Automated Segmentation of Knee MRI Scans for Measuring Cartilage Thickness and Running-Induced Strain

Patrick X. Bradley, Sophia Y. Kim-Wang, Brooke S. Blaisdell, Alexie D. Riofrio, Amber T. Collins, Lauren N. Heckelman, Eziamaka C. Okafor, Margaret R. Widmyer, Chinmay S. Paranjape, Bryan S. Crook, Nimit K. Lad, Edward G. Sutter, Charles E. Spritzer, Louis E. DeFrate
Duke University, Durham, North Carolina, USA
Patrick.bradley@duke.edu

Disclosures: P. X. Bradley (N), S. Y. Kim-Wang (N), B. S. Blaisdell (N), A. D. Riofrio (N), A. T. Collins (N), L. N. Heckelman (N), E. C. Okafor (N), M. R. Widmyer (N), C. S. Paranjape (N), B. S. Crook (N), N. K. Lad (N), E. G. Sutter (N), C. E. Spritzer (N), L. E. DeFrate (AJSM, JoB, JOR)

Introduction: The capacity to effectively probe the mechanical function of cartilage is important due to the suggested link between osteoarthritis (OA) development and altered cartilage mechanics. Techniques that assess the mechanical function of cartilage *in vivo*, such as measuring cartilage strain in response to loading, commonly require participant-specific magnetic resonance imaging (MRI)-based 3D models of bone and cartilage [1,2]. However, creating these models typically involves manual segmentation, which is a labor-intensive process that ultimately hinders research throughput. Encouragingly, deep learning has recently shown the potential to automate some medical imaging segmentation tasks [3]. Thus, the objectives of this work were to train a suite of deep learning models to auto-segment the bone and cartilage from knee MRI scans, validate these models for measuring tibiofemoral cartilage thickness, and apply them to measure cartilage strains induced by running. Our hypotheses were that the deep learning models would produce repeatable measures of cartilage thickness and that running would induce both tibial and femoral cartilage compressive strain.

Methods: Four separate supervised deep learning models were trained, validated, and externally tested for automated segmentation of the tibia, femur, tibial cartilage, and femoral cartilage. ***Data*** **–** All data utilized to train and test our deep learning models were obtained from six previously published institutional review board-approved studies [4-9]. In total, model development utilized data from 21 participants (sex: 13M/8F; injury status: healthy; age: 22-48 years; BMI: 20.0-27.9 kg/m^2) for the bone models and 72 participants (sex: 51M/21F; injury status: healthy, ACL deficient, ACL reconstructed, or ACL and meniscus deficient; age: 22-48 years; BMI: 18.5-34.7 kg/m^2) for the cartilage models. All double-echo steady-state (DESS) MRI scans were obtained on the same scanner with identical imaging parameters (field of view: 16 cm x 16 cm; image resolution: 0.3 x 0.3 x 1.0 mm; flip angle: 25°; repetition time: 17 ms; echo time: 6 ms). Each participant contributed two MRI scans which were previously segmented by experienced researchers and reviewed by a musculoskeletal radiologist. Segmentations were converted into binary masks, cropped to 256 x 256-pixel regions of interest, paired with their respective image slice, grouped by participant, and divided into training/validation/testing datasets (Bone models = 32/6/4 scans; Cartilage models = 100/22/22 scans). ***Model Training*** **–** A 2D-UNet architecture was used for model training [3]. Training was performed using the Duke University high-performance computing cluster and consisted of a grid search performed over the following hyperparameters: batch size, filter size, kernel size, and learning rate. Model depth (5), loss function (dice coefficient loss), and epochs trained (500) all remained constant. Optimal models were determined by the highest validation set dice score achieved during training and were subsequently applied to the testing set. ***Cartilage Thickness Validation & External Application*** **–** Following model development, we validated our trained segmentation models for measuring tibiofemoral cartilage thickness and subsequently performed a measurement of running-induced cartilage strain. We utilized a previously published dataset investigating patellofemoral cartilage thickness before, immediately after, and 24 hours following a 3-mile treadmill run in 8 asymptomatic males (age: 27-40 years; BMI: 18-25 kg/m^2) [1]. In our analysis, we utilized the trained deep learning models to predict the tibiofemoral bone and cartilage masks, removed any clear outliers, created the 3D models, and measured cartilage thickness using a previously validated technique (Figure 1) [10]. We then evaluated cartilage thickness repeatability by calculating a two-way, mixed effects, multiple measurement, absolute agreement intraclass correlation coefficient (ICC) and the difference in group means between the two unloaded time points (pre-exercise and recovery). Additionally, we calculated cartilage strain at the post-exercise time point, where cartilage strain was defined as the change in cartilage thickness normalized to the pre-exercise thickness [10]. Two repeated measures ANOVAs with Fisher's Least Significant Difference post-hoc tests were performed to determine the influence of time point on tibial and femoral cartilage thickness. Differences were considered statistically significant where $p < 0.05$.

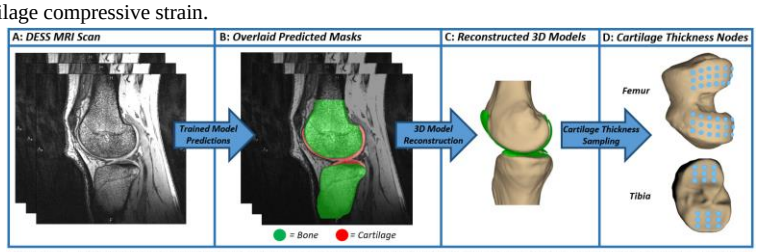

Figure 1. Cartilage Thickness Measurement Pipeline (A) The MRI scan is first input into the four trained deep learning models. **(B)** The models then output mask predictions for the tibia and femur (green) and tibial and femoral cartilage (red). **(C)** Masks for each tissue are stacked together, converted into point clouds of the tissue contours, and reconstructed into 3D models. **(D)** Cartilage thickness is measured at 18 locations across the tibial plateaus and 36 locations across the femoral condyles and then averaged together for each bone.

Results: Both bone segmentation models achieved testing set dice scores (DSC) > 0.980 (DSC_{tibia} = 0.988, DSC_{femur} = 0.990) and both cartilage segmentation models achieved testing set dice scores > 0.900 ($DSC_{tibial\ cartilage}$ = 0.901, $DSC_{femoral\ cartilage}$ = 0.913). Regarding day-to-day cartilage thickness repeatability, tibial and femoral cartilage measurements achieved ICCs of 0.984 and 0.987, respectively, and comparison of group means resulted in differences of 0.02 mm (0.7% of thickness) and 0.01 mm (0.4% of thickness) for the tibial and femoral cartilage. In our analysis of running-induced changes to cartilage thickness, we detected a significant effect of time point (Tibia: $p < 0.00001$, Femur: $p < 0.01$) with a mean post-exercise compressive strain of 5.4 ± 1.3% (mean ± 95% CI, $p < 0.0001$) for the tibial cartilage and 2.3 ± 1.7% (mean ± 95% CI, $p < 0.01$) for the femoral cartilage (Figure 2).

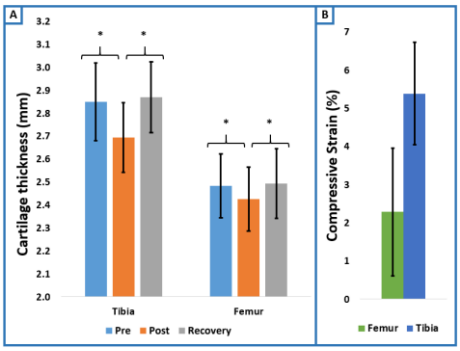

Figure 2. Cartilage thickness and strain (A) Mean ± 95% CI tibial and femoral cartilage thickness before, immediately after, and 24 hours after a 3-mile run. *$p < 0.05$ **(B)** Mean ± 95% CI tibial and femoral cartilage strain following a 3-mile run.

Discussion: In summary, we successfully trained, validated, and externally applied a suite of deep learning models for segmenting the tibia, femur, tibial cartilage, and femoral cartilage from knee MRI scans. Each of the four models achieved testing set DSCs indicative of substantial agreement between automatic and manual segmentations. Following model development, we evaluated cartilage thickness measurement repeatability by comparing thickness between the pre-exercise and recovery time points. This comparison encompassed both measurement error and day-to-day variance in cartilage thickness and resulted in ICCs that indicate excellent measurement repeatability, and differences in group means that fall within the previously reported manual measurement resolution of this technique (<1%) [10]. Further, in our external application we measured significant tibiofemoral cartilage compressive strains in response to a 3-mile run. Importantly, application of our trained deep learning models reduces the segmentation time from an order of days to minutes.

Significance: In this work, we demonstrated the efficacy of using trained deep learning segmentation models to expedite *in vivo* measures of cartilage function and further performed a novel measurement of tibiofemoral cartilage strains in response to a 3-mile run.

References: [1] Heckelman 2020. [2] Wang 2015. [3] Ronneberger 2015. [4] Widmyer 2013. [5] Sutter 2018. [6] Okafor 2014. [7] Lad 2016. [8] Paranjape 2019. [9] Crook 2021. [10] Coleman 2013.

Acknowledgements: Funding provided by NIH grants R01AR074800, R01AR065527, and R01AR079184.