Detection of Differences in Longitudinal Cartilage Thickness Loss Using a Deep-Learning Automated Segmentation Algorithm: Data From the Foundation for the National Institutes of Health Biomarkers Study of the Osteoarthritis Initiative

Journal Article

Author(s):
Eckstein, Felix; Chaudhari, Akshay S.; Fuerst, David; Gaisberger, Martin; Kemnitz, Jana; Baumgartner, Christian F.; Konukoglu, Ender; Hunter, David J.; Wirth, Wolfgang

Publication date:
2022-06

Permanent link:
https://doi.org/10.3929/ethz-b-000541786

Arthritis Care & Research
Vol. 74, No. 6, June 2022, pp 929–936
DOI 10.1002/acr.24539

American College of Rheumatology
Empowering Rheumatology Professionals

BRIEF REPORT

Detection of Differences in Longitudinal Cartilage Thickness Loss Using a Deep-Learning Automated Segmentation Algorithm: Data From the Foundation for the National Institutes of Health Biomarkers Study of the Osteoarthritis Initiative

Felix Eckstein,[1] Akshay S. Chaudhari,[2] David Fuerst,[1] Martin Gaisberger,[3] Jana Kemnitz,[4] Christian F. Baumgartner,[5] Ender Konukoglu,[5] David J. Hunter,[6] and Wolfgang Wirth[1]

Objective. To study the longitudinal performance of fully automated cartilage segmentation in knees with radiographic osteoarthritis (OA), we evaluated the sensitivity to change in progressor knees from the Foundation for the National Institutes of Health OA Biomarkers Consortium between the automated and previously reported manual expert segmentation, and we determined whether differences in progression rates between predefined cohorts can be detected by the fully automated approach.

Methods. The OA Initiative Biomarker Consortium was a nested case–control study. Progressor knees had both medial tibiofemoral radiographic joint space width loss (≥0.7 mm) and a persistent increase in Western Ontario and McMaster Universities Osteoarthritis Index pain scores (≥9 on a 0–100 scale) after 2 years from baseline (n = 194), whereas non-progressor knees did not have either of both (n = 200). Deep-learning automated algorithms trained on radiographic OA knees or knees of a healthy reference cohort (HRC) were used to automatically segment medial femorotibial compartment (MFTC) and lateral femorotibial cartilage on baseline and 2-year follow-up magnetic resonance imaging. Findings were compared with previously published manual expert segmentation.

Results. The mean ± SD MFTC cartilage loss in the progressor cohort was –181 ± 245 µm by manual segmentation (standardized response mean [SRM] –0.74), –144 ± 200 µm by the radiographic OA–based model (SRM –0.72), and –69 ± 231 µm by HRC-based model segmentation (SRM –0.30). Cohen's *d* for rates of progression between progressor versus the non-progressor cohort was –0.84 ($P < 0.001$) for manual, –0.68 ($P < 0.001$) for the automated radiographic OA model, and –0.14 ($P = 0.18$) for automated HRC model segmentation.

Conclusion. A fully automated deep-learning segmentation approach not only displays similar sensitivity to change of longitudinal cartilage thickness loss in knee OA as did manual expert segmentation but also effectively differentiates longitudinal rates of loss of cartilage thickness between cohorts with different progression profiles.

INTRODUCTION

Knee osteoarthritis (OA) severely affects the quality of life in an aging population and is responsible for substantial health care utilization and cost (1). OA treatments that go beyond symptomatic amelioration and modify the pathophysiology of the disease are an unmet clinical need. Quantitative magnetic resonance imaging (qMRI) had an important recent impact on the conduct of clinical

ClinicalTrials.gov identifier: NCT00080171.

Scientific and financial support for the Foundation for the NIH (FNIH) Osteoarthritis (OA) Biomarkers Consortium and for this study has been made possible through grants as well as direct and in-kind contributions from AbbVie, Amgen, Inc., the Arthritis Foundation, Bioiberica SA, DePuy Mitek, Inc., Flexion Therapeutics, Inc., GlaxoSmithKline, Merck Serono, Rottapharm | Madaus, Sanofi, Stryker, and the Pivotal Osteoarthritis Initiative Magnetic Resonance Imaging Analyses (POMA) study (NIH/National Institute of Arthritis and Musculoskeletal and Skin Diseases grant HHSN2682010000). The Osteoarthritis Initiative (OAI) is a public–private partnership between the NIH (contracts N01-AR-2-2258, N01-AR-2-2259, N01-AR-2-2260, N01-AR-2-2261, and N01-AR-2-2262) and private funding partners (Merck Research Laboratories, Novartis Pharmaceuticals, GlaxoSmithKline, and Pfizer, Inc.) and is conducted by the OAI Study Investigators. Private sector funding for the Biomarkers Consortium and the OAI is managed by the FNIH. Supported by Paracelsus Medical University (research fund grant E-18/27/146-WIK) and the Ludwig Boltzmann Institute for Arthritis and Rehabilitation, Austria.

[1]Felix Eckstein, MD, David Fuerst, PhD, Wolfgang Wirth, PhD: Paracelsus Medical University, Salzburg and Nuremberg, Salzburg, Austria, and Chondrometrics, Ainring, Germany; [2]Akshay S. Chaudhari, PhD: Stanford University, Stanford, California; [3]Martin Gaisberger, PhD: Paracelsus Medical University, Salzburg and Nuremberg, Salzburg, Austria; [4]Jana Kemnitz,

SIGNIFICANCE & INNOVATIONS

- This study investigated the longitudinal performance characteristics of an automated, convolutional neural network–based cartilage segmentation method using magnetic resonance imaging, including the sensitivity to change of cartilage thickness loss, and the ability to efficiently differentiate rates of cartilage loss between different strata (e.g., progressor knees versus non-progressor knees).
- The fully automated segmentation approach not only displayed similar sensitivity to change of longitudinal cartilage thickness loss in knee osteoarthritis (OA) for automated versus manual expert segmentation but was also able to effectively differentiate longitudinal rates of cartilage thickness loss between cohorts with different progression profiles.
- The method therefore shows great promise in leveraging the application of quantitative analysis methods of cartilage thickness loss in clinical trials investigating the structural progression of knee OA.

trials on potential disease-modifying OA drugs (DMOADs) (2–4), with longitudinal qMRI cartilage thickness change being increasingly used as the primary structural end point for potential regulatory approval (2–4).

An (imaging) biomarker exhibiting near-term change that is associated with longer-term, clinically important outcomes has potential as a marker of the treatment efficacy of DMOADs. Therefore, the Foundation for the National Institutes of Health (FNIH) OA Biomarkers Consortium study was conducted to evaluate the association of imaging and molecular biomarkers with structural (radiographic) and symptomatic (pain) progression in knee OA (5). Medial femorotibial compartment (MFTC) cartilage thickness loss over 24 months was shown to be associated with combined radiographic and symptomatic progression, and the association was shown to be stronger for radiographic progression than for pain progression (6). Use of imaging biomarkers in clinical trials and eventually in clinical practice will be greatly facilitated by the availability of fully automated measurement technology that can be scaled to faster turnaround. With a potential DMOAD coming to market, large-scale cartilage morphometric measuring in clinical practice may become a high demand for monitoring individual treatment response and the need for intermittent versus continuous treatment.

Quantitative MRI of articular cartilage currently requires time-consuming expert image segmentation. Various semiautomated or fully automated analysis methods have been proposed to overcome this limitation; among these are deep-learning convolutional neural networks (CNNs) (7) and, specifically, U-Net architectures (8). However, no CNN-based method has thus far tested the sensitivity to longitudinal change or the ability to efficiently differentiate rates of cartilage loss between progressor knees versus non-progressor knees (5,6). Yet, this is of crucial importance, given that quantitative measurement of cartilage is almost exclusively used in longitudinal context in clinical research, with only small changes being observed over time (2–4).

Recently, we examined the accuracy of U-Net–based automated cartilage segmentation in the OA Initiative (OAI) healthy reference cohort (HRC) and reported high correlations between cartilage morphometry using automated versus expert manual segmentation as well as similar test–retest precision between both (9). The purpose of the current study was to examine the longitudinal performance of automated deep-learning U-Net–based cartilage segmentation in knees with radiographic OA (OA). Specifically, we tested whether the sensitivity to change in knees with combined radiographic and symptomatic progression from the FNIH Biomarkers Consortium (5), and whether differences in rates of progression between progressor and non-progressor knees (5), are similar between the novel automated, and previously reported (6), manual expert segmentation approach.

MATERIALS AND METHODS

Study Design. The FNIH Biomarker Consortium was a nested case–control study (5,6) using data from the OAI (10). Eligible participants had at least 1 knee with baseline Kellgren/Lawrence (K/L) grade 1–3 from central radiographic readings, baseline and 24-month knee radiographs and knee MRI, serum and urine specimens, and clinical data (5,6). Radiographs of the knee in fixed flexion were assessed for K/L grade and OA Research Society International (OARSI) joint space narrowing (JSN) grades (5,6). Medial radiographic progression was defined by a loss in minimum radiographic joint space width (JSW) of ≥0.7 mm from baseline to 24, 36, or 48 months; knee pain was assessed using the Western Ontario

PhD: Paracelsus Medical University, Salzburg and Nuremberg, and Siemens, Vienna, Austria; [5]Christian F. Baumgartner, PhD, Ender Konukoglu, PhD: ETH, Zurich, Switzerland; [6]David J. Hunter, MBBS: Royal North Shore Hospital and University of Sydney, Sydney, New South Wales, Australia.

Dr. Eckstein has received consulting fees from Merck, Samumed, Kolon-Tissuegene, Servier, Galapagos, Roche, Novartis, and ICM (less than $10,000 each) and owns stock or stock options in Chondrometrics. Dr. Chaudhari has received consulting fees from Skope MR, Chondrometrics, Image Analysis Group, Edge Analytics, Culvert Engineering (less than $10,000 each), and Subtle Medical (more than $10,000) and owns stock or stock options in Subtle Medical, LVIS Corporation, and Brain Key. Dr. Kemnitz owns stock or stock options in Siemens. Dr. Hunter has received consulting fees from Merck Serono, Pfizer, Eli Lilly and Company, and TLCBio (less than $10,000 each). Dr. Wirth has received consulting fees from Galapagos (less than $10,000) and owns stock or stock options in Chondrometrics. No other disclosures relevant to this article were reported.

Address correspondence to Felix Eckstein, MD, Institute of Anatomy & Cell Biology, Paracelsus Medical University, Strubergasse 21, A-5020 Salzburg, Austria. Email: felix.eckstein@pmu.ac.at.

Submitted for publication August 6, 2020; accepted in revised form December 15, 2020.

Table 1. Demographic and baseline cartilage thickness data for the analyzed set of the Foundation for the National Institutes of Health cohort*

Characteristic	JSW + pain progression	Non-progression	JSW progression only	Pain progression only
Analyzed set, no.	192	200	103	102
Age, years	62.0 ± 8.8	61.5 ± 9.1	63.1 ± 8.3	59.2 ± 8.7
Female, no. (%)	109 (57)	130 (65)	46 (45)	66 (65)
BMI, kg/m^2	30.7 ± 4.8	30.5 ± 4.8	30.7 ± 4.7	31.0 ± 5.0
K/L grade 1/2/3, no.	24/82/86	24/114/62	14/47/42	13/60/29
Medial JSN grade 0/1/2, no.	42/64/86	67/71/62	19/42/42	32/41/29
Lateral JSN grade 0/1/2, no.	188/4/0	196/4/0	102/1/0	99/3/0
Cartilage thickness				
MFTC, mm				
Expert manual	3.2 ± 0.7	3.4 ± 0.6	3.4 ± 0.6	3.3 ± 0.6
Auto ROA	3.5 ± 0.6	3.6 ± 0.5	3.7 ± 0.6	3.6 ± 0.6
Auto HRC	3.6 ± 0.5	3.6 ± 0.5	3.7 ± 0.5	3.5 ± 0.5
Auto ROA + HRC	3.5 ± 0.6	3.6 ± 0.5	3.7 ± 0.5	3.5 ± 0.6
LFTC, mm				
Expert manual	3.9 ± 0.6	3.8 ± 0.6	4.0 ± 0.6	3.8 ± 0.6
Auto ROA	4.1 ± 0.6	4.0 ± 0.6	4.2 ± 0.6	4.0 ± 0.6
Auto HRC	4.0 ± 0.6	3.9 ± 0.5	4.0 ± 0.5	3.9 ± 0.5
Auto ROA + HRC	4.0 ± 0.6	3.9 ± 0.6	4.1 ± 0.6	3.9 ± 0.5

* Values are the mean ± SD unless indicated otherwise. Auto HRC = automatic segmentation algorithm trained on a sample of healthy reference cohort knees; auto ROA = automatic segmentation algorithm trained on a sample of knees with radiographic osteoarthritis; auto ROA + HRC = automatic segmentation algorithm trained on a sample of a combined set of ROA and HRC knees (only n = 101 knees in joint space width [JSW] progressors group); BMI = body mass index; expert manual = expert manual segmentation; JSN = joint space narrowing (according to Osteoarthritis Research Society International Atlas); K/L = Kellgren/Lawrence; LFTC = lateral femorotibial compartment; MFTC = medial femorotibial compartment.

and McMaster Universities Osteoarthritis Index (WOMAC) pain subscale, with progression defined as a persistent (≥2 time points) increase of ≥9 points on a 0–100 normalized score from baseline to 24, 36, 48, or 60 months (5,6).

In the FNIH Biomarker Consortium study, primary cases were as follows: 1) knees that had both radiographic and pain progression (progressor cohort; n = 194); 2) control knees that did not have this combination and included knees with neither radiographic nor pain progression (n = 200); 3) knees with radiographic but not pain progression (n = 103); and 4) knees with pain but not radiographic progression (n = 103) (5,6). For better covariate balance, the knees selected for the 4 groups were frequency matched, using K/L grade and body mass index. Cartilage thickness and bone shape biomarkers were previously shown to be strongly associated with radiographic (but not pain) progression (6,11); therefore, only knees with neither radiographic nor pain progression were used in this study as non-progressor controls. Sensitivity analyses were conducted for partial progressors (i.e., knees with minimum JSW or pain progression only).

Expert manual and automated deep-learning cartilage thickness measurement. Manual expert segmentation of femorotibial cartilage thickness published in the FNIH Biomarker study had relied on double-echo steady-state (DESS) imaging, with blinding to group assignment and order of acquisition (6). Segmentation encompassed the total medial and lateral tibia and the weight-bearing (central) medial and lateral femoral condyles. All segmentations had been quality controlled by an expert (6), and a 75% femoral region of interest (ROI) (distance between the trochlear notch and the posterior ends of the condyles) was used (6).

The automated segmentation method used here was based on a 2-D U-Net architecture (8,12) and was trained, validated, and tested also using sagittal DESS images. The U-Net was trained using a weighted cross entropy loss function with equal weights for each of the foreground features (i.e., cartilages), and with the background weight set to one-half of the one used for the foreground, which was minimized, using Adam optimization (initial learning rate 0.01), as published previously (9). The software was implemented in Python (Python Software Foundation) using the Tensorflow framework (Google) (9). Three algorithms were used: one trained on 52 knees of the HRC of the OAI without radiographic signs, symptoms, or risk factors of knee OA (9); one trained on 86 OAI knees with radiographic OA (K/L grade 2, 3, and 4 = 35%, 34%, and 31%, respectively); and one trained on all of the above 138 knees (combined model). The training was performed using full-resolution, full-sized MRI slices on an RTX 2080TI graphics processing unit (Nvidia) (9). The performance of the algorithm trained on HRC knees was validated and tested in 21 of 21 HRC knees, with the automated segmentations displaying high agreement (Dice similarity coefficients), high accuracy, and test–retest precision of cartilage thickness computations (9).

The algorithms trained on the HRC knees, radiographic OA knees, or the combined set were then applied to the FNIH sample (5,6), with none of the FNIH knees being including in the training sets. Automated segmentations were neither quality controlled nor manually corrected to explore the performance of the automated approach without manual intervention. Yet, some fully

Table 2. Cartilage thickness loss over 24 months in the medial femorotibial compartment (MFTC) in the 4 Foundation for the National Institutes of Health cohorts*

	JSW + pain progression	Non-progression	JSW progression only	Pain progression only
MFTC, μm				
Expert manual	−181 ± 245	−22 ± 108	−184 ± 252	−8 ± 119
95% CI	−216, −146	−37, −7	−233, −135	−32, 15
SRM (95% CI)	−0.74 (−0.85, −0.63)	−0.21 (−0.36, −0.07)	−0.73 (−0.88, −0.55)	−0.07 (−0.28, 0.13)
Auto ROA	−144 ± 200	−33 ± 121	−151 ± 251	−15 ± 113
95% CI	−172, −116	−49, −16	−200, −102	−38, 7
SRM (95% CI)	−0.72 (−0.86, −0.57)	−0.27 (−0.39, −0.12)	−0.60 (−0.75, −0.44)	−0.14 (−0.32, 0.07)
Auto HRC	−69 ± 231	−42 ± 130	−96 ± 227	−5 ± 146
95% CI	−102, −36	−60, −24	−140, −51	−34, 23
SRM (95% CI)	−0.30 (−0.47, −0.05)	−0.33 (−0.45, −0.19)	−0.42 (−0.59, −0.23)	−0.04 (−0.22, 0.17)
Auto ROA + HRC	−116 ± 284	−29 ± 122	−150 ± 245	12 ± 134
95% CI	−157, −76	−46, −12	−199, −102	−15, 38
SRM (95% CI)	−0.41 (−0.72, −0.08)	−0.24 (−0.37, −0.11)	−0.61 (−0.77, −0.45)	0.09 (−0.11, 0.28)
MT, μm				
Expert manual	−55 ± 100	−11 ± 54	−48 ± 97	−4 ± 51
95% CI	−70, −41	−18, −3	−67, −29	−14, 6
SRM (95% CI)	−0.56 (−0.67, −0.43)	−0.20 (−0.34, −0.06)	−0.49 (−0.65, −0.32)	−0.09 (−0.31, 0.09)
Auto ROA	−31 ± 79	−6 ± 57	−27 ± 91	−2 ± 50
95% CI	−42, −20	−14, 2	−45, −9	−12, 8
SRM (95% CI)	−0.40 (−0.54, −0.25)	−0.10 (−0.23, 0.04)	−0.30 (−0.45, −0.13)	−0.04 (−0.25, 0.17)
Auto HRC	−2 ± 158	−12 ± 58	−8 ± 91	0 ± 60
95% CI	−25, 21	−20, −4	−25, 10	−12, 12
SRM (95% CI)	−0.01 (−0.19, 0.15)	−0.21 (−0.33, −0.05)	−0.08 (−0.26, 0.14)	0.00 (−0.20, 0.19)
Auto ROA + HRC	−14 ± 185	−7 ± 62	−24 ± 97	10 ± 82
95% CI	−41, 12	−16, 2	−44, −5	−6, 26
SRM (95% CI)	−0.08 (−0.35, 0.11)	−0.11 (−0.24, 0.02)	−0.25 (−0.41, −0.05)	0.12 (−0.07, 0.28)
cMF, μm				
Expert manual	−126 ± 175	−12 ± 79	−136 ± 184	−4 ± 93
95% CI	−151, −101	−23, −1	−172, −100	−22, 14
SRM (95% CI)	−0.72 (−0.82, −0.61)	−0.15 (−0.31, −0.01)	−0.74 (−0.88, −0.58)	−0.04 (−0.25, 0.16)
Auto ROA	−113 ± 148	−27 ± 88	−124 ± 185	−13 ± 88
95% CI	−134, −92	−39, −14	−160, −87	−31, 4
SRM (95% CI)	−0.76 (−0.90, −0.56)	−0.30 (−0.44, −0.16)	−0.67 (−0.82, −0.51)	−0.15 (−0.33, 0.06)
Auto HRC	−67 ± 136	−30 ± 93	−88 ± 166	−5 ± 103
95% CI	−86, −47	−43, −17	−121, −56	−25, 15
SRM (95% CI)	−0.49 (−0.65, −0.34)	−0.33 (−0.44, −0.19)	−0.53 (−0.71, −0.35)	−0.05 (−0.23, 0.15)
Auto ROA + HRC	−102 ± 145	−22 ± 83	−126 ± 181	2 ± 82
95% CI	−123, −82	−34, −11	−162, −90	−14, 18
SRM (95% CI)	−0.71 (−0.85, −0.53)	−0.27 (−0.41, −0.14)	−0.70 (−0.84, −0.53)	0.02 (−0.17, 0.22)

* Values are the mean ± SD unless indicated otherwise. Data shown for combined radiographic joint space width and pain (JSW + pain) progressors, non-progressors, radiographic JSW but not pain progressors (JSW progressors only), and pain but not JSW progressors (pain progressors only). Analyzed by expert manual segmentation, by an automated algorithm trained on a radiographic osteoarthritis knee sample (auto ROA), a healthy reference cohort (auto HRC) sample, and by a combined (auto ROA + HRC) sample. 95% CI = 95% confidence interval; cMF = medial weight-bearing femoral condyle; MT = medial tibia; SRM = standardized response mean.
† ROA + HRC: 101 knees in the JSW progression only group.

automated postprocessing was applied, such as filling small gaps by detecting enclosed unsegmented areas, removal of implausible segmentations (e.g., fragments not connected to the main segmentation and those sticking out of the cartilage surface), and removal of femoral cartilage segmentations outside the ROI (9). Cartilage thickness was computed from the automatically segmented contours in the same way as from the manual ones using Chondrometrics software (9).

Statistical analysis. The primary descriptive analytic end point was the comparison of the standardized response mean (SRM; the mean change from baseline to 24 months' follow-up divided by the SD of the change) for the automated radiographic OA model versus expert manual analysis in the progressor cohort in the MFTC (the sum of medial tibia and medial femoral condyles). The SRMs and 95% confidence intervals (95% CIs) were computed using bias-corrected and accelerated bootstrapping (1,000 iterations). The primary analytic end point was the difference in longitudinal MFTC cartilage thickness loss over 24 months between the progressor versus non-progressor cohort and between the radiographic OA model and the expert manual analysis. These were compared between the automated and expert manual analysis using *t*-tests and Cohen's *d* as a measure of effect size, including their 95% CIs (13).

Table 3. Cartilage thickness loss over 24 months in the lateral femorotibial compartment (LFTC) in the 4 Foundation for the National Institutes of Health cohorts*

	JSW + pain progression	Non-progression	JSW progression only	Pain progression only
LFTC, μm				
Expert manual	−16 ± 124	−21 ± 113	−18 ± 111	−12 ± 95
95% CI	−34, 1	−37, −5	−40, 4	−30, 7
SRM (95% CI)	−0.13 (−0.26, 0.02)	−0.19 (−0.32, −0.04)	−0.16 (−0.40, 0.04)	−0.12 (−0.33, 0.07)
Auto ROA	−60 ± 175	−71 ± 176	−51 ± 188	−44 ± 135
95% CI	−85, −35	−96, −47	−88, −15	−71, −18
SRM (95% CI)	−0.34 (−0.47, −0.20)	−0.41 (−0.51, −0.28)	−0.27 (−0.45, −0.06)	−0.33 (−0.50, −0.12)
Auto HRC	−73 ± 209	−73 ± 202	−72 ± 201	−52 ± 168
95% CI	−103, −43	−102, −45	−111, −32	−85, −19
SRM (95% CI)	−0.35 (−0.44, −0.25)	−0.36 (−0.48, −0.23)	−0.36 (−0.51, −0.18)	−0.31 (−0.48, −0.08)
Auto ROA + HRC†	−47 ± 152	−45 ± 130	−39 ± 149	−27 ± 100
95% CI	−68, −25	−63, −27	−69, −10	−47, −7
SRM (95% CI)	−0.31 (−0.42, −0.19)	−0.35 (−0.47, −0.23)	−0.26 (−0.43, −0.06)	−0.27 (−0.49, −0.06)
LT, μm				
Expert manual	−26 ± 61	−26 ± 63	−24 ± 65	−18 ± 57
95% CI	−34, −17	−34, −17	−36, −11	−29, −6
SRM (95% CI)	−0.42 (−0.56, −0.26)	−0.41 (−0.54, −0.27)	−0.36 (−0.59, −0.13)	−0.31 (−0.50, −0.12)
Auto ROA	−43 ± 96	−48 ± 108	−33 ± 104	−37 ± 74
95% CI	−57, −30	−63, −33	−54, −13	−51, −22
SRM (95% CI)	−0.45 (−0.57, −0.33)	−0.44 (−0.55, −0.33)	−0.32 (−0.48, −0.10)	−0.50 (−0.68, −0.32)
Auto HRC	−37 ± 104	−39 ± 100	−30 ± 92	−31 ± 104
95% CI	−52, −22	−53, −25	−48, −12	−51, −10
SRM (95% CI)	−0.35 (−0.45, −0.23)	−0.39 (−0.51, −0.26)	−0.33 (−0.51, −0.12)	−0.30 (−0.47, −0.03)
Auto ROA + HRC†	−33 ± 81	−30 ± 75	−28 ± 72	−23 ± 55
95% CI	−44, −21	−41, −20	−43, −14	−34, −12
SRM (95% CI)	−0.40 (−0.51, −0.28)	−0.40 (−0.53, −0.26)	−0.39 (−0.59, −0.19)	−0.42 (−0.62, −0.23)
cLF, μm				
Expert manual	9 ± 84	4 ± 74	6 ± 70	6 ± 59
95% CI	−3, 21	−6, 15	−8, 19	−6, 17
SRM (95% CI)	0.11 (−0.03, 0.27)	0.06 (−0.08, 0.20)	0.08 (−0.12, 0.27)	0.10 (−0.11, 0.29)
Auto ROA	−16 ± 101	−24 ± 95	−18 ± 106	−8 ± 77
95% CI	−31, −2	−37, −10	−38, 3	−23, 8
SRM (95% CI)	−0.16 (−0.30, −0.01)	−0.25 (−0.37, −0.11)	−0.17 (−0.37, 0.02)	−0.10 (−0.29, 0.11)
Auto HRC	−36 ± 124	−35 ± 117	−41 ± 134	−21 ± 82
95% CI	−54, −19	−51, −18	−67, −15	−37, −5
SRM (95% CI)	−0.29 (−0.41, −0.18)	−0.30 (−0.40, −0.18)	−0.31 (−0.46, −0.14)	−0.26 (−0.43, −0.06)
Auto ROA + HRC†	−14 ± 94	−15 ± 81	−11 ± 100	−4 ± 65
95% CI	−28, −1	−26, −4	−31, 9	−17, 9
SRM (95% CI)	−0.15 (−0.28, −0.02)	−0.18 (−0.31, −0.05)	−0.11 (−0.29, 0.09)	−0.06 (−0.25, 0.15)

* Values are the mean ± SD unless indicated otherwise. Data shown for combined radiographic joint space width and pain (JSW + pain) progressors, non-progressors, radiographic JSW but not pain progressors (JSW progressors only), and pain but not JSW progressors (pain progressors only). Analyzed by expert manual segmentation, by an automated algorithm trained on a radiographic osteoarthritis knee sample (auto ROA), a healthy reference cohort (auto HRC) sample, and by a combined (auto ROA + HRC) sample. 95% CI = 95% confidence interval; cLF = lateral weight-bearing femoral condyle; LT = lateral tibia; SRM = standardized response mean.
† ROA + HRC: 101 knees in the JSW progression only group.

Furthermore, we determined the number (proportion) of individual progressors in both cohorts, defined by published thresholds from OAI pilot study test–retest analyses (14).

RESULTS

From the 600 FNIH Consortium knees, automated computation was successful for 597: 192 (of 194) progressor, 200 (of 200) non-progressor, 103 (of 103) partial JSW progressor, and 102 (of 103) partial pain progressor knees, of which 1 also did not have manual expert segmentation due to insufficient image quality. The computation time for the automated segmentation, postprocessing, and morphometric analysis was <1 minute per visit. Table 1 lists the demographic characteristics and baseline cartilage thickness values of the analyzed set, obtained from expert manual segmentation, radiographic OA model segmentation, and HRC model segmentation, respectively. The automated algorithms somewhat overestimated the baseline cartilage thickness in the MFTC and lateral femorotibial compartment (LFTC) (Table 1).

The mean ± SD MFTC cartilage loss in the progressor cohort was −181 ± 245 μm by manual expert segmentation (SRM −0.74), −144 ± 200 μm by radiographic OA–based model segmentation (SRM −0.72), and −69 ± 231 μm by HRC-based

model segmentation (SRM –0.30) (Table 2). The SRMs in the non-progressor cohort were –0.21, –0.27, and –0.33, respectively (Table 2). Overall, the SRM was greatest for the medial femoral condyles when applying the radiographic OA model algorithm (–0.76) and was somewhat less for manual segmentation (–0.72). The SRMs were less for the medial tibia (–0.40 for the radiographic OA model and –0.56 for expert manual segmentation) than for the medial femoral condyles (Table 2). Interestingly, the SRMs for the combined (radiographic OA plus HRC) model revealed smaller sensitivity to change than for the radiographic OA–based model but greater sensitivity than for the HRC-based model (Table 2).

Cohen's *d* for differences in rates of MFTC progression between the progressor versus non-progressor cohort was –0.85 (95% CI –1.05, –0.64; $P < 0.001$), for expert manual segmentation, –0.68 (95% CI –0.88, –0.47; $P < 0.001$), for the radiographic OA model, and –0.15 (95% CI –0.34, –0.05; $P = 0.16$), for the HRC-model automated segmentation. For the medial femoral condyles, these values were as follows: expert manual segmentation –0.85 (95% CI –1.05, –0.64; $P < 0.001$); radiographic OA model –0.71 (95% CI –0.91, –0.51; $P < 0.001$); and HRC-model automated segmentation –0.32 (95% CI –0.52, –0.12; $P = 0.002$); and for the medial tibia: expert manual segmentation –0.55 (95% CI –0.75, –0.34; $P < 0.001$); radiographic OA model –0.36 (95% CI –0.56, –0.16; $P < 0.001$); and HRC-model automated segmentation 0.09 (95% CI –0.11, 0.28; $P = 0.41$). Results for the combined (radiographic OA plus HRC) model revealed less discrimination compared with the radiographic OA–based model but a larger effect size compared with the HRC-based model algorithm (Table 2).

Application of test–retest thresholds identified 102 (53%) MFTC progressors in the progressor cohort versus 41 (21%) in the non-progressor cohort using expert manual segmentation, and 102 (53%) versus 45 (23%), respectively, using the radiographic OA model algorithm. Of the 102 progressor knees identified by expert manual segmentation, 78 (76%) were also detected using the radiographic OA model algorithm.

Observations in the MFTC in the partial JSW progressor cohort were consistent with those made in the progressor cohort, whereas those made in the partial pain progressor cohort were consistent with those made in the non-progressor cohort (Table 2). Results for the LFTC are shown in Table 3, with the automated algorithms producing higher SRMs than manual segmentation in all 4 cohorts. No relevant or statistically significant differences in the rates of LFTC cartilage thickness loss were observed between both cohorts using either expert manual or automated segmentation methods.

DISCUSSION

The current study evaluated the performance of deep-learning algorithms for the longitudinal measurement of articular cartilage thickness in knee OA and the ability of automated segmentation methodology to discriminate longitudinal rates of cartilage thickness loss between 2 cohorts with different progression profiles. The sensitivity to change of MFTC cartilage thickness measurements in the progressor cohort of the FNIH biomarker study was similar for the automated radiographic OA model algorithm when compared to expert manual segmentation, with 95% CIs of the SRM completely overlapping, whereas it was considerably less for an algorithm trained on HRC data, with the 95% CIs of the SRMs not overlapping at all. The discrimination of the rates of MFTC cartilage loss between progressor versus non-progressor knees (Cohen's *d*) was only slightly less for the automated algorithm of the radiographic OA model than for expert manual segmentation due to relatively greater sensitivity to change in the non-progressor cohort observed with the automated algorithm compared with manual expert segmentation. A high proportion of individual progressor knees identified by expert manual segmentation also was identified by automated segmentation. No satisfactory discrimination between progressor versus non-progressor knees was observed for the algorithm trained on HRC data. Furthermore, the discrimination was less for a combined model trained on both radiographic OA and HRC data together than for the one trained on radiographic OA knees alone.

Only 2 knees in the FNIH biomarker sample (0.3%) that had expert manual segmentation could not be successfully analyzed using the automated algorithm; one with extensive osteophytes, and one in which automated segmentation included portions of non-cartilage tissue that precluded successful thickness computation. This is encouraging, as the FNIH biomarker sample displayed a similar distribution of radiographic knee OA as do clinical trials that test the efficacy of DMOADs (2–4). Dice similarity coefficients and other (cross-sectional) performance metrics for the algorithm used have been reported previously (9).

A limitation of the current study is that the results are specific to the sagittal DESS MRI sequences, whereas clinical trials often rely on spoiled gradient-recalled acquisition in the steady state (SPGR), fast low-angle shoot (FLASH), and fast field echo (FFE) MRI sequences. However, manual expert segmentation of the coronal FLASH images was not available for the FNIH study, and coronal FLASH MRIs are only available for right knees in the OAI (11), whereas the FNIH biomarker study included a balanced mix of left and right knees. Yet, future work should compare the differential sensitivity to change between different MRI protocols and image orientations.

The strengths of the current study include the use of a radiographic OA–based, HRC-based, as well as a combined model. The results are interesting in that the larger (radiographic OA plus HRC) model was less effective in detecting longitudinal change than the radiographic OA model alone, suggesting that the specificity of the training set may be more important than its size. Future work should explore whether training sets specific to K/L grade or JSN perform better than a single model using various

K/L and JSN grades. Another strength of the current study is that the sensitivity to change of the automated algorithms was compared with expert manual segmentation that had thorough expert quality control in the relatively large FNIH biomarkers study, which contained several cohorts with differential rates of progression, and for which other biomarker results have been documented.

A recent study reported a sensitivity to change of –0.43 (SRM) over 2 years for medial femoral condyles in manual segmentation and of –0.67 in fully automated segmentation relying on a shape-based segmentation approach (15), whereas we found an SRM of –0.76 for automated and –0.72 for manual segmentation. Although these results cannot be directly compared across different OAI samples, the previous study (15) did not explore the performance of automated measurement technology in differentiating subpopulations with different progression profiles. Given the limited space and references in this brief report, semi- (rather than fully) automated approaches of cartilage segmentation from MRI, as well as those including only local cartilage measurements rather than the complete MFTC and/or LFTC, are not included in this discussion, nor are the findings of studies that have explored other (imaging) biomarkers than cartilage thickness change in the FNIH Consortium.

In conclusion, we found that a fully automated segmentation approach using deep learning not only displayed similar sensitivity to change of longitudinal cartilage thickness loss in knee OA compared with manual expert segmentation and expert quality control but also effectively differentiated longitudinal rates of cartilage thickness loss between cohorts with different progression profiles. These results are promising in that such automated measurement technology can be scaled to large reading volumes in clinical practice and in clinical trials testing the efficacy of DMOAD therapy, where rates of progression of participants treated with drugs versus placebo need to be tracked accurately.

ACKNOWLEDGMENTS

We thank the expert readers at Chondrometrics (Gudrun Goldmann, Linda Jakobi, Manuela Kunz, Dr. Susanne Maschek, Jana Matthes, Sabine Mühlsimer, Annette Thebis, and Dr. Barbara Wehr) for the manual expert segmentation; we particularly thank Dr. Susanne Maschek for quality control readings of the segmentations. Further, we thank the readers of the fixed flexion radiographs at Boston University for the central K/L grading, the OAI investigators, clinic staff, and OAI participants at each of the OAI clinical centers for their contributions in acquiring the publicly available clinical and imaging data, the team at the OAI coordinating center, FNIH, and the members of the FNIH OA Biomarker Consortium and the OARSI for their leadership and expertise on the FNIH OA Consortium project.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Eckstein had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Eckstein, Chaudhari, Hunter, Wirth.
Acquisition of data. Hunter, Wirth.
Analysis and interpretation of data. Eckstein, Chaudhari, Fuerst, Gaisberger, Kemnitz, Baumgartner, Konukoglu, Hunter, Wirth.

ADDITIONAL DISCLOSURES

Authors Eckstein, Fuerst, and Wirth are employees of Chondrometrics. Author Kemnitz is an employee of Siemens.

REFERENCES

1. Wright EA, Katz JN, Cisternas MG, Kessler CL, Wagenseller A, Losina E. Impact of knee osteoarthritis on health care resource utilization in a US population-based national sample. Med Care 2010;48:785–91.
2. Hochberg MC, Guermazi A, Guehring H, Aydemir A, Wax S, Fleuranceau-Morel P, et al. Effect of intra-articular sprifermin vs placebo on femorotibial joint cartilage thickness in patients with osteoarthritis. JAMA 2019;322:1360.
3. Conaghan PG, Bowes MA, Kingsbury SR, Brett A, Guillard G, Rizoska B, et al. Disease-modifying effects of a novel cathepsin k inhibitor in osteoarthritis: a randomized controlled trial. Ann Intern Med 2020; 172:86–95.
4. Cai G, Aitken D, Laslett LL, Pelletier JP, Martel-Pelletier J, Hill C, et al. Effect of intravenous zoledronic acid on tibiofemoral cartilage volume among patients with knee osteoarthritis with bone marrow lesions: a randomized clinical trial. JAMA 2020;323:1456–66.
5. Hunter DJ, Nevitt M, Losina E, Kraus V. Biomarkers for osteoarthritis: current position and steps towards further validation. Best Pract Res Clin Rheumatol 2014;28:61–71.
6. Eckstein F, Collins JE, Nevitt MC, Lynch JA, Kraus VB, Katz JN, et al. Cartilage thickness change as an imaging biomarker of knee osteoarthritis progression: data from the Foundation for the National Institutes of Health Osteoarthritis Biomarkers Consortium. Arthritis Rheumatol 2015;67:3184–9.
7. 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 2018;79:2379–91.
8. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Int Conf Med image Comput Comput Interv 2015:234–41.
9. Wirth W, Eckstein F, Kemnitz J, Baumgartner CF, Konukoglu E, Fuerst D, 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. MAGMA 2021;34:337–54.
10. Eckstein F, Kwoh CK, Link TM. Imaging research results from the Osteoarthritis Initiative (OAI): a review and lessons learned 10 years after start of enrolment. Ann Rheum Dis 2014;73:1289–300.
11. Hunter D, Nevitt M, Lynch J, Kraus VB, Katz JN, Collins JE, et al. Longitudinal validation of periarticular bone area and 3D shape as biomarkers for knee OA progression? Data from the FNIH OA Biomarkers Consortium. Ann Rheum Dis 2016;75:1607–14.
12. Baumgartner CF, Koch LM, Pollefeys M, Konukoglu E. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In: Pop M, Sermesant M, Jodoin PM, Lalande A, Zhuang X, Yang G, et al, editors. Statistical atlases and computational models of the heart: ACDC and MMWHS challenges. Springer International Publishing; 2018 p. 111–9.
13. Hedges LV, Olkin I. Statistical methods for meta-analysis. Elsevier; 1985.

14. Wirth W, Larroque S, Davies RY, Nevitt M, Gimona A, Baribaud F, et al. Comparison of 1-year vs 2-year change in regional cartilage thickness in osteoarthritis results from 346 participants from the Osteoarthritis Initiative. Osteoarthritis Cartilage 2011;19: 74–83.

15. Bowes MA, Guillard GA, Vincent GR, Brett AD, Wolstenholme CB, Conaghan PG. Precision, reliability, and responsiveness of a novel automated quantification tool for cartilage thickness: data from the osteoarthritis initiative. J Rheumatol 2020;47: 282–9.