Deep Learning Models Show Greater Longitudinal Sensitivity to Cartilage Thickness Change than Manual Segmentations in Osteoarthritis Initiative Foundation for the National Institutes of Health Cohort

McKenzie S. White[1], Anoosha Pai S[1], Kenneth T. Gao[2], Valentina Pedoia[2], Sharmila Majumdar[2], Garry E. Gold[1], Akshay S. Chaudhari[1], Anthony A. Gatti[1]
[1]Department of Radiology, Stanford University, Stanford, CA
[2]Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA
kenziew@stanford.edu

DISCLOSURES: A.C. has provided consulting services to Patient Square Capital, Chondrometrics GmbH, and Elucid Bioimaging; is co-founder of Cognita; has equity interest in Cognita, Subtle Medical, LVIS Corp, Brain Key. A.A.G is a shareholder of NeuralSeg, GeminiOV, and NodeAI.

INTRODUCTION: Accurate detection of longitudinal cartilage thickness change is essential, as thickness serves as a marker of osteoarthritis progression and an endpoint in clinical trials. Current practices rely heavily on cartilage thickness measurement derived from manual segmentations, which are not scalable for large cohorts and are subject to inter- and intra-reader variability. Previous studies demonstrate that deep learning (DL) enables rapid, automated segmentation at scale. Yet, beyond segmentation accuracy, it is critical that DL models capture true longitudinal changes in thickness with the same or greater sensitivity as manual methods. Moreover, prior DL models exclude denuded regions (where thickness = 0), inflating thickness estimates, which may make them insensitive to late-stage OA progression, which is typically the focus of clinical trials. The FNIH Biomarkers Consortium sub-cohort from the Osteoarthritis Initiative (OAI) includes patients stratified by progression of joint space loss (JSL) and pain: non-progressors, JSL progression only, pain progression only, both pain and JSL progression. The FNIH cohort has manual cartilage thickness measurements available with and without denuded regions. Whether including denuded regions affects sensitivity, and whether DL models match manual performance for detecting longitudinal cartilage thickness changes remains unclear. To address this, we compared the sensitivity of cartilage thickness changes between baseline and 24 months using bone and cartilage segmentations produced by DL models previously published from Stanford and UCSF[1,2,3], benchmarking them against manual reference measurements in the OAI FNIH cohort.

METHODS: We analyzed 586 knees from the OAI FNIH cohort with MRIs at baseline and 24-months. Bone and cartilage segmentations were obtained from two previously released DL models (Stanford and UCSF), and surface meshes were generated at both timepoints. To enable consistent regions of interest (ROI) across subjects, including denuded areas (thickness = 0), we fit a neural shape model (NSM) that was previously trained on 6,325 knees from the broader OAI cohort to 505 knees without structural damage (MRI Osteoarthritis Knee Score for cartilage morphology= 0) and defined subchondral bone ROIs following established standards [4]. ROIs only included bone vertices where ≥99% of the healthy cohort had overlying cartilage. Within each ROI, mean thickness was computed in two ways: (i) using segmentation-derived labels that exclude denuded areas and (ii) using the NSM-based standardized ROIs, which included denuded areas (thickness = 0 mm) in the mean calculation. Longitudinal sensitivity to change was quantified using the standardized response mean (SRM = meanΔ/SDΔ) with 95% bootstrap intervals. We assessed the impact of including denuded regions within each method by limb-level bootstrapping the difference in SRM between mean calculations that included and excluded denuded areas. Comparisons between methods (Stanford, UCSF, manual) were performed using the mean thickness measurement (denuded included vs. excluded) with the best SRM. For each DL method (Stanford, UCSF), mean thickness in the central medial and lateral femur at baseline and 24 months was compared with manual FNIH measurements using Pearson correlations. Pairwise differences in SRM between methods (Stanford, UCSF, manual) were tested using limb-level bootstrapping. Analyses were conducted overall and within the four FNIH clinical subgroups (non-progressors, JSL only, pain only, JSL + pain).

Table 1. Correlations in Mean Thickness

	Medial	Lateral
Stanford vs UCSF	0.978	0.976
Stanford vs manual	0.923	0.913
UCSF vs manual	0.902	0.888

RESULTS: Including denuded regions increased sensitivity to longitudinal change in the medial femur for all methods ($p \leq 0.008$), therefore, subsequent analyses compared thickness measures that included denuded regions. Correlations in mean thickness between the Stanford and UCSF models were excellent (r = 0.976-0.978), and manual measures correlated well but to a lower degree with each DL model (Table 1). All DL models detected cartilage thinning over 24 months in both the medial and lateral femur (Fig 1). In the medial femur, pairwise SRM differences were all significant ($p<0.05$), with values ranging from -0.65 (Stanford) to -0.45 (manual). In the lateral femur, the Stanford and UCSF models showed weak responsiveness (-0.166, -0.133, respectively) yet were significantly more sensitive to longitudinal change than manual segmentation ($p<0.05$); manual SRM=0.081(Fig 1). In the medial femur, the Stanford model was more sensitive than manual analyses in the JSL only (ΔSRM=−0.19, p=0.009) and JSL+pain subgroups (ΔSRM=−0.16, p=0.005), while in non-progressors and pain only groups, both DL models exceeded manual ($p \leq 0.006$). In the lateral femur, both DL models were more sensitive to longitudinal change than manual in all groups except JSL only progressors (Fig 2).

DISCUSSION: Ensuring standardized ROIs that included denuded regions (thickness = 0) improved sensitivity to longitudinal changes in thickness across all methods (DL and manual). Independently trained DL models correlated strongly with each other and, and to a lesser extent, with manual thickness measurements. In the medial femur, DL models showed greater sensitivity than manual measurements for detecting cartilage thinning overall and within predefined clinical subgroups.

SIGNIFICANCE/CLINICAL RELEVANCE: Automated DL pipelines offer a scalable, sensitive, and clinically meaningful alternative to manual cartilage measurements in large OA cohorts. Our results indicate that DL based approaches accelerate image processing, while also enhancing sensitivity to early structural progression at 24 months. Moving forward, we will expand this approach to the entire OAI and will publicly release all datasets, including segmentations, surface meshes, NSMs, and thickness measurements.

REFERENCES: [1] Gatti, 2021 [2] Iriondo, 2021 [3] Martinez, 2020 [4] Eckstein and Wirth, 2011

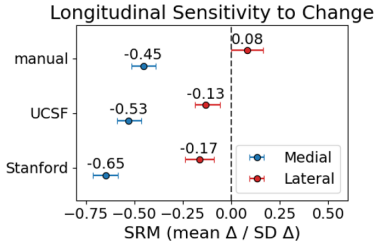

Figure 1. Standardized response means (SRMs) with 95% CIs for automated and manual methods, collapsed across all clinical subgroups.

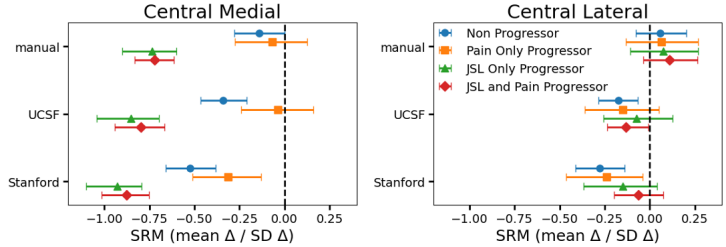

Figure 2. Standardized response means (SRMs) with 95% CIs for automated and manual methods stratified by clinical progression subgroups: joint space loss (JSL) only, pain only, combined JSL and pain, and non-progressors. The panels display SRMs for central medial (left) and central lateral (right) femoral cartilage.