between -0.12 to 0.33 [-0.28, 0.96] mm with 2D U-Net and 0.07 to 0.14 [-0.14, 0.39] mm with 3D U-Net. For $T_2$, the mean bias [95% limits of agreement] ranged between -0.16 to 1.32 [-4.71, 4.83] ms with 2D U-Net and -0.05 to 0.46 [-2.47, 3.39] ms with 3D U-Net.

**Conclusions:** While both 2D and 3D U-Nets exemplified the time-efficiency benefit of using deep learning methods for generating the required segmentations, segmentations from 3D U-Nets demonstrated higher accuracy in the extracted thickness and $T_2$ features using 3D-CaSM compared to the segmentations from 2D U-Nets.

**Keywords:** Articular Cartilage, Bone, Segmentation, Musculoskeletal, Convolutional Neural Network, Magnetic Resonance Imaging (MRI)

**Abbreviations**: 3D-CaSM, three-dimensional cartilage surface mapping; ASSD, average symmetric surfaces distance; CNN, Convolutional Neural Network; DESS, double echo steady state; DL, deep learning; DSC, Dice similarity coefficient; KL, Kellgren-Lawrence; OA, osteoarthritis; OAI, Osteoarthritis Initiative; QMRI, quantitative magnetic resonance imaging; RMS, root-mean-squared; VOE, volumetric overlap error; ZIB, Zuse Institute Berlin.

## Introduction

Osteoarthritis (OA) is the most common disabling joint disorder. It is characterised by the progressive deterioration of articular cartilage, subchondral bone and other tissues of diarthrotic joints [1,2]. A full understanding of all the effects that contribute to OA development is lacking, hindering the development of effective interventions at early stages of disease. Quantitative magnetic resonance imaging (qMRI) methods can assist in the non-invasive detection and quantification of morphological and compositional changes present in diseases such as OA. Although several promising techniques have been validated, to date, no qMRI method has yet been regulatorily qualified as a sensitive and reliable disease biomarker. Potential qMRI biomarkers include measurement of cartilage volume and thickness, as well as the spin lattice relaxation time in the rotating frame ($T_{1\rho}$) and the transverse relaxation time ($T_2$) [3–5]. However, their clinical translation has been affected by the laborious post-processing required which almost always includes some form of image segmentation for detailed region- or tissue-specific analysis, limiting their use to primarily early-phase clinical trials [6–9].

Traditionally, compartmental measurements of cartilage morphology and composition have been performed in studies aimed at determining biomarkers for early OA detection and progression. Nevertheless, such measurements over large regions-of-interest could mask heterogeneous focal changes and are prone to inter- and intra-observer errors [10,11]. More recent studies have begun to perform analysis on

multiple smaller cartilage subregions and layers aimed at avoiding masking of important focal variations, although measurements extracted from these sub-regional compartments are even more prone to observer error [11]. As a result, cartilage surface-based analysis techniques have progressively gained more interest over recent years as these methods can determine and visualise the heterogeneity and bidirectionality of morphological and compositional changes occurring during OA progression [8,12,13]. A surface-based method called 3D cartilage surface mapping (3D-CaSM) was recently described and validated for analysing cartilage thickness and composition on MRI [8]. However, its utility for analysing large imaging cohorts is limited by requiring manually segmented bone-cartilage structures from which cartilage patches are determined and correspondent cartilage thickness maps calculated.

While manual tissue segmentation continues to be the gold-standard method for analysing qMRI data, it is very time-consuming. Therefore, interest has grown in developing consistent automated methods for multi-tissue segmentation of MR images [10,14]. Prior to the recent advent of machine learning, most proposed methods have been shape-, region- or atlas-based which require a priori knowledge of the image structures. Developments in deep learning (DL) using convolutional neural networks (CNNs) have shown great promise in overcoming the repetitive and laborious nature of manual tissue segmentation. The convolutional encoder-decoder network U-Net [15], and its 3-dimensional (3D) analogues, 3D U-Net [16] and V-Net [17], are currently regarded as state-of-the-art methods, showing high segmentation accuracy in musculoskeletal segmentation tasks [18–22].

Increasingly, studies are investigating the utility of DL methods for segmentation of musculoskeletal MR images. However very few have assessed the efficacy of the segmentations for extracting accurate qMRI values. In a study by Paproki *et al*. [23], measurements of tissue volume and $T_2$ relaxation times, extracted from automated posterior cruciate ligament segmentations using a patch-based method, achieved high correlations with those from manual segmentations. Similarly, Norman *et al*. [19] achieved high correlations between manual and automatic quantifications of cartilage and menisci morphology (volume, thickness) and composition ($T_{1\rho}$, $T_2$) using a U-Net. Liu *et al*. [18] showed no significant difference between the $T_2$ relaxation times extracted from manual and automated segmentations using SegNet [24] in combination with a 3D deformable model of femoral, tibial and patellar cartilage. A

recent study by Wirth *et al.* [25] evaluated the use of U-Net generated segmentations of femorotibial cartilage from two different high resolution 3D sequences for deriving accurate and longitudinally reproducible measures of cartilage morphology such as thickness, volume and surface area. However, these studies have used compartmental rather than surface-based analysis which may have masked the heterogeneity in segmentation accuracy for different cartilage regions.

We hypothesise that CNNs can provide a fast and accurate approach for the automated segmentation of knee MRIs that overcome the limitations of manual segmentation within the 3D-CaSM method. The purpose of this study was to analyse the usage of 2D and 3D U-Net generated segmentations of knee MR images within the 3D-CaSM method for extracting accurate regional measurements of cartilage morphology and relaxometry. By comparing the extracted morphological and compositional cartilage measures from manual and automated segmentations we expect to be able to identify those focal regions where the networks' segmentations deviate from manual.

# Methods

## Image Datasets

The cross-sectional MRI dataset used was OAI ZIB [26] consisting of multi-class tissue segmentations of femoral and tibial cartilage and bone from 507 patients from the publicly available Osteoarthritis Initiative (OAI) baseline dataset [27]. All images were acquired on Siemens 3T Trio systems using a 3D double echo steady state (DESS) sequence with water excitation. After automatically generating outlines of femoral and tibial bone and cartilage using a statistical shape model, manual adjustments were performed by experts at Zuse Institute Berlin (ZIB) [28]. The OAI ZIB cohort consisted of patients with various degrees of OA progression (Kellgren-Lawrence (KL) 0 – 4), with a majority having moderate-to-severe OA (KL ≥ 3). Detailed characteristics of the OAI ZIB cohort are in Ambellan et al [26].

### *OAI $T_2$ Mapping*

$T_2$-weighted images from the OAI dataset were acquired with a sagittal 2D multi-slice, multi-echo (MSME) spin echo (SE) sequence. Sequence parameters were: acquisition time = 10.6 min; FOV = 120 mm; matrix = 384 x 269 interpolated to 384x384; slice

thickness = 3 mm; approximately 27 slices per TE; TR = 2700 ms; TEs = 10, 20, 30, 40, 50, 60, 70 ms; bandwidth = 250 Hz/pixel.

For the analysis of cartilage $T_2$ relaxation times, all $T_2$-weighted images were rigidly registered to the 3D DESS images using the Elastix registration toolbox before calculating the quantitative $T_2$ maps.

$T_2$ maps were calculated using a log-linearised least-squares algorithm to fit a mono-exponential decay function to the signal intensities

$$S = S_0 \cdot e^{-TE/T_2}$$

$T_2$ relaxation times > 100ms in $T_2$ maps were excluded from analysis to avoid including partial volume artefacts with synovial fluid [29]–[31].

## 2D and 3D U-Net Model Specifications

The 2D U-Net adapted a previously described model used for multi-tissue segmentation of knee MR images [21]. To overcome memory constraints when training the 3D U-Net while maintaining an effective comparison between both networks, we modified the input layer of the U-Net to an input size of 256 x 256.

The 3D U-Net used a similar architecture to the 2D U-Net, except that all 2D operations were replaced with their 3D counterparts. The detailed architectures of both networks are shown in Figure 1.

## Data Preparation and Network Training

The MRIs and image masks of the OAI ZIB data were converted from their respective DICOM and ITK MetaImage Header (MHD) formats [32,33] to a MAT-file format (binary MATLAB files) before training. Each major structure was stored in a separate channel in the segmentation map. The 3D cartilage surface mapping (3D-CaSM) method described below requires delineation of the bone and cartilage in one structure. To allow a simple downstream integration within the 3D-CaSM pipeline, the OAI ZIB femoral bone and cartilage segmentations were combined into one structure, given a value code of 1 and stored in the first channel. Similarly, the tibial bone-cartilage combination was given a value code of 1 and stored in the second channel of the volume. The two-channel input to the networks had identical DESS image slices

stored in both channels. While the 2D U-Nets were trained on individual slices, for 3D network training, each dataset was split into 256 x 256 x 10 sub-volumes of overlapping slices. For instance, a single OAI 3D DESS dataset consisting of 160 slices was divided into 31 sub-volumes with each sub-volume overlapping the previous by 5 slices (except for the first and last sub-volumes, that started at slice 1 and ended at slice 160, respectively). We divided the OAI ZIB dataset into 467 cases for training, 20 cases for validation and 20 cases for testing, on which the 3D-CaSM analysis would be performed. The validation and testing sets consisted of patients covering all degrees of OA progression (KL 0 – 4), with the majority having minimal-to-moderate OA (KL 1-3). Characteristics of the testing set cohort are in Table 1.

Both networks were implemented using PyTorch (Torch v1.0.1) and all training was performed on an Nvidia P6000 GPU card (3840 CUDA cores, 24 GB GDDR5X). All models were trained for 100 epochs with the Adam solver [34] used for network optimisation with a learning rate of 0.0002 and momentum parameters, $\gamma_1$ = 0.5, $\gamma_2$ = 0.999. To avoid overfitting, we employed early stopping once the network's performance did not improve over 20 epochs assessed with the validation set. Batch sizes of 50 and 5 were used for 2D and 3D network training, respectively. The input DESS images were first resized from their original size (384 x 384) to 286 x 286 using bi-cubic interpolation followed by random cropping of the images to 256 x 256 to resemble the input layer size of the networks. The first resampling step was performed to avoid potentially cropping out regions or tissues of interest if the knee joint was not positioned centrally during acquisition. The second step allowed us to introduce random jitter for data augmentation during training.

We evaluated the performance 2D and 3D U-Nets by training with different loss functions. We trained the networks with binary cross entropy loss ($BCE$), $Dice$ loss, as well as a weighted combination of the cross entropy ($CE$) and $Dice$ loss within the $Combo$ loss ($\beta \cdot CE - (1 - \beta) \cdot Dice$) [35]. The weighting hyperparameter $\beta$ between the $CE$ and $Dice$ loss was altered to vary the balance between the two losses. We investigated values for $\beta$ = 0.25, 0.5 and 0.75. The segmentations from the best-scoring 2D and 3D networks were chosen to perform the 3D-CaSM analysis described below.

## Evaluation Metrics

After network training, a tissue-specific Boolean mask was created on the predicted test images.

We evaluated the network segmentation performances with the widely used Sørensen–Dice Similarity Coefficient ($DSC$) [36], [37], Volumetric Overlap Error ($VOE$) and the boundary distance-based metric Average Symmetric Surfaces Distance ($ASSD$). The $DSC$ ranges between 0 and 1, higher $DSC$ values representing greater overlap between the manual (X) and network generated (Y) segmentations. The $DSC$ is given as

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

for Boolean metrics. The $VOE$ also ranges between 0 and 1, however with small values for $VOE$ expressing greater segmentation accuracy. It can be calculated as

$$VOE = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

The $ASSD$ calculates the average of all distances from each pixel on the boundary of the manual segmentation X to the boundary of the automated segmentation Y and vice versa. It is expressed in mm and is defined as

$$ASSD = \frac{1}{N_X + N_Y}\left(\sum_{i=1}^{N_X} D_X(y) + \sum_{i=1}^{N_Y} D_Y(x)\right)$$

where $D_X(y) = \min_{x \in X} \|y - x\|$ and $D_Y(x) = \min_{y \in Y} \|x - y\|$.

## Postprocessing and Cartilage Surface-based Analysis

*3D Cartilage Surface Mapping*

Contours from the 2D and 3D U-Net generated masks of femoral and tibial bone-cartilage structures were extracted and converted into tissue-specific polygons within

Python (v3.7.6, Python Software Foundation, Wilmington, Delaware, United States). Detailed surface-based analysis (3D Cartilage Surface Mapping, 3D-CaSM) of the contours generated from manual and network-automated segmentations of femoral and tibial bone-cartilage structures was performed on the OAI DESS images using the StradView software v6.1 (freely available at http://mi.eng.cam.ac.uk/Main/StradView/) [8,31]. The full 3D-CaSM pipeline is illustrated in Figure 2. The femoral and tibial bone-cartilage contours are first used to generate 3D triangulated mesh surface objects (Figure 2, Step 1). By displaying the signal intensities along the normal surface on the inside of the mesh object onto the object's surface, the cartilage surfaces can be visualised. Triangulated femoral, lateral tibial and medial tibial cartilage surface patches can then be manually extracted, taking approximately two minutes for all three cartilage surfaces (Figure 2, Step 2). These patches are then used to calculate the cartilage thickness at each surface vertex by generating inner and outer cartilage surfaces (guided by the contrast difference between high (bright) cartilage and low (dark) bone signal intensities) and sampling the length of the perpendicular line between the two surfaces (Figure 2, Step 3). The thickness measurements can then be displayed onto the surface (Figure 2, Step 4).

*Analysis of Cartilage $T_2$*

Following the generation of inner and outer cartilage surfaces through the thickness measurement procedure using the 3D-CaSM method, these surfaces were used to analyse the registered quantitative $T_2$ maps. At each triangulated surface vertex, the $T_2$ relaxation time values along a perpendicular line (normal surface) between inner and outer surface were sampled and averaged (Figure 2, Steps 5 and 6).

*Cohort Analysis*

For an effective comparison, each participant's unique triangulated femoral, medial tibial and lateral tibial cartilage surface mesh was registered to a canonical (average) femoral, medial tibial and lateral tibial mesh. Subsequently, the individual thickness and $T_2$ data were mapped onto the canonical surface. Both surface registration and mapping of the individual surface data to the canonical surfaces were performed using the freely available wxRegSurf software v20 (http://mi.eng.cam.ac.uk/~ahg/wxRegSurf). Since all data were registered to a

canonical surface, vertex-wise cohort-averaged analysis was performed by calculating the root-mean-squared (RMS) thickness and $T_2$ at each canonical surface vertex from the individual cartilage thickness and $T_2$ measurements of the 20 OAI ZIB testing datasets.

### Statistical Analysis

Two-sample t-tests were used to determine whether differences in segmentation scores between 2D U-Nets and 3D U-Nets trained with equivalent loss functions were significant. For all t-test analyses, a significance level of 0.05 was used.

The RMS thickness and $T_2$ measurements at each vertex from cartilage surfaces extracted from manual segmentations were compared to the matching vertex measurements on the corresponding cartilage surfaces extracted from automated segmentations. Scatterplots and Pearson correlation coefficients were determined to evaluate any related differences. Bland-Altman analysis was performed to determine the mean bias and the 95% limits of agreement between manual and automated RMS thickness and $T_2$ measurements.

## Results

### Network Training and Testing

Training of the 2D and 3D U-Nets on the OAI ZIB dataset required approximately 500 s/epoch and 2500 s/epoch, respectively. Segmenting an entire OAI 3D DESS volume took approximately 3.2 s with 2D U-Net and 4.6 s with 3D U-Net.

### Segmentation Performance

Average scores on the OAI ZIB testing dataset from all networks trained with different loss objectives are shown in Table 2. Both 2D and 3D networks achieve high segmentation performance for femoral and tibial bone-cartilage structures compared to manual segmentations. All segmentation scores between 2D and 3D U-Nets trained with the same loss functions differed significantly ($p < 0.05$). The 2D U-Net and the 3D U-Net trained with $Combo$ loss ($\beta = 0.25$) achieved the highest segmentation performance for the femoral bone-cartilage structure. The 2D U-Net trained with $Combo$ loss ($\beta = 0.75$) and the 3D U-Net trained with $Dice$ loss-only achieved the

highest segmentation performance for the tibial bone-cartilage structure. The 3D U-Net achieved higher overall segmentation performance compared to the 2D U-Net for both femoral (DSC: 0.980 vs 0.971; ASSD: 0.314 mm vs 0.543 mm) and tibial (DSC: 0.982 vs 0.974; ASSD: 0.282 mm vs 0.412 mm) bone-cartilage structures. The segmentations from the 2D U-Net and the 3D U-Net trained with $Combo$ loss (β = 0.25) were used to perform the 3D-CaSM.

## Cartilage Surface-based Analysis

Results of the Pearson correlation and Bland-Altman analysis of thickness and $T_2$ measurements are listed in Table 3 as well as visualised in Figures 3 and 4 for cartilage thickness, and Figures 6 and 7 for cartilage $T_2$ measurements. Bland-Altman analysis showed that the mean bias [95% limits of agreement] for femoral cartilage thickness measurements was 0.33 [-0.28; 0.96] with 2D U-Net, but only 0.07 [-0.11; 0.25] mm with 3D U-Net. The 2D U-Net lateral tibial thickness measurement demonstrated a very small underestimation while those from 3D U-Net showed a slightly larger overestimation (Table 3, Figure 4). Both 2D and 3D U-Net demonstrated a similarly small systematic overestimation of medial tibial cartilage thickness. The mean biases and 95% limits of agreement for femoral and tibial $T_2$ measurements showed similar trends between 2D and 3D U-Net but were generally lower for 3D U-Net (Table 3, Figure 7).

Vertex-wise root-mean-squared (RMS) thickness measurements extracted from all cartilage surfaces generated from 2D U-Net segmentations of the OAI ZIB testing set demonstrated moderate-to-high linear correlations (range: R = 0.74 - 0.89, Figure 3), while those extracted from 3D U-Net segmentations demonstrated high correlations (range: R = 0.83 - 0.98, Figure 3).

Vertex-wise RMS $T_2$ measurements extracted from all cartilage surfaces generated from 2D U-Net (range: R = 0.94 - 0.97, Figure 6) and 3D U-Net (range: R = 0.97 - 0.99, Figure 6) segmentations demonstrated high linear correlations with the measurements obtained from manual segmentations.

Surface-averaged RMS thickness and $T_2$ measurements of all cartilage surfaces extracted from manual and automatic segmentations of the OAI ZIB testing set using 3D-CaSM are listed in Table 3. Vertex-wise RMS thickness measurements and $T_2$ relaxation times extracted from manual and automated segmentations displayed on

the canonical femoral, medial tibial and lateral tibial cartilage surfaces are shown in Figures 5 and 8, respectively. Difference (Δ) maps between data extracted from manual and 2D and 3D U-Net automated segmentations are also shown.

## Discussion

This study has shown that CNN-automated segmentations of femoral and tibial bone-cartilage structures can be used within the 3D cartilage surface mapping (3D-CaSM) pipeline for a fast, accurate and quantitative surface-based analysis of cartilage morphology and relaxometry.

Compared to an expert human reader requiring ~3-4 hours to manually delineate cartilage for a single knee joint, both the 2D and 3D U-Net showed substantially increased time-efficiency by segmenting a complete sagittal 3D DESS volume from the OAI dataset in under 5 s [8]. While both networks showed high agreement with manual segmentation (DSC > 0.966), the best performing 3D U-Net (DSC ≥ 0.980, ASSD ≤ 0.314 mm) showed a slightly higher accuracy than the best performing 2D network (DSC ≥ 0.971, ASSD ≤ 0.543 mm). Therefore, although the training duration for 3D U-Net was substantially larger than that of the 2D U-Nets due to training on a ten-fold larger volume, applying this fully trained 3D U-Net in practice achieves high performance with only a minor increase in segmentation time. The International Workshop on Osteoarthritis Imaging (IWOAI) Knee MRI Segmentation Challenge showed promising results for cartilage segmentations with all participating teams using 2D or 3D CNNs and achieving DSCs of approximately 0.81 – 0.90 [38]. The 3D U-Net used here achieved similar DSCs for both the femoral and tibial bone-cartilage structures compared with those presented by Ambellan et al for femoral and tibial bone-only segmentations of the OAI DESS images [26]. However, lower ASSD scores were achieved in this study compared to Ambellan et al, which could stem from the more challenging task of segmenting both bone and cartilage in one volume and both tissues having very different contrasts on the DESS images. Due to the difference in segmentation strategy used in this study compared to previous studies (bone-cartilage structures vs bone-only and cartilage-only) a simple, direct comparison is challenging. Nevertheless, the benefit of using 3D CNNs over 2D CNNs to achieve greater segmentation accuracy became evident, primarily through the inclusion of adjacent slices in the training volume.

We performed 3D-CaSM to determine cartilage regions where the networks might experience segmentation difficulties compared to manual. When averaging the thickness measurements across the whole femoral, lateral tibial and medial tibial cartilage surfaces, the average femoral cartilage thickness determined from 2D U-Net segmentations was shown to be substantially larger than the average thickness determined from expert-defined manual segmentations (difference = +0.33 mm). By using the 3D-CaSM technique, this average over-estimation of cartilage thickness by the 2D U-Net was regionally localised to the lateral and medial femoral condyle as well as the patellofemoral groove and where these regions transition into the intercondylar notch. These areas are particularly difficult to segment with different tissues of similar contrast on the DESS images having direct contact with each other such as femorotibial cartilage-cartilage contact and cartilage-meniscus contact. Previous studies have also shown similar cartilage thickness over-estimation in these regions when using 2D U-Net generated segmentations [19,25]. However, this regional over-estimation of femoral cartilage thickness was not observed with the 3D U-Net. This could be explained by the 3D network being more volumetrically consistent by training on a larger sub-volume of adjacent image slices instead of individual slices as with 2D networks.

When averaging the $T_2$ measurements over the entire femoral, lateral tibial and medial tibial cartilage surfaces, differences between manual and 3D U-Net T2 measurements were slightly less than those between manual and 2D U-Net. However, when looking at vertex-wise $T_2$ difference maps between data extracted from manual and automated segmentations, the $T_2$ 2D U-Net difference maps show higher focal disagreements ($T_2$ under-estimations at the medial and lateral femoral condyles; $T_2$ over-estimation at the patellofemoral groove) than the $T_2$ difference maps with 3D U-Net. While the $T_2$ 3D U-Net difference maps of femoral, medial tibial and lateral tibial cartilage show similar patterns to the corresponding $T_2$ 2D U-Net difference maps, the overall vertex-wise $T_2$ differences with manual segmentations are substantially less pronounced.

While segmentation networks are usually trained on a large amount of imaging data to achieve a sense of generalisability, the networks remain sensitive to small distributional shifts between training and testing data and could produce inaccurate segmentations. Although automated methods lead to more repeatable segmentation results, the proposed method could also be used interactively. Following the

automated generation of the structure polygons, they could be manually adjusted by the user to fine-tune and correct the network-generated results. This could not only increase segmentation accuracy, but also substantially reduce the time required to fully delineate structures manually. This could allow the models to be employed by clinical end users and enable the analysis of large medical imaging datasets in an acceptable time.

Although we automated the segmentation of the bone-cartilage structures from which triangulated surface mesh objects were generated, the 3D-CaSM technique used in this work still required manual outlining of cartilage surface patches. Although this step only required approximately 2 minutes to extract all cartilage patches per subject, future work could focus on fully automating the 3D-CaSM pipeline. Another limitation is the lack of longitudinal validation of 3D-CaSM using contours generated from both manual and network-automated segmentations. This study only assessed the segmentation performance of 2D and 3D U-Net on the OAI baseline dataset as the OAI ZIB dataset only consists of segmentations from baseline OAI scans. Generating manual segmentations from multiple OAI follow-up imaging data might mitigate this problem in the future. By registering the OAI $T_2$-weighted MSME images to the OAI DESS images, we may be including systematic variations in the quantified $T_2$ relaxation time measurements [39,40]. However, in the present study, we evaluated the similarities between $T_2$ relaxation time measurements extracted using manual and automated segmentations within the 3D-CaSM pipeline rather than the accuracy of the extracted $T_2$ relaxation times themselves. Further, the two-channel input containing duplicate slices in this work could be redundant for model training. Future investigations to reduce dimensionality within our models are warranted. However, the current dimensionality has shown in this and prior work to provide accurate results which provides a good baseline for future optimisations [21]. Lastly, the networks were trained, validated and tested on imaging data from the highly controlled OAI observational study. Although the OAI is a multi-centre study, images were acquired on MR systems from a single manufacturer. Additionally, the chosen sample size for testing was relatively small, mainly attributed to limited availability of resources. Future work would aim to expand towards larger and more clinically representational datasets and to evaluate the clinical significance of imaging biomarkers.

In the current study we evaluate the use of automated femoral and tibial bone-cartilage segmentations in combination with 3D-CaSM for fast and accurate

quantification of cartilage thickness and $T_2$ relaxation time measurements in comparison with manual segmentation. This method allowed us to highlight localized cartilage regions in which both 2D and 3D U-Nets experienced segmentation difficulties. The presented results demonstrate the validity of using automated segmentations within the 3D-CaSM pipeline to improve the time-efficiency for the MRI-based evaluation of knee joint health.

## Acknowledgements

The authors acknowledge research support from GlaxoSmithKline, the National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014) and the Addenbrooke's Charitable Trust (ACT). JDK acknowledges support from Horizon EU 2020 (grant agreement no. 761214). The views expressed in this article are those of the author(s) and not necessarily those of GlaxoSmithKline, the NIHR, the ACT or the Department of Health and Social Care.

This manuscript was prepared using public use data sets from the Osteoarthritis Initiative (OAI) and Zuse Institute Berlin (ZIB) and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, the private funding partners or the ZIB.

## CRediT authorship contribution statement

**Dimitri A. Kessler:** Conceptualization, Data curation, Software, Analysis, Data interpretation, Writing – original draft, Writing – review & editing. **James W. MacKay:** Conceptualization, Software, Data interpretation, Writing – review & editing. **Stephen M. McDonnell:** Data interpretation, Writing – review & editing. **Robert L. Janiczek:** Data interpretation, Writing – review & editing. **Martin J. Graves:** Data interpretation, Writing – review & editing. **Joshua D. Kaggie:** Conceptualization, Software, Data interpretation, Writing – review & editing. **Fiona J. Gilbert:** Conceptualization, Data interpretation, Writing – review & editing.

## Conflict of Interest

Dimitri A. Kessler acknowledges funding support from GlaxoSmithKline for his studentship. Dimitri A. Kessler is an employee of Image Analysis Group. James W. MacKay is an employee of AstraZeneca. Stephen M. McDonnell, Robert L. Janiczek, Martin J. Graves, Joshua D. Kaggie and Fiona J. Gilbert have no competing interests to declare.

## References

[1] D. J. Hunter and S. Bierma-Zeinstra, Osteoarthritis, Lancet 393 (2019) 1745–1759. https://doi.org/10.1016/S0140-6736(19)30417-9.

[2] D. J. Hunter, D. Schofield, and E. Callander, The individual and socioeconomic impact of osteoarthritis, Nat. Rev. Rheumatol. 10 (2014) 437–441. https://doi.org/10.1038/nrrheum.2014.44.

[3] M. T. Nieminen, J. Rieppo, J. Töyräs, *et al.*, T 2 relaxation reveals spatial collagen architecture in articular cartilage: A comparative quantitative MRI and polarized light microscopic study, Magn. Reson. Med. 46 (2001) 487–493. https://doi.org/10.1002/mrm.1218.

[4] S. V. S. Akella, R. R. Regatte, A. J. Gougoutas, *et al.*, Proteoglycan-Induced Changes in T1ρ-Relaxation of Articular Cartilage at 4T, Magn. Reson. Med. 46 (2001) 419–423. https://doi.org/10.1002/mrm.1208.

[5] U. Duvvuri, R. Reddy, S. D. Patel, J. H. Kaufman, J. B. Kneeland, and J. S. Leigh, T1ρ-relaxation in articular cartilage: Effects of enzymatic degradation, Magn. Reson. Med. 38 (1997) 863–867. https://doi.org/10.1002/mrm.1910380602.

[6] M. J. Nissi, J. Töyräs, M. S. Laasanen, *et al.*, Proteoglycan and collagen sensitive MRI evaluation of normal and degenerated articular cartilage, J. Orthop. Res. 22 (2004) 557–564. https://doi.org/10.1016/j.orthres.2003.09.008.

[7] R. R. Regatte, S. V. S. Akella, J. H. Lonner, J. B. Kneeland, and R. Reddy, T1ρ relaxation mapping in human osteoarthritis (OA) cartilage: Comparison of T1ρ with T2, J. Magn. Reson. Imaging 23 (2006) 547–553. https://doi.org/10.1002/jmri.20536.

[8] J. W. MacKay, J. D. Kaggie, G. M. Treece, *et al.*, Three-Dimensional Surface-

Based Analysis of Cartilage MRI Data in Knee Osteoarthritis: Validation and Initial Clinical Application, J. Magn. Reson. Imaging 52 (2020) 1139–1151. https://doi.org/10.1002/jmri.27193.

[9] F. Eckstein, W. Wirth, A. Guermazi, S. Maschek, and A. Aydemir, Brief Report: Intraarticular Sprifermin Not Only Increases Cartilage Thickness, but Also Reduces Cartilage Loss: Location-Independent Post Hoc Analysis Using Magnetic Resonance Imaging, Arthritis Rheumatol. 67 (2015) 2916–2922. https://doi.org/10.1002/art.39265.

[10] V. Pedoia, S. Majumdar, and T. M. Link, Segmentation of joint and musculoskeletal tissue in the study of arthritis, Magn. Reson. Mater. Physics, Biol. Med. 29 (2016) 207–221. https://doi.org/10.1007/s10334-016-0532-9.

[11] J. W. MacKay, S. B. L. Low, T. O. Smith, A. P. Toms, A. W. McCaskie, and F. J. Gilbert, Systematic review and meta-analysis of the reliability and discriminative validity of cartilage compositional MRI in knee osteoarthritis, Osteoarthr. Cartil. 26 (2018) 1140–1152. https://doi.org/10.1016/j.joca.2017.11.018.

[12] T. G. Williams, A. P. Holmes, J. C. Waterton, *et al.*, Anatomically corresponded regional analysis of cartilage in asymptomatic and osteoarthritic knees by statistical shape modelling of the bone, IEEE Trans. Med. Imaging 29 (2010) 1541–1559. https://doi.org/10.1109/TMI.2010.2047653.

[13] U. D. Monu, C. D. Jordan, B. L. Samuelson, B. A. Hargreaves, G. E. Gold, and E. J. McWalter, Cluster analysis of quantitative MRI T 2 and T 1ρ relaxation times of cartilage identifies differences between healthy and ACL-injured individuals at 3T, Osteoarthr. Cartil. 25 (2017) 513–520. https://doi.org/10.1016/j.joca.2016.09.015.

[14] A. Norouzi, M. S. M. Rahim, A. Altameem, *et al.*, Medical Image Segmentation Methods, Algorithms, and Applications, IETE Tech. Rev. 31 (2014) 199–213. https://doi.org/10.1080/02564602.2014.906861.

[15] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.*, Lecture No., N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Springer, 2015 234–241., 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, 3D U-

Net: Learning Dense Volumetric Segmentation from Sparse Annotation, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9901 LNCS, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016 424–432., 424–432. https://doi.org/10.1007/978-3-319-46723-8_49.

[17] F. Milletari, N. Navab, and S. A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, Proc. - 4th Int. Conf. 3D Vision, 3DV 2016 (2016) 565–571. https://doi.org/10.1109/3DV.2016.79.

[18] F. Liu, Z. Zhou, H. Jang, A. Samsonov, G. Zhao, and R. Kijowski, Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging, Magn. Reson. Med. 79 (2017) 2379–2391. https://doi.org/10.1002/mrm.26841.

[19] B. Norman, V. Pedoia, and S. Majumdar, 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 288 (2018) 177–185. https://doi.org/10.1148/radiol.2018172322.

[20] S. Gaj, M. Yang, K. Nakamura, and X. Li, Automated cartilage and meniscus segmentation of knee MRI with conditional generative adversarial networks, Magn. Reson. Med. (2019) 1–13. https://doi.org/10.1002/mrm.28111.

[21] D. A. Kessler, J. W. MacKay, V. A. Crowe, *et al.*, The optimisation of deep neural networks for segmenting multiple knee joint tissues from MRIs, Comput. Med. Imaging Graph. 86 (2020) https://doi.org/10.1016/j.compmedimag.2020.101793.

[22] Z. Zhou, G. Zhao, R. Kijowski, and F. Liu, Deep convolutional neural network for segmentation of knee joint anatomy, Magn. Reson. Med. 80 (2018) 2759–2770. https://doi.org/10.1002/mrm.27229.

[23] A. Paproki, K. J. Wilson, R. K. Surowiec, *et al.*, Automated segmentation and T2-mapping of the posterior cruciate ligament from MRI of the knee: Data from the osteoarthritis initiative, Proc. - 2016 IEEE 13th Int. Symp. Biomed. Imaging (2016) 424–427. https://doi.org/10.1109/ISBI.2016.7493298.

[24] V. Badrinarayanan, A. Kendall, and R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495.

https://doi.org/10.1109/TPAMI.2016.2644615.

[25] W. Wirth, F. Eckstein, J. Kemnitz, *et al.*, Accuracy and longitudinal reproducibility of quantitative femorotibial cartilage measures derived from automated U-Net-based segmentation of two different MRI contrasts: data from the osteoarthritis initiative healthy reference cohort, Magn. Reson. Mater. Physics, Biol. Med. (2020) https://doi.org/10.1007/s10334-020-00889-7.

[26] F. Ambellan, A. Tack, M. Ehlke, and S. Zachow, Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative, Med. Image Anal. 52 (2019) 109–118. https://doi.org/10.1016/j.media.2018.11.009.

[27] The Osteoarthritis Initiative, *https://nda.nih.gov/oai/*.

[28] H. Seim, D. Kainmueller, H. Lamecker, M. Bindernagel, J. Malinowski, and S. Zachow, Model-based auto-segmentation of knee bones and cartilage in MRI data, Med. Image Anal. Clin. A Gd. Chall. (2010) 215–223. [Online]. Available: http://www.diagnijmegen.nl/~bram/grandchallenge2010/215.pdf

[29] R. B. Souza, D. Kumar, N. Calixto, *et al.*, Response of knee cartilage T1rho and T2 relaxation times to in vivo mechanical loading in individuals with and without knee osteoarthritis, Osteoarthr. Cartil. 22 (2014) 1367–1376. https://doi.org/10.1016/j.joca.2014.04.017.

[30] S. Van Rossom, C. R. Smith, L. Zevenbergen, *et al.*, Knee cartilage thickness, T1ρ and T2 relaxation time are related to articular cartilage loading in healthy adults, PLoS One 12 (2017) 1–16. https://doi.org/10.1371/journal.pone.0170002.

[31] D. A. Kessler, J. W. MacKay, S. McDonald, *et al.*, Effectively Measuring Exercise-Related Variations in T1ρ and T2 Relaxation Times of Healthy Articular Cartilage, J. Magn. Reson. Imaging 52 (2020) 1753–1764. https://doi.org/10.1002/jmri.27278.

[32] M. Larobina and L. Murino, Medical image file formats, J. Digit. Imaging 27 (2014) 200–206. https://doi.org/10.1007/s10278-013-9657-9.

[33] M. McCormick, X. Liu, J. Jomier, C. Marion, and L. Ibanez, ITK: enabling reproducible research and open science, Front. Neuroinform. 8 (2014) 1–11. https://doi.org/10.3389/fninf.2014.00013.

[34] D. P. Kingma and J. L. Ba, Adam: A Method for Stochastic Optimization, 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, USA (2015) 1–15..

[35] S. A. Taghanaki, Y. Zheng, S. Kevin Zhou, *et al.*, Combo loss: Handling input and output imbalance in multi-organ segmentation, Comput. Med. Imaging Graph. 75 (2019) 24–33. https://doi.org/10.1016/j.compmedimag.2019.04.005.

[36] L. R. Dice, Measures of the Amount of Ecologic Association Between Species., Ecology 26 (1945) 297–302. https://doi.org/10.2307/1932409.

[37] T. J. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1–34..

[38] A. D. Desai, F. Caliva, C. Iriondo, *et al.*, The International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge: A Multi-Institute Evaluation and Analysis Framework on a Standardized Dataset, Radiol. Artif. Intell. 3 (2021) e200078. https://doi.org/10.1148/ryai.2021200078.

[39] K. L. Urish, A. A. Williams, J. R. Durkin, and C. R. Chu, Registration of Magnetic Resonance Image Series for Knee Articular Cartilage Analysis, Cartilage 4 (2013) 20–27. https://doi.org/10.1177/1947603512451745.

[40] D. Fürst, W. Wirth, A. Chaudhari, and F. Eckstein, Layer-specific analysis of femorotibial cartilage t2 relaxation time based on registration of segmented double echo steady state (dess) to multi-echo-spin-echo (mese) images, Magn. Reson. Mater. Physics, Biol. Med. 33 (2020) 819–828. https://doi.org/10.1007/s10334-020-00852-6.



## Tables

**Table 1** - Summary of testing dataset used to evaluate networks and to perform quantitative cartilage surface analysis.

| | | |
|---|---|---|
| Testing | Number of Subjects | 20 |
| | Sex (Female / Male) | 12 / 8 |
| | Age (years) | 66.8 ± 9.0 |
| | Subjects per KL Grade (0 / 1 / 2 / 3 / 4) | 2 / 4 / 6 / 6 / 2 |

**Table 2 -** Comparison of segmentation performance of the proposed 2D and 3D U-Net when trained and tested on the OAI ZIB dataset using different loss objective functions. Data are means ± standard deviations. The asterisk (*) indicates that all differences in segmentation scores between the 2D U-Nets and the corresponding 3D U-Nets of the same loss function were statistically significant ($p < 0.05$). Abbreviations: ASSD, average symmetric surfaces distance; BCE, binary cross entropy; β, weighting hyperparameter within Combo loss; DSC, Sørensen–Dice similarity coefficient; VOE, volumetric overlap error.

| Network | Loss | β | Femur | | | Tibia | | |
|---|---|---|---|---|---|---|---|---|
| | | | DSC | VOE (%) | ASSD (mm) | DSC | VOE (%) | ASSD (mm) |
| 2D U-Net | BCE | | 0.966 ± 0.008 | 6.598 ± 1.568 | 0.701 ± 0.232 | 0.974 ± 0.004 | 5.143 ± 0.687 | 0.449 ± 0.093 |
| | Dice | | 0.968 ± 0.007 | 6.118 ± 1.295 | 1.030 ± 0.717 | 0.968 ± 0.005 | 6.166 ± 0.906 | 0.548 ± 0.109 |
| | Combo | 0.25 | **0.971 ± 0.006** | **5.662 ± 1.150** | **0.543 ± 0.164** | 0.974 ± 0.004 | 5.129 ± 0.685 | 0.437 ± 0.068 |
| | | 0.50 | 0.966 ± 0.008 | 6.522 ± 1.446 | 0.779 ± 0.335 | 0.972 ± 0.004 | 5.512 ± 0.706 | 0.459 ± 0.060 |
| | | 0.75 | 0.969 ± 0.007 | 6.070 ± 1.258 | 0.581 ± 0.168 | **0.974 ± 0.003** | **5.044 ± 0.646** | **0.412 ± 0.055** |
| 3D U-Net* | BCE | | 0.978 ± 0.004 | 4.379 ± 0.840 | 0.353 ± 0.058 | 0.980 ± 0.003 | 3.895 ± 0.506 | 0.305 ± 0.039 |
| | Dice | | 0.979 ± 0.004 | 4.069 ± 0.723 | 0.327 ± 0.052 | **0.982 ± 0.002** | **3.621 ± 0.465** | **0.282 ± 0.034** |
| | Combo | 0.25 | **0.980 ± 0.004** | **3.902 ± 0.784** | **0.314 ± 0.054** | 0.980 ± 0.003 | 3.893 ± 0.545 | 0.304 ± 0.040 |
| | | 0.50 | 0.980 ± 0.004 | 3.979 ± 0.685 | 0.320 ± 0.045 | 0.981 ± 0.003 | 3.788 ± 0.512 | 0.296 ± 0.041 |
| | | 0.75 | 0.978 ± 0.004 | 4.277 ± 0.702 | 0.345 ± 0.049 | 0.978 ± 0.003 | 4.221 ± 0.511 | 0.331 ± 0.038 |

**Table 3 –** Results of Pearson correlation (R) and Bland-Altman analysis. Vertex-wise thickness and T2 measurements from manual segmentations were compared to the respective vertex-wise measurements from automated segmentations of both networks. *Mean bias and 95% limits of agreement (LoA) are in mm. ** Mean bias and 95% LoA are in ms

| Parameter | Network | Cartilage Surface | R-Value | Mean Bias [95% LoA] |
|---|---|---|---|---|
| Thickness* | 2D U-Net | Femoral | 0.75 | 0.33 [-0.28; 0.96] |
| | | Lateral Tibial | 0.89 | -0.03 [-0.21; 0.15] |
| | | Medial Tibial | 0.74 | 0.12 [-0.16; 0.40] |
| | 3D U-Net | Femoral | 0.98 | 0.07 [-0.11; 0.25] |
| | | Lateral Tibial | 0.83 | 0.11 [-0.14; 0.36] |
| | | Medial Tibial | 0.85 | 0.14 [-0.10; 0.39] |
| $T_2$** | 2D U-Net | Femoral | 0.94 | -0.16 [-4.71; 4.40] |
| | | Lateral Tibial | 0.97 | 0.47 [-2.68; 3.62] |
| | | Medial Tibial | 0.96 | 1.32 [-2.19; 4.83] |
| | 3D U-Net | Femoral | 0.99 | -0.05 [-2.06; 1.95] |
| | | Lateral Tibial | 0.98 | 0.38 [-2.12; 2.87] |
| | | Medial Tibial | 0.97 | 0.46 [-2.47; 3.39] |

**Table 4 -** Results of thickness and $T_2$ analysis of the baseline OAI ZIB testing dataset. Cartilage surface means (standard deviations) are calculated from all vertex-wise RMS thickness and $T_2$ measurements. *Measurements are in mm. **Measurements are in ms.

| Parameter | Cartilage Surface | Manual Surface Mean (SD) | Automated Surface Mean (SD) | | Absolute Difference | |
|---|---|---|---|---|---|---|
| | | | 2D U-Net | 3D U-Net | 2D U-Net | 3D U-Net |
| Thickness* | Femur | 1.97 (0.47) | 2.30 (0.41) | 2.04 (0.46) | 0.33 | 0.07 |
| | Lateral Tibia | 1.71 (0.19) | 1.68 (0.20) | 1.82 (0.23) | 0.03 | 0.11 |
| | Medial Tibia | 1.43 (0.13) | 1.54 (0.21) | 1.57 (0.21) | 0.11 | 0.14 |
| $T_2$** | Femur | 59.07 (5.95) | 58.91 (6.70) | 59.02 (6.30) | 0.16 | 0.05 |
| | Lateral Tibia | 51.10 (6.55) | 51.57 (6.63) | 51.48 (6.25) | 0.47 | 0.38 |
| | Medial Tibia | 52.63 (6.25) | 53.95 (5.44) | 53.09 (5.71) | 1.33 | 0.46 |



## Figure Legends

**Figure 1** - **A)** The 2D U-Net architecture. The encoding part of the network consists of the repeated application of eight 4x4 convolutions with stride 2 and padding 1, downsampling the input by a factor of 2 at each layer. With exception of the first layer, each convolution is followed by a batch normalisation layer (BatchNorm) and a leaky rectified linear unit (leaky ReLU) with slope 0.2. After the first encoding step, the number of feature channels is doubled

during the next three convolutions (64 – 512), while the ensuring four are kept at 512. In the subsequent decoding part, the input is repeatedly up sampled by a factor of 2 by eight 4x4 convolutional layers with stride 2 and padding 1, and additional skip connections (concatenations) between each major layer, changing the number of feature channels at each step. The first transpose convolution is followed by BatchNorm and ReLU without dropout while the succeeding three transpose convolutions are followed by BatchNorm, dropout (50%) and a ReLU. The next three decoder up-convolutions are followed by BatchNorm and a ReLU without dropout. The final up-convolution is only followed by a Sigmoid activation layer to generate the label map. Total number of parameters: 66.999 M. **B)** The 3D U-Net architecture is very similar to that of the 2D U-Net in which all 2D operations are replaced by their 3D counterparts, i.e. 4x4x4 convolutions with 1x2x2 stride and 1x1x1 padding as well as 3D BatchNorm. Total number of parameters: 217.614 M

**Figure 2** – 3D Cartilage Surface Mapping (3D-CaSM) pipeline used for quantitative measurements of cartilage thickness and $T_2$ relaxation time determined from manual and network-generated segmentations. Femoral cartilage measurements used for demonstration purposes; same process used for tibial cartilages. Following the conversion of the manual and automated segmentation maps into tissue-specific polygon contours, 3D triangulated surface mesh objects were generated from the bone-cartilage structures (step 1). Through data compounding (displaying mean image intensities within 10 pixels from the 3D object onto its surface), cartilage patches were identified and manually extracted (step 2). Inner and outer cartilage surfaces were determined by calculating the cartilage thickness across the patch (step 3). These vertex-wise thickness measurements can then be displayed onto the cartilage surface patch (step 4). The same inner and outer cartilage surfaces can then be used to sample the registered $T_2$ maps and obtain vertex-wise measurements of cartilage $T_2$ relaxation times (steps 5 and 6). After registering individual patches to a canonical surface, we can visually compare the measurements obtained from manual and network-generated contours and calculate vertex-wise thickness and $T_2$ differences (step not shown). The focus of this study is highlighted in blue in which we compare the 3D-CaSM pipeline using contours from manual segmentations with the 3D-CaSM pipeline using contours determined from automated segmentations. The successive steps are kept identical for an effective comparison.

**Figure 3** - Scatterplots show comparison of vertex-wise thickness measurements determined from cartilage surfaces extracted from manual and automatic segmentations (**A**: 2D U-Net, **B**: 3D U-Net).

**Figure 4** - Bland-Altman plots show comparison of vertex-wise thickness measurements determined from cartilage surfaces extracted from manual and automatic segmentations (**A**: 2D U-Net, **B**: 3D U-Net).

**Figure 5** - Vertex-wise RMS thickness data extracted from manual and automated segmentations from the OAI ZIB test set displayed on the canonical femoral, medial tibial and lateral tibial cartilage surfaces. Difference maps between data extracted from manual and 2D U-Net-automated (2D U-Net Δ) / 3D U-Net-automated (3D U-Net Δ) segmentations highlight spatial regions where the networks experience segmentation difficulties.

**Figure 6** - Scatterplots show comparison of vertex-wise $T_2$ measurements determined from cartilage surfaces extracted from manual and automatic segmentations (**A**: 2D U-Net, **B**: 3D U-Net).

**Figure 7** - Bland-Altman plots show comparison of vertex-wise $T_2$ measurements determined from cartilage surfaces extracted from manual and automatic segmentations (**A**: 2D U-Net, **B**: 3D U-Net).

**Figure 8** - Vertex-wise RMS $T_2$ data extracted from manual and automated segmentations from the OAI ZIB test set displayed on the canonical femoral, medial tibial and lateral tibial cartilage surfaces. Difference maps between data extracted from manual and 2D U-Net-automated (2D U-Net Δ) / 3D U-Net-automated (3D U-Net Δ) segmentations highlight focal regions of $T_2$ discrepancies.

## Figures

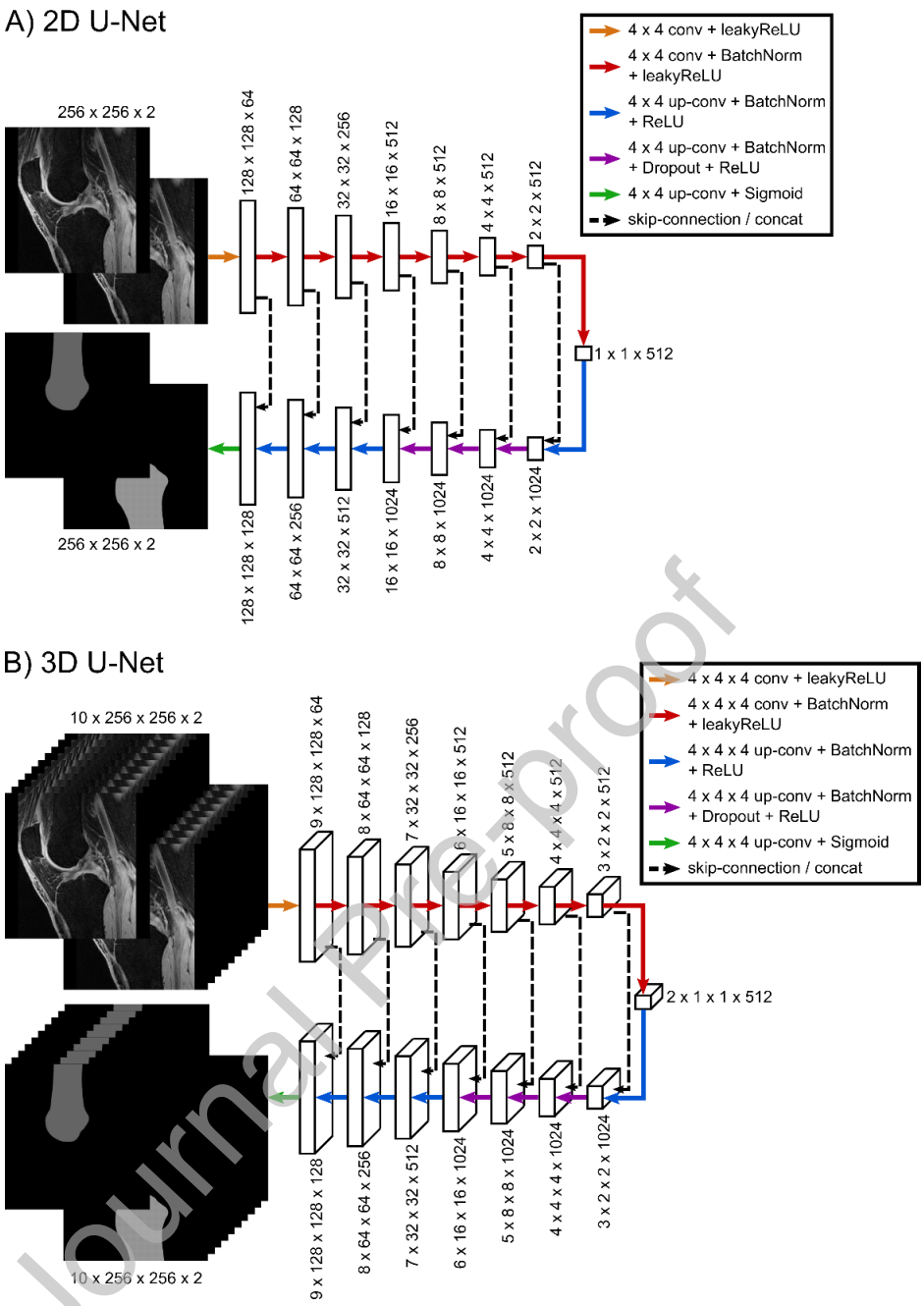



**Figure 1 - A)** The 2D U-Net architecture. The encoding part of the network consists of the repeated application of eight 4x4 convolutions with stride 2 and padding 1, downsampling the input by a factor of 2 at each layer. With the exception of the first layer, each convolution is followed by a batch normalisation layer (BatchNorm) and a leaky rectified linear unit (leaky ReLU) with slope 0.2. After the first encoding step, the number of feature channels is doubled during the next three convolutions (64 – 512), while the ensuring four are kept at 512. In the subsequent decoding part, the input is repeatedly up sampled by a factor of 2 by eight 4x4 convolutional layers with stride 2 and padding 1, and additional skip connections (concatenations) between each major layer, changing the number of feature channels at each step. The first transpose convolution is followed by BatchNorm and ReLU without dropout while the succeeding three transpose convolutions are followed by BatchNorm, dropout (50%) and a ReLU. The next three decoder up-convolutions are followed by BatchNorm and a ReLU without dropout. The final up-convolution is only followed by a Sigmoid activation layer to generate the label map. Total number of parameters: 66.999 M. **B)** The 3D U-Net architecture is very similar to that of the 2D U-Net in which all 2D operations are replaced by their 3D counterparts, i.e. 4x4x4 convolutions with 1x2x2 stride and 1x1x1 padding as well as 3D BatchNorm. Total number of parameters: 217.614 M

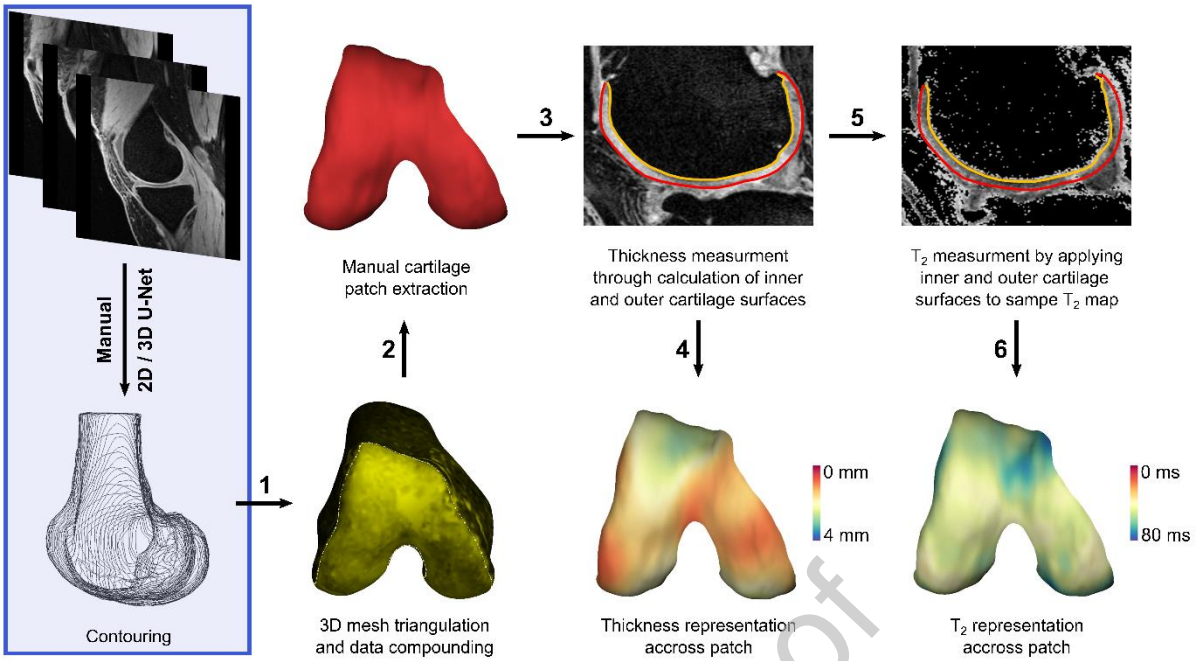



**Figure 2 -** 3D Cartilage Surface Mapping (3D-CaSM) pipeline used for quantitative measurements of cartilage thickness and $T_2$ relaxation time determined from manual and network-generated segmentations. Femoral cartilage measurements used for demonstration purposes; same process used for tibial cartilages. Following the conversion of the manual and automated segmentation maps into tissue-specific polygon contours, 3D triangulated surface mesh objects were generated from the bone-cartilage structures (step 1). Through data compounding (displaying mean image intensities within 10 pixels from the 3D object onto its surface), cartilage patches were identified and manually extracted (step 2). Inner and outer cartilage surfaces were determined by calculating the cartilage thickness across the patch (step 3). These vertex-wise thickness measurements can then be displayed onto the cartilage surface patch (step 4). The same inner and outer cartilage surfaces can then be used to sample the registered $T_2$ maps and obtain vertex-wise measurements of cartilage $T_2$ relaxation times (steps 5 and 6). After registering individual patches to a canonical surface, we can visually compare the measurements obtained from manual and network-generated contours and calculate vertex-wise thickness and $T_2$ differences (step not shown). The focus area of this study is highlighted in blue when we compare the 3D-CaSM pipeline using contours from manual segmentations and the 3D-CaSM pipeline using contours from automated segmentations. The successive steps are kept identical for an effective comparison.

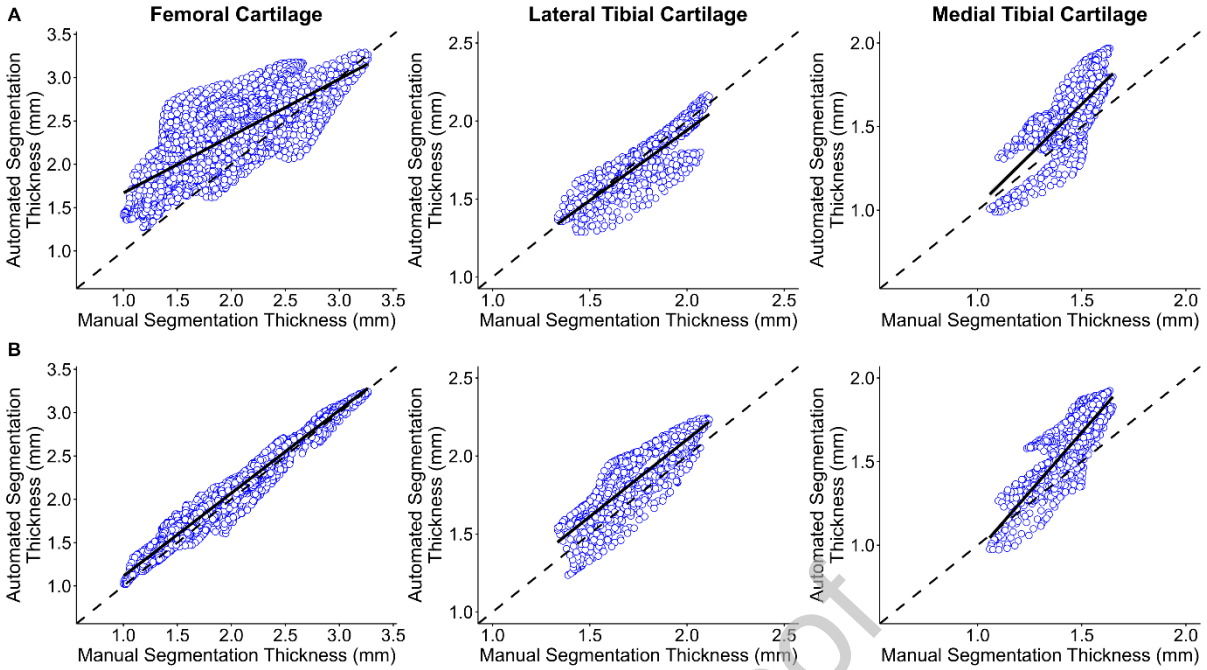



**Figure 3 -** Scatterplots show comparison of vertex-wise thickness measurements determined from cartilage surfaces extracted from manual and automatic segmentations (**A**: 2D U-Net, **B**: 3D U-Net).

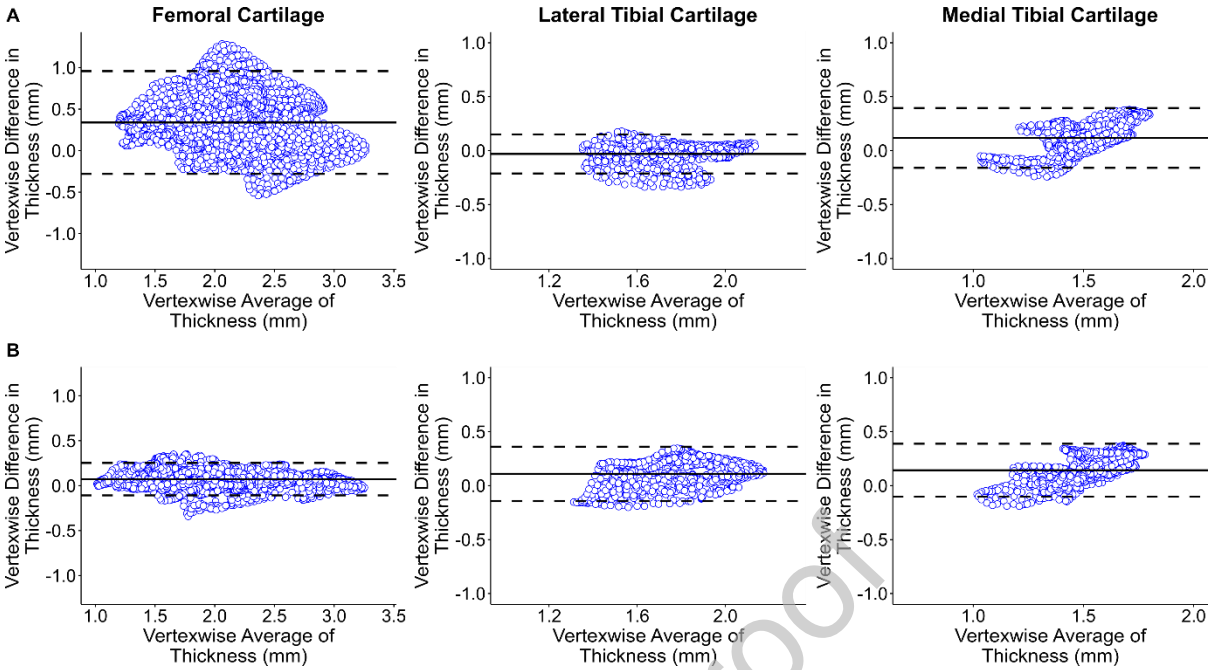



**Figure 4 -** Bland-Altman plots show comparison of vertex-wise thickness measurements determined from cartilage surfaces extracted from manual and automatic segmentations (**A**: 2D U-Net, **B**: 3D U-Net).

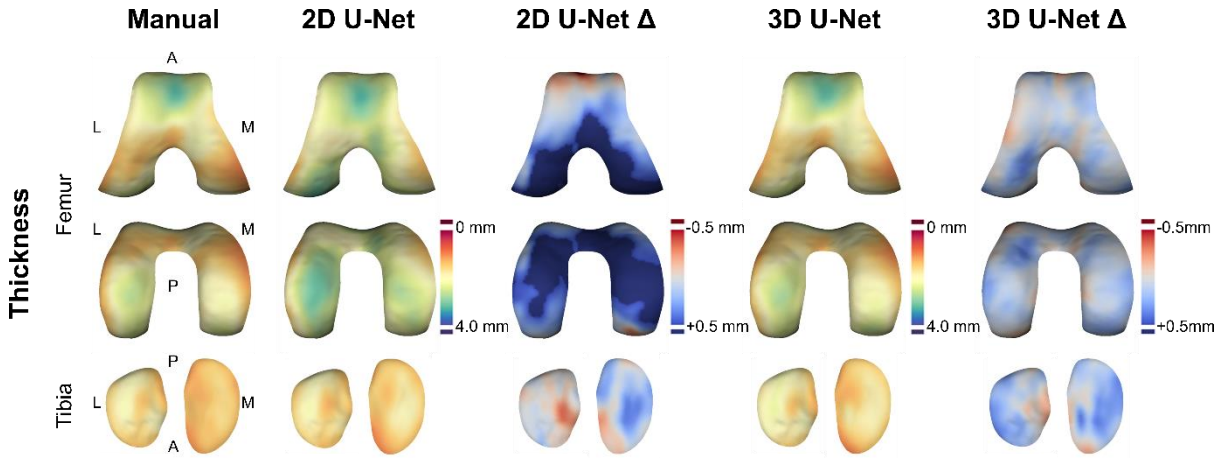



**Figure 5 –** Vertex-wise RMS thickness data extracted from manual and automated segmentations from the OAI ZIB test set displayed on the canonical femoral, medial tibial and lateral tibial cartilage surfaces. Difference maps between data extracted from manual and 2D U-Net-automated (2D U-Net Δ) / 3D U-Net-automated (3D U-Net Δ) segmentations highlight spatial regions where the networks experience segmentation difficulties.

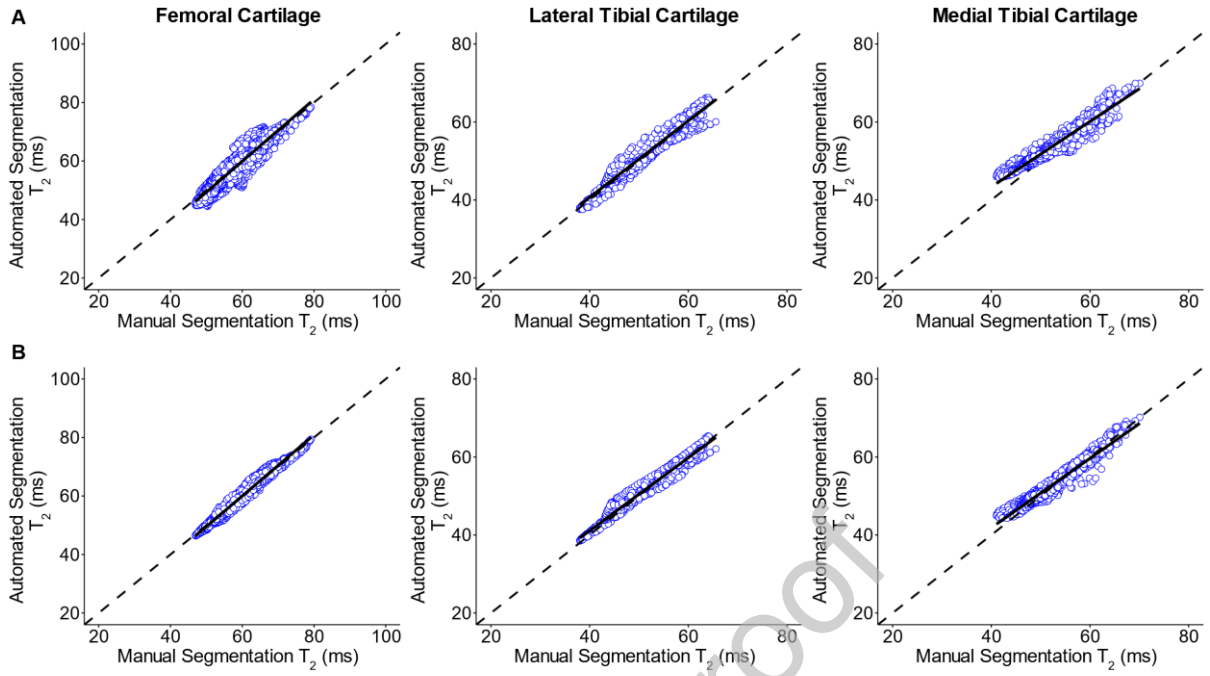



**Figure 6 -** Scatterplots show comparison of vertex-wise $T_2$ measurements determined from cartilage surfaces extracted from manual and automatic segmentations (**A**: 2D U-Net, **B**: 3D U-Net).

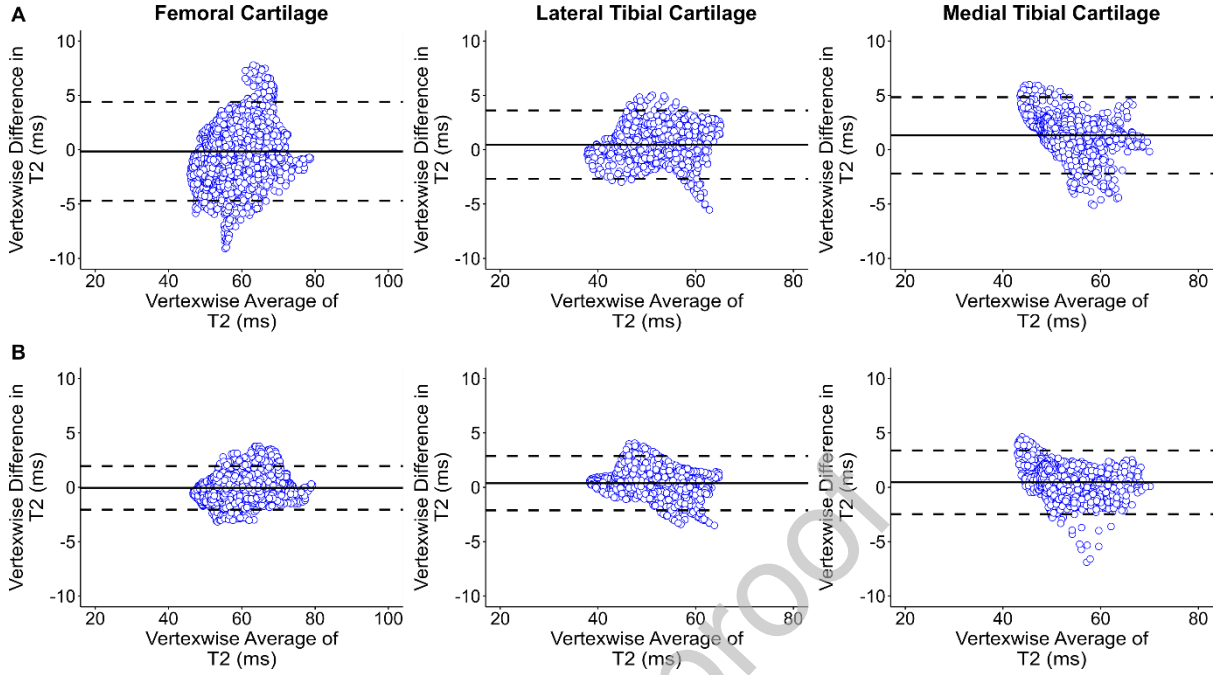



**Figure 7 -** Bland-Altman plots show comparison of vertex-wise $T_2$ measurements determined from cartilage surfaces extracted from manual and automatic segmentations (**A**: 2D U-Net, **B**: 3D U-Net).

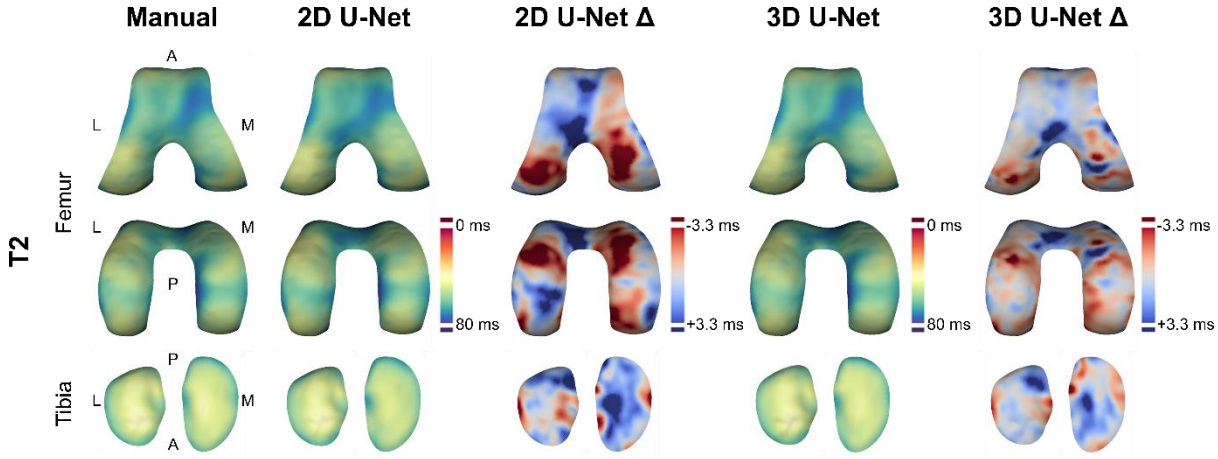



**Figure 8 –** Vertex-wise RMS $T_2$ data extracted from manual and automated segmentations from the OAI ZIB test set displayed on the canonical femoral, medial tibial and lateral tibial cartilage surfaces. Difference maps between data extracted from manual and 2D U-Net-automated (2D U-Net Δ) / 3D U-Net-automated (3D U-Net Δ) segmentations highlight focal regions of $T_2$ discrepancies.