
Article:
Bowes, MA, Guillard, GA, Vincent, GR et al. (2020) Precision, Reliability, and Responsiveness of a Novel Automated Quantification Tool for Cartilage Thickness: Data from the Osteoarthritis Initiative. The Journal of Rheumatology, 47 (2). pp. 282-289. ISSN: 0315-162X

https://doi.org/10.3899/jrheum.180541

Precision, Reliability and Responsiveness of a Novel Automated Quantification Tool for Cartilage Thickness: Data from the Osteoarthritis Initiative

Michael Antony Bowes, Gwenael Alain Guillard, Graham Richard Vincent, Alan Donald Brett, Christopher Brian Hartley Wolstenholme and Philip Gerard Conaghan

DOI: 10.3899/jrheum.180541
http://www.jrheum.org/content/early/2019/04/09/jrheum.180541

Title: **Precision, Reliability and Responsiveness of a Novel Automated Quantification Tool for Cartilage Thickness: Data from the Osteoarthritis Initiative**

Author Full Names: Michael Antony Bowes, 0000-0001-5774-847; Gwenael Alain Guillard, 0000-0002-2468-5652, Graham Richard Vincent; 0000-0001-6927-0138 ; Alan Donald Brett, 0000-0002-1671-9277 ; Christopher Brian Hartley Wolstenholme, 0000-0002-0488-6510; Philip Gerard Conaghan, 0000-0002-3478-5665

Key Indexing Terms: Knee osteoarthritis cartilage MRI

Affiliations: [1]Imorphics Ltd, Manchester, UK [2] Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds and NIHR Leeds Biomedical Research Centre, Leeds, UK

Source of support: Internal funding. Scientific and financial support for the FNIH OA Biomarkers Consortium and the study are made possible through grants and direct contributions provided by AbbVie; Amgen; Arthritis Foundation; Bioiberica S.A; DePuy Mitek; Flexion Therapeutics; GlaxoSmithKline; Merck Serono; Rottapharm Madaus; Sanofi; and Stryker; The Pivotal OAI MRI Analyses (POMA) Study, NIH HHSN2682010000. The OAI is a public–private partnership comprised of five contracts (N01-AR-2-2258;N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health. Funding partners include Merck Research Laboratories;Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer. Private sector funding for the Consortium and OAI is managed by the FNIH.

Conflict of interest: PGC has undertaken consultancies or speakers bureaus for Abbvie, Flexion, Galapagos, GSK, Medivir, Novartis, Pfizer, Samumed, Servier, TissueGene, MAB, GG, GRV, CBW, AB are employees of Imorphics Ltd, a wholly owned subsidiary of Stryker Corp

Authors: [1]Bowes M.A PhD, [1]Guillard G. PhD, [1]Vincent G.R. PhD, [1]Brett, A. PhD, [1]Wolstenholme, C.B. BSc, [2]Conaghan, P.G. MBBS, PhD, FRACP, FRCP

Address for correspondence:

Philip G Conaghan

Leeds Institute of Rheumatic and Musculoskeletal Medicine,

Chapel Allerton Hospital,

Chapeltown Rd,

Leeds LS7 4SA, UK

Phone: +44 113 3924884

Fax: +44 113 3924991

Email: p.conaghan@leeds.ac.uk

Running head: Automated cartilage segmentation

Abstract

Objective: Accurate automated segmentation of cartilage should provide rapid reliable outcomes for both epidemiological studies and clinical trials. We aimed to assess the precision and responsiveness of cartilage thickness measured with careful manual segmentation or a novel automated technique.

Methods:

Agreement of automated segmentation was assessed against two manual segmentation datasets: 379 MR images manually segmented in-house (Training set), and 582 from the OAI with data available at 0, 1, and 2 years (Biomarkers set). Agreement of mean thickness was assessed using Bland-Altman plots, change with pairwise Students t-test, in the central medial femur and tibia regions (cMF, cMT). Repeatability was assessed on a set of 19 knees imaged twice on the same day. Responsiveness was assessed using standardised response means (SRMs).

Results:

Agreement of manual vs automated methods was excellent with no meaningful systematic bias (Training set cMF bias 0.1mm 95%CI ±0.35, Biomarkers set bias 0.1mm ±0.4). The smallest detectable difference (SDD) for cMF was 0.13mm, coefficient of variation (CoV) 3.1%; cMT 0.16 mm, 2.65%. Reported change using manual segmentations in the cMF region at 1 year was -0.031mm, confidence limit (-0.022, -0.039), $p<10^{-4}$, SRM -0.31 (-0.23,-0.38); at 2 years was -0.071 (-0.058,-0.085), $p<10^{-4}$, SRM -0.43(-0.36,-0.49). Reported change using automated segmentations in the cMF at 1 year was -0.059 (-0.047, -0.071), $p<10^{-4}$, SRM -0.41(-0.34,-0.48) ; 2 years: -0.14 (-0.123,-0.157), $p<10^{-4}$, SRM -0.67 (-0.6,-0.72).

Conclusion: A novel cartilage segmentation method provides highly accurate and repeatable measures with comparable cartilage thickness measurements to careful manual segmentation, but with improved responsiveness.

Introduction

Cartilage is a key tissue of interest in structure-modification trials of osteoarthritis (OA). Although radiographic joint space width, a surrogate for cartilage loss, is the regulatory endpoint in these trials there is increasing evidence of the benefits of direct measures of cartilage morphology using magnetic resonance imaging (MRI)(1).

Techniques employing manual segmentation of cartilage have been explored with respect to a number of morphological characteristics, including volume and thickness, and extensively validated, including construct validity against radiographic joint space width, predictive and concurrent validity, and clinical outcomes (2-5). MRI cartilage thickness measures are associated with OA progression and joint replacement, and provide more responsive measures of progression than radiographic joint space narrowing (JSN) (5-7)

However, manual segmentation of cartilage morphology is time-consuming, tedious and challenging as careful attention must be paid to detecting the eroding outer margin of the cartilage. It therefore takes considerable time (hours) to carefully segment a single MR image, being composed in this case of 160 slices, limiting the utility of the method in analysing large datasets such as the Osteoarthritis Initiative (OAI), which includes data from over 9,000 knees at multiple time points. Additionally, the average amount of cartilage lost on each bone in the medial tibiofemoral joint of an OA knee is very small, typically around 50 – 100 microns per annum. This equates to a change of around 1/5 to 1/10 of a pixel in a typical MR image. To improve the speed of segmentation, some techniques for analysis have incorporated varying degrees of user input into semi-automated cartilage assessment(8).

Fully automated segmentation is desirable but the reliability and responsiveness of any such methods need to be established in a method that does not rely upon any user interaction. Fully

automated methods based on active appearance modelling (AAM) have demonstrated good measurement accuracy for a number of MRI-assessed tissues including knee cartilage, bone area and bone shape (9, 10) The addition of supervised machine learning to the AAM methodology offers potential enhancement in terms of improved voxel classification resulting in improved accuracy and responsiveness. A previous exercise used a preliminary version of this technology (10) but utilised a training set that had relatively crude manual segmentation, was not widely reflective of an OA population, used different MRI sequences to those in this study (making it impossible to run the older technology on the new dataset), and contained no longitudinal data.

In this study, we examined the performance metrics of a novel extension of AAM technology which incorporated a final refinement stage using supervised machine learning (AQ-CART). We assessed mean cartilage thickness in the anatomical locations which are commonly used in OA studies; we examined the accuracy and reliability of the method, agreement with careful manual segmentation and relative responsiveness.

Method

A number of comparisons were used in this study. For convenience, a summary of the datasets used, and the analyses performed are provided in Table 1.

Patients and Imaging

Image selection

A training set of 379 patient single-knee MRI images (the "Training" setwere used as input data for the supervised machine learning step of AQ-CART. These were selected to represent the entire range of radiographic OA structural severity, including medial compartment Kellgren-Lawrence grades 0-4, lateral compartment OA, together with young healthy knees which tend to have thicker cartilage. 287 images were acquired using a 3D double-echo-in-steady-state sequence (DESS-we) from the OAI (voxel size 0.3 x 0.3 x 0.7mm), but were not members of the Biomarkers set. 92 images were

acquired using a Philips 3D T2* weighted 3D gradient-echo sequence with water excitation (voxel size 0.3 x 0.3 x 1.5mm). The AAM training set has been described previously (11, 12).

Repeatability was performed on the Repeatability image set, a group of 19 subjects with and without radiographic OA that had test-retest single knee images acquired as a pilot study for the OAI (13).

For agreement and responsiveness, we used patient datasets from the OA Biomarkers Consortium FNIH sub-study of the OAI (https://oai.epi-ucsf.org/datarelease/FNIH.asp). Of 600 patients in the study, 582 patient datasets had manual cartilage measurements (Biomarkers image set) recorded at baseline, 1 and 2 years, resulting in sub-groups of 196 non-progressors and 386 progressors for either pain or structure or both, according to the FNIH subgroups. All images employed in these analyses used the Dual Echo Steady-State (DESS) MRI sequence: Additional parameters of the full OAI pulse sequence protocol and sequence parameters have been published in detail (14).

Ethics Approval

The OAI study received ethical approval from the UCSF OAI Coordinating Center IRB number 10-00532, reference 210064, Federalwide Assurance #00000068, and the OAI Clinical Sites Single IRB of Record was for study number 2017H0487, Federalwide Assurance #00006378. All patients provide informed consent to the OAI. Some of the Training set were collected under a study approved by the ethics committee of Lund University (LU-535)

Selection of regions for comparison

A number of anatomical regions of cartilage were provided on the OAI website – for convenience we chose the regions usually considered the most responsive – the central medial femur (cMF) and central medial tibia (cMT) (15) (https://oai.epiucsf.org/datarelease/SASDocs/kMRI_FNIH_QCart_Chondrometrics_Descrip.pdf). The mean thickness measure (ThCtAB) from each region was compared with the mean thickness from the automated segmentation. For automated segmentation, regions were selected on the mean

shape model to match the anatomical definition used for the manual method (Figure 1A). For reference the variable names of the baseline cartilage measures for the manual method were V00BMFMTH (cMF.ThCtAB) and V00CMTMTH (cMT.ThCtAB).

Manual segmentation method – Biomarkers dataset

Cartilage thickness was measured in the Biomarkers image set, using manual segmentation of the femorotibial cartilage surfaces by experienced segmenters, and reviewed by an expert as has been described previously ((16, 17),Chondrometrics GmBH).

Manual segmentation and surface building – Training dataset

For the supervised learning algorithm training set, cartilage was manually segmented by experienced segmenters, using Imorphics EndPoint software (Imorphics, Manchester, UK) using the Training image set. 3D surfaces were generated from the cartilage contours in each image slice using a marching cubes algorithm, followed by geometric smoothing.

AQ-CART method

Each image was automatically segmented using 3D AAMs of bone and cartilage using a multi-start optimisation. Active appearance models are widely used in medical imaging, and fit the shape and grey-level variations of a training set to a 3D image, and are capable of rapid and accurate 3D segmentation, with sub-voxel accuracy (18). Initially, this fits low-density low-resolution deformable models but ends in a robust matching of detailed high-resolution models. Finally, in a novel step, the voxels contained in the cartilage region are assigned with a non–linear regression function, based on a bootstrap aggregation, chosen using a probably approximately correct (PAC) learning method.

Cartilage thickness was measured using the Anatomically Corresponded Regional Analysis of Cartilage (ACRAC) (11, 19), which is summarised in Figure 1B. From each correspondence point on the 3D bone surface, which is the result of an AAM bone search, we measure the distance from the bone to the outer cartilage surface, along a line normal to the bone surface. In addition to providing

accurate and repeatable measurement, this process fits all examples with a consistent dense set of anatomical landmarks, which can be used to take a measurement at the same point across a population and between time points, correcting for both the size and shape of each bone.

Accuracy, reliability and comparative analyses

Accuracy of AQ-CART was determined using the Training image set, using leave-25%-out models. In this method, 4 models are built, each of which leaves out 25% of the training examples. Each image is then searched using the single model which does not contain itself as a training example. This means that each image is searched using an unbiased model.

ACRAC cartilage thickness maps (Figure 1C) were then prepared for both manual and automated segmentations and used to calculate the mean thickness within each region. Correlation and agreement of the mean thickness measure was assessed using least-squares linear fits and Bland-Altman plots.

Repeatability of AQ-CART was assessed on the Repeatability set, using the smallest detectable difference (SDD) defined as the 95% confidence interval (CI) on the Bland-Altman plot, and the coefficient of variation (CoV) using the root-mean-square method.

Agreement of the mean thickness reported by the manual and automated segmentation methods using the baseline images of the Biomarkers image set was assessed using Bland-Altman plots. We then compared change from baseline of both methods using pairwise student t-tests of mean thickness of the central medial femur and tibia (cMF and cMT) in the 582 knees. Agreement of 2-year change from baseline, as reported by the manual and automated segmentation methods, was assessed using a Bland Altman plot. Responsiveness was assessed using standardised response means (SRMs). Confidence limits for the SRMs were calculated using a bootstrap method (MedCalc Software, Ostend, Belgium). Results were calculated separately for the 4 FNIH Biomarkers

subgroups, which were JSN progressors, pain progressors, combined JSN and pain progressors, and non-progressors [5]

Results

Correlation and agreement mean cartilage thickness using the Training set.

Correlation of the mean thickness reported by the manual and automated methods was $r^2 = 0.97$ for the cMF region, and 0.84 for the cMT. The equation for the linear least squares fit between the manual and automated methods for the cMF region was y = 0.81x + 0.44; for the cMT region was y = 0.81x + 0.35 (Figure 2, top row). The automated segmentation had a small tendency to under-segment thicker cartilage and over-segment thinner cartilage, when compared with the Training set. Systematic bias for the cMF region was 0.098 mm, 95% limits of agreement were 0.354 mm; for the cMT region bias was -0.026 and 95% limits of agreement were 0.420 (Figure 2, bottom row).

Repeatability

The smallest detectable difference (SDD) in the Repeatability image set for the cMF region was 0.13mm, coefficient of variation (CoV) 3.1%; for the cMT region the SDD was 0.16 mm, CoV 2.65% (Bland Altman plot not shown)

Agreement between baseline manual segmentations (Biomarkers set)

Systematic bias of the mean thickness reported by the manual and automated methods for the cMF region at baseline was +0.09mm, 95% confidence limits were ±0.35mm; for the cMT region bias was -0.2mm, 95% confidence limits were ±0.39mm (Figure 3)

Agreement of 2-year change (Biomarkers set)

In the Biomarkers set of 582 knees, the reported change in mean thickness measured with automated segmentation was around twice that reported by that with manual segmentation. SRM values were also higher for the automated method. For example, change in **manual** cMF at 1 year was 0.03mm, 95% confidence limit (-0.02, 0.09), SRM 0.34 (-0.38), at 2 years this

was -0.071 (-0.058, -0.085), $p<10^{-4}$, SRM -0.43 (-0.36,-0.49). Change in **automated** cMF at 1 years was -0.059 (-0.047, -0.071), $p<10^{-4}$, SRM -0.41 (-0.34,-0.48); at 2 years change was- 0.14 (-0.123, -0.157), $p<10^{-4}$, SRM -0.67 (-0.6,-0.72) (Figure 4).

The detection of greater change with the automated method was consistent in both of the joint space narrowing subgroups (Figure 1); additionally, the automated method detected a significant change in cartilage thickness at both timepoints in those who did not progress with either pain or joint space narrowing. Neither method showed any change in those who progressed only in their pain score. Detailed results for all subgroups of femur and tibia using manual and automated methods is shown in Table 2

Time required for automatic analysis

Automatic segmentation of a single image, using a single CPU core of a PC, took on average 45 seconds, and calculation of cartilage thickness and volume required a further 7 seconds.

Discussion

The novel automated segmentation technique reported here demonstrated excellent accuracy and reliability in assessing cartilage thickness in the medial tibiofemoral joint, the most commonly used region assessed in clinical trials. There was also excellent agreement with both cross-sectional measurement and longitudinal change in cartilage thickness when compared with a well-established manual segmentation method.

The agreement of automated segmentation measurements using the Training set was excellent with no meaningful systematic bias. The automated segmentation had a small tendency to under-segment the thickest cartilage, and to over-segment denuded cartilage when compared with the Training set. . In the central medial femur, cartilage with mean thickness of 3mm (approximately the 95th percentile of cartilage thickness distribution in the training set) would be under-segmented by 0.25mm, or about half of the average length of a voxel edge. Completely denuded cartilage (mean

thickness of 0mm) would be over-segmented by 0.44m. Repeatability of the automated method (SDD of around 0.14mm, and CoV of 2.5 and 3.1%) was excellent, and comparable with values reported for manual segmentation methods (11, 13)

When comparing automated segmentation with the careful manual segmentation method of another group in the Biomarkers dataset, the automated method reported a slightly thicker average measure than the manual method of about 0.1mm. This small difference is not particularly surprising for a few reasons: the 2 measures are calculated in very different ways; the regions to be measured were prepared independently; and the manual segmentation of the automated training set and manual set were also prepared independently. However, despite these differences in methodology, the agreement between the two methods was excellent, as illustrated by the Bland Altman plot.

The correlation of longitudinal change in the femur and tibia for the Biomarkers set was excellent, although the correlation of tibia measures was lower (0.87 vs 0.95 for the femur). We cannot be certain of why the tibia has a lower correlation; as noted above, the methodologies are different, and both correlation coefficients are acceptable.

We did not perform a correlation of the individual longitudinal changes, as these would not be expected to correlate, given the amount of change found here, and the reported measurement errors of the methods. Given 2 methods, with measurement SD of 0.075 mm (approximately the SD for the two methods, and a test set which contains changes of between 0 and 0.15mm (the approximate range of annual changes found here), the correlation of the 2 methods will be very low (less than 0.02) assuming perfect agreement between the methods. Any single measurement will contain the actual change, plus a normally-distributed error ranging from -0.14 mm to +0.14mm (the 95th percentile, or 1.96 x SD). Most of the differences found are dominated by noise, and do not reflect true change. In a larger group, these differences in noise cancel each other out.

Automated segmentation of tissues which change by small fractional amounts are often insensitive to any such change; such methods are often repeatable because of regression to the mean during the automated search. This causes potential over-segmentation of thin cartilage, and under-segmentation of thick cartilage. However automated segmentation with AQ-CART was at least as sensitive to change as careful manual segmentation, and this responsiveness was seen across the clinical progression subgroups. Additionally, the "Non-Progressor" group demonstrated significant cartilage thickness loss at both 1 and 2 years with the automated method, whereas no change was measured using the manual method.

The improved responsiveness was a consequence of the automated method identifying about twice as much change (in the femur), with similar levels of measurement noise. A typical amount of average cartilage thickness loss is tiny, much less than one voxel width in a year. This means that cartilage loss is fundamentally a change in what becomes a partial volume in an MR image sampling voxel at the outer edge of the cartilage. Human measurement is not capable of dealing with these partial volumes and It is likely that a human reader at a standard computer display cannot adequately resolve such differences in partial volume, whereas an algorithm can. All measurement methods contain errors, and there is no "ground truth" in this study, such as an independent measure of cartilage thickness using more accurate methods; it is not possible to be certain that improved responsiveness is certainly caused by cartilage changing by an additional 50 microns per year.

The short time required for analysis of an image (52 seconds), compared with the preparation of a manual segmentation (typically around 4 hours for our in-house segmenters), allows for the segmentation of large numbers of images. In actuality, this time is shorter; 52 seconds are required for a single CPU core of a PC; however a typical desktop machine can run 8 threads simultaneously, reducing the average time for a single segmentation to around 10 seconds per image, with no requirement for user input.

A potential limitation of this work was that the models were trained and tested on on 2 particular MRI sequences, and these were obtained using the same manufacturers and models of MRI machines, from an observational study in which image quality was tightly controlled. The accuracy, repeatability and responsiveness of these models may not provide the same results when using other MR imaging sequences.

In summary, application of a novel AAM-based cartilage segmentation incorporating a supervised machine learning step provided highly accurate and repeatable measurement of cartilage thickness with excellent agreement with careful manual segmentation, but with improved responsiveness.

Acknowledgements

PGC is supported in part by the National Institute for Health Research (NIHR) infrastructure at Leeds. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

The OAI is a public-private partnership comprising five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partner

References

1. Hunter DJ, Zhang W, Conaghan PG, Hirko K, Menashe L, Reichmann WM, et al. Responsiveness and reliability of mri in knee osteoarthritis: A meta-analysis of published evidence. Osteoarthritis Cartilage 2011;19:589-605.

2. Hunter DJ, Arden N, Conaghan PG, Eckstein F, Gold G, Grainger A, et al. Definition of osteoarthritis on mri: Results of a delphi exercise. Osteoarthritis Cartilage 2011;19:963-9.

3. 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.

4. Wirth W, Duryea J, Hellio Le Graverand MP, John MR, Nevitt M, Buck RJ, et al. Direct comparison of fixed flexion, radiography and mri in knee osteoarthritis: Responsiveness data from the osteoarthritis initiative. Osteoarthritis Cartilage 2013;21:117-25.

5. Eckstein F, Kwoh CK, Boudreau RM, Wang Z, Hannon MJ, Cotofana S, et al. Quantitative mri measures of cartilage predict knee replacement: A case-control study from the osteoarthritis initiative. Ann Rheum Dis 2013;72:707-14.

6. Eckstein F, Collins JE, Nevitt MC, Lynch JA, Kraus VB, Katz JN, et al. Brief report: 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. Cicuttini FM, Jones G, Forbes A, Wluka AE. Rate of cartilage loss at two years predicts subsequent total knee arthroplasty: A prospective study. Ann Rheum Dis 2004;63:1124-7.

8. Duryea J, Iranpour-Boroujeni T, Collins JE, Vanwynngaarden C, Guermazi A, Katz JN, et al. Local area cartilage segmentation: A semiautomated novel method of measuring cartilage loss in knee osteoarthritis. Arthritis Care Res 2014;66:1560-5.

9. Williams TG, Holmes AP, Bowes M, Vincent G, Hutchinson CE, Waterton JC, et al. Measurement and visualisation of focal cartilage thickness change by mri in a study of knee osteoarthritis using a novel image analysis tool. Br J Radiol 2010;83:940-8.

10. Vincent G, Wolstenholme C, Scott I, Bowes M. Fully automatic segmentation of the knee joint using active appearance models. Proceedings of MICCAI Workshop on Medical Image Analysis for the Clinic 2010:224-30.

11. Hunter DJ, Bowes MA, Eaton CB, Holmes AP, Mann H, Kwoh CK, et al. Can cartilage loss be detected in knee osteoarthritis (OA) patients with 3-6 months' observation using advanced image analysis of 3t mri? Osteoarthritis Cartilage 2010;18:677-83.

12. Bowes MA, Vincent GR, Wolstenholme CB, Conaghan PG. A novel method for bone area measurement provides new insights into osteoarthritis and its progression. Ann Rheum Dis 2015;74:519-25.

13. Eckstein F, Hudelmaier M, Wirth W, Kiefer B, Jackson R, Yu J, et al. Double echo steady state magnetic resonance imaging of knee articular cartilage at 3 tesla: A pilot study for the osteoarthritis initiative. Ann Rheum Dis 2006;65:433-41.

14. Peterfy CG, Schneider E, Nevitt M. The osteoarthritis initiative: Report on the design rationale for the magnetic resonance imaging protocol for the knee. Osteoarthritis Cartilage 2008;16:1433-41.

15. Eckstein F, Guermazi A, Gold G, Duryea J, Hellio Le Graverand MP, Wirth W, et al. Imaging of cartilage and bone: Promises and pitfalls in clinical trials of osteoarthritis. Osteoarthritis Cartilage 2014;22:1516-32.

16. Eckstein F, Maschek S, Wirth W, Hudelmaier M, Hitzl W, Wyman B, et al. One year change of knee cartilage morphology in the first release of participants from the osteoarthritis initiative progression subcohort: Association with sex, body mass index, symptoms and radiographic osteoarthritis status. Ann Rheum Dis 2009;68:674-9.

17. Wirth W, Hellio Le Graverand MP, Wyman BT, Maschek S, Hudelmaier M, Hitzl W, et al. Regional analysis of femorotibial cartilage loss in a subsample from the osteoarthritis initiative progression subcohort. Osteoarthritis Cartilage 2009;17:291-7.

18. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Trans Patt Anal Mach Intell 2001;23:681-5.

19. Williams TG, Holmes AP, Waterton JC, Maciewicz RA, Hutchinson CE, Moots RJ, et al. Anatomically corresponded regional analysis of cartilage in asymptomatic and osteoarthritic knees by statistical shape modelling of the bone. IEEE Trans Med Imaging 2010;29:1541-59.

Figure and Table Legends

Figure 1: Measurement Methodology

Figure (A) shows the selected regions of the central medial femur (cMF, top) and the central medial tibia (cMT, bottom). Each correspondence point within the shape model is shown as a red sphere on the surface of the mean bone shapes; there are 1527 correspondence points in the cMF region, and 828 in the cMT regions. Figure (B) schematically shows the method by which cartilage thickness is measured using the Anatomically Corresponded Regional Analysis of Cartilage (ACRAC) method. From each correspondence point the distance along a line normal to the surface, and the distance from the bone to the outer cartilage surface is recorded (note normals are shown schematically, all in the same direction – in practise normal direction varies slightly with the curvature of the bone surface). Figure (C) shows typical examples of cartilage thickness in the femur of a healthy knee (left), and an OA knee (right). Note that the OA knee is denuded in part of the cMF region (dotted green)

Figure 2: Correlation and agreement of mean thickness in the Training set

Top graphs show a scatter plot of mean thickness values, comparing reported mean thickness values for manual and automated segmentations in the Training set, using miss-25%-out models, for the cMF region (left) and cMT region (right), together with the results of a linear fit, plus the r-squared value for the correlation of the datasets. The same data is displayed in the lower graphs using a Bland Altman plot to assess agreement; bias is shown with a thickly dashed line, and the 95th percentile confidence limits are shown using a dotted line

Figure 3: Agreement of mean cartilage thickness in the Biomarkers set

Systematic bias is shown with a thickly dashed line, and the 95th percentile confidence limits are shown using a dotted line for the central medial femur (left) and central medial tibia (right)

Figure 4: Graphical representation of 2-year change in central medial femur region by FNIH Biomarkers Subgroup

Results are shown for all 582 knees ("All"), together with the 4 subgroups; joint space narrowing progressors ("JSN Only Progressor", n=102), both joint space narrowing and pain progressors ("JSN and Pain Progressors", n=183), pain progressors ("Pain Only Progressors", n=101), and non-progressors ("No JSN or Pain Progression", n=196). Further detail is provided in Table 2, along with results for the central medial tibia. Error bars represent 95% confidence intervals.

Table 1: Datasets and analysis methods used in this study

Key to 4 subgroups; joint space narrowing progressors ("JSN Only Progressors"), both joint space narrowing and pain progressors ("JSN and Pain Progressors"), pain progressors ("Pain Only Progressors"), and non-progressors ("No JSN or Pain Progressors").

Table 2: Comparison of 1-year and 2-year change in cartilage thickness from baseline in the Biomarkers set

Results are shown for all 582 knees ("All"), together with the 4 subgroups; joint space narrowing progressors ("JSN Only Progressor", n=102), both joint space narrowing and pain progressors ("JSN and Pain

n=183), pain progressors ("Pain", n=101), and non-progressors ("Non-Progressors", n=196). SRM 95% confidence limits were estimated using a bootstrap method.

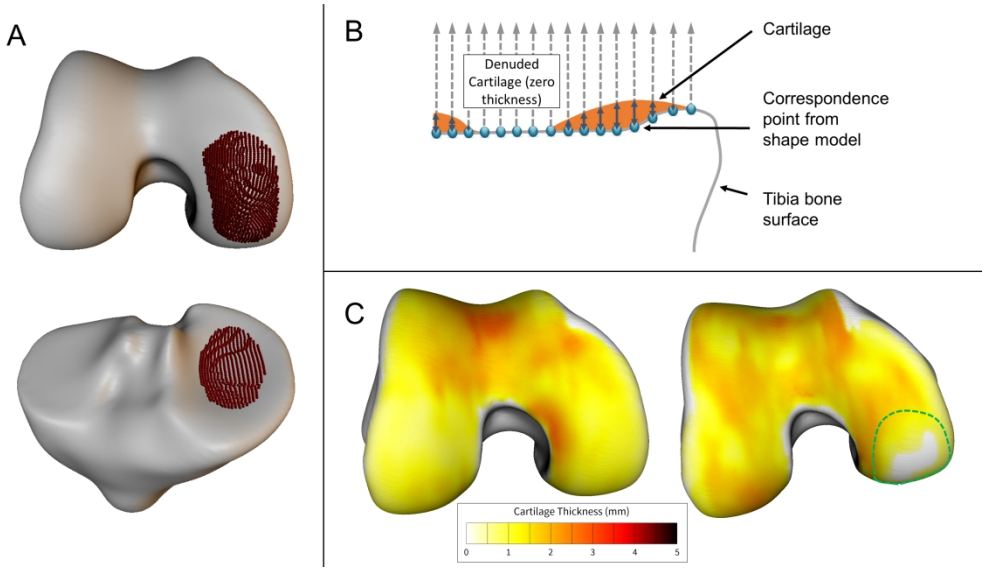

Figure 1: Measurement Methodology

Figure A shows the selected regions of the central medial femur (cMF, top) and the central medial tibia (cMT, bottom). Each correspondence point within the shape model is shown as a red sphere on the surface of the mean bone shapes there are 1527 correspondence points in the cMF region, and 828 in the cMT regions

(B) Schematically shows the method by which cartilage thickness is measured using the Anatomically Corresponded Regional Analysis (ACRAC) method.. From each correspondence point the distance along a line normal to the surface, and the distance from the bone to the outer cartilage surface is recorded (note normals are shown schematically, all in the same direction – in practise normal direction varies slightly with the curvature of the bone surface)

(C) Shows typical examples of cartilage thickness in the femur of a healthy knee (left), and an OA knee (right). Note that the OA knee is denuded in part of the cMF region

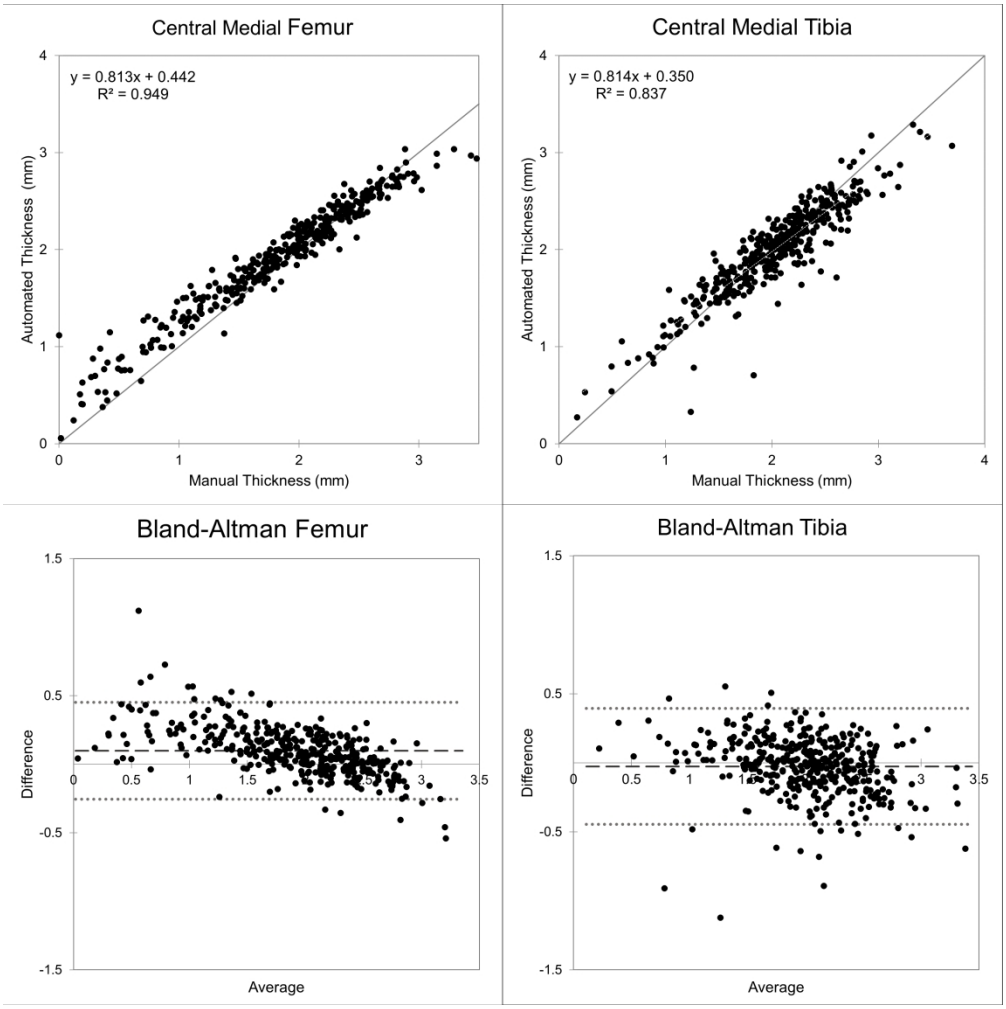

Figure 2: Correlation and agreement of mean thickness in the Reference set
Top graphs show a scatter plot of mean thickness values, comparing reported mean thickness values for manual and automated segmentations in the Reference set, using miss-25%-out models, for the cMF region (left) and cMT region (right), together with the results of a linear fit, plus the r-squared value for the correlation of the datasets. The same data is displayed in the lower graphs using a Bland Altman plot to assess agreement; bias is shown with a thickly dashed line, and the 95th percentile confidence limits are shown using a dotted line

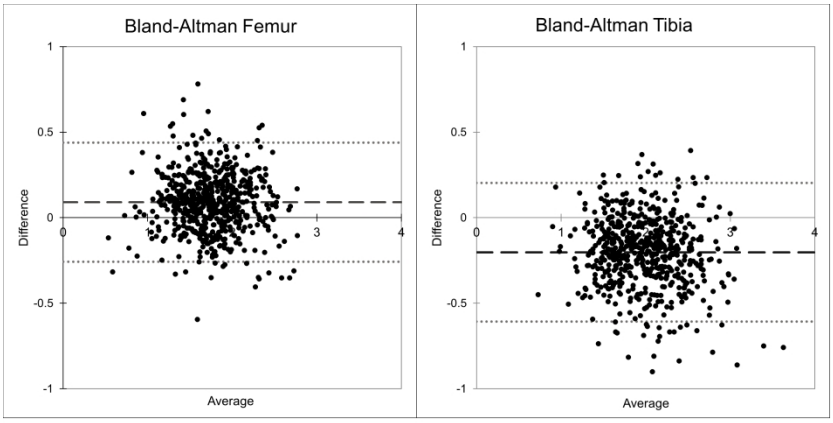

Figure 3: Agreement of mean cartilage thickness in the Biomarkers set
Systematic bias is shown with a thickly dashed line, and the 95th percentile confidence limits are shown using a dotted line for the central medial femur (left) and central medial tibia (right)

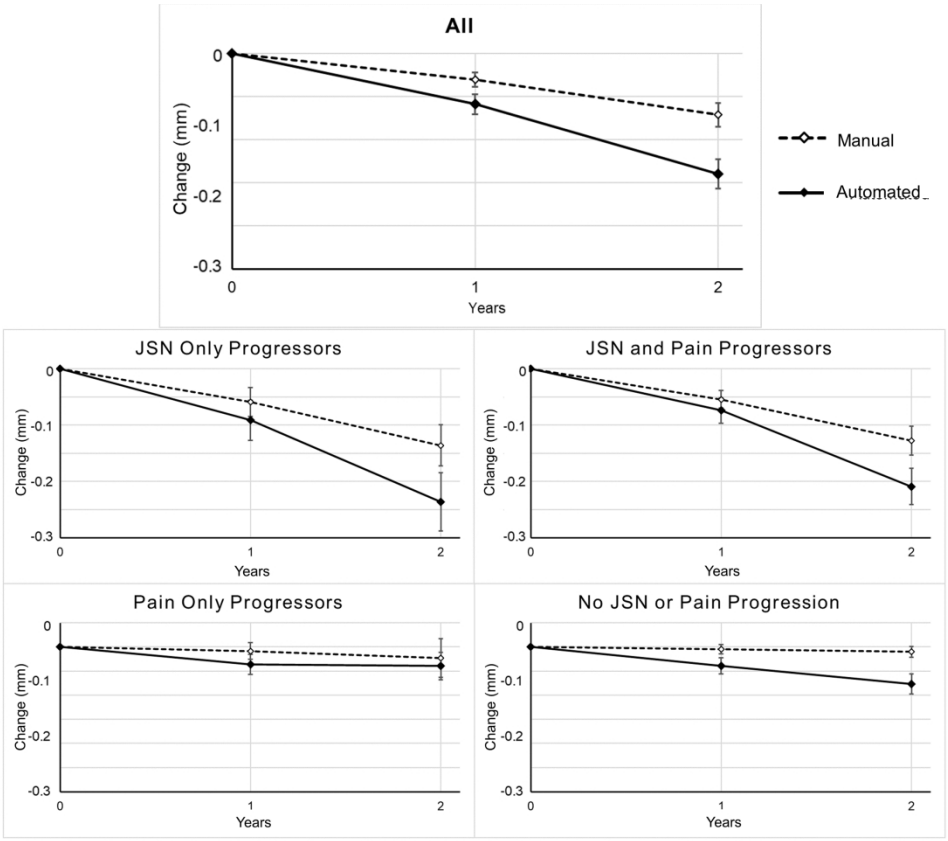

Figure 4: Graphical representation of 2-year change in central medial femur region by FNIH Biomarkers Subgroup

Results are shown for all 582 knees ("All"), together with the 4 subgroups; joint space narrowing progressors ("JSN Only Progressor", n=102), both joint space narrowing and pain progressors ("JSN and Pain Progressors", n=183), pain progressors ("Pain Only Progressors",n=101), and non-progressors ("No JSN or Pain Progression", n=196). Further detail is provided in Table 2, along with results for the central medial tibia.

Table 1: Datasets and analysis methods used in this study

Image Dataset	Dataset	Segmentation of Cartilage Surfaces	Calculation of Cartilage Thickness	Used For
Training	379 segmentations of femur and tibial cartilage at a single time point, on fat-saturated 3D MR images *Range of radiographic OA structural severity, including medial compartment Kellgren-Lawrence grades 0-4, lateral compartment OA, plus young healthy knees*	Manual segmentation using EndPoint (Imorphics), supervised by experienced segmenter (1)	Anatomically Corresponded Regional Analysis of Cartilage Thickness (ACRAC). Thickness is measured at multiple points along normals from the bone surface (Figure 1B,(2))	Training set for supervised machine learning step in AQ-CART Correlation and agreement of mean cartilage thickness in cMF and cMT regions, automated or manual segmentations, miss-25%-out models
Repeatability	19 test-retest images of knees with and without radiographic OA - pilot study for the OAI (3)	n/a	ACRAC	Repeatability of automated segmentation
Biomarkers	582 segmentations of femur and tibial cartilage at baseline, 1 and 2 years *JSN Only Progressors, n=102* *JSN and Pain Progressors, n=183)* *Pain Only Progressors, n=101)* *No JSN or Pain Progression, n=196)*	Manual segmentation by Chondrometrics, supervised by experienced segmenter (4, 5)	Volume of cartilage divided by region of bone ((5))	Cross-sectional agreement of mean cartilage thickness in cMF and cMT regions, using automated or manual segmentation, baseline images only Longitudinal agreement of change in mean cartilage thickness from baseline in the same regions, using automated or manual segmentation Responsiveness of automated and manual segmentation in the same regions using pairwise Student's t-test and SRM

Key to 4 subgroups; joint space narrowing progressors ("JSN Only Progressors"), both joint space narrowing and pain progressors ("JSN and Pain Progressors"), pain progressors ("Pain Only Progressors"), and non-progressors ("No JSN or Pain Progressors").

Table 2: Comparison of 1-year and 2-year change in cartilage thickness from baseline in the Biomarkers set

FNIH Biomarkers Group		1—year change from baseline			2-year change from baseline		
		Mean Change [95% CL]	SRM [95% CL]	p-value	Mean Change [95% CL]	SRM [95% CL]	p-value
Femur Change (cMF region)							
All	**Manual**	-0.031 [-0.022,-0.039]	-0.31 [-0.23,-0.38]	5.797E-13	-0.071 [-0.058,-0.085]	-0.43 [-0.36,-0.49]	9.72E-23
All	**Automated**	-0.059 [-0.047,-0.071]	-0.41 [-0.34,-0.48]	8.330E-22	-0.14 [-0.123,-0.157]	-0.67 [-0.6,-0.72]	2.54E-48
JSN	**Manual**	-0.059 [-0.033,-0.084]	-0.45 [-0.28,-0.6]	1.320E-05	-0.136 [-0.099,-0.173]	-0.74 [-0.59,-0.89]	4.07E-11
JSN	**Automated**	-0.092 [-0.056,-0.128]	-0.5 [-0.32,-0.67]	1.620E-06	-0.236 [-0.184,-0.288]	-0.9 [-0.73,-1.05]	9.10E-15
JSN and Pain	**Manual**	-0.055 [-0.039,-0.07]	-0.5 [-0.34,-0.63]	1.228E-10	-0.128 [-0.102,-0.154]	-0.73 [-0.62,-0.83]	9.72E-23
JSN and Pain	**Automated**	-0.074 [-0.052,-0.097]	-0.48 [-0.36,0.6]	8.330E-22	-0.209 [-0.177,-0.241]	-0.96 [-0.82,-1.09]	2.54E-48
Pain	**Manual**	-0.008 [0.009,-0.026]	-0.1 [0.1,-0.28]	3.398E-01	-0.023 [0.017,-0.063]	-0.12 [0.08,-0.24]	2.63E-01
Pain	**Automated**	-0.036 [-0.016,-0.057]	-0.35 [-0.16,-0.52]	6.703E-04	-0.04 [-0.012,-0.068]	-0.28 [-0.06,-0.43]	5.97E-03
Non-Progressors	**Manual**	-0.005 [0.004,-0.014]	-0.07 [0.07,-0.21]	3.057E-01	-0.01 [0.001,-0.021]	-0.13 [0.01,-0.27]	7.98E-02
Non-Progressors	**Automated**	-0.039 [-0.023,-0.056]	-0.33 [-0.2,-0.45]	6.924E-06	-0.077 [-0.056,-0.098]	-0.52 [-0.4,-0.62]	1.14E-11
Tibia Change (cMF region)							
All	**Manual**	-0.036 [-0.026,-0.045]	-0.3 [-0.23,-0.38]	2.264E-12	-0.073 [-0.059,-0.086]	-0.43 [-0.35,-0.49]	1.14E-22
All	**Automated**	-0.055 [-0.043,-0.067]	-0.39 [-0.31,-0.45]	1.829E-19	-0.114 [-0.097,-0.131]	-0.55 [-0.48,-0.61]	3.21E-35
JSN	**Manual**	-0.057 [-0.03,-0.084]	-0.42 [-0.22,-0.6]	4.223E-05	-0.117 [-0.083,-0.15]	-0.7 [-0.52,-0.85]	3.17E-10
JSN	**Automated**	-0.08 [-0.05,-0.11]	-0.52 [-0.33,-0.72]	7.201E-07	-0.179 [-0.132,-0.225]	-0.76 [-0.58,-0.91]	1.43E-11
JSN and Pain	**Manual**	-0.05 [-0.03,-0.07]	-0.37 [-0.23,-0.49]	1.287E-06	-0.117 [-0.088,-0.146]	-0.6 [-0.47,-0.7]	1.14E-22
JSN and Pain	**Automated**	-0.068 [-0.043,-0.093]	-0.4 [-0.26,-0.51]	1.829E-19	-0.172 [-0.137,-0.207]	-0.72 [-0.6,-0.82]	3.21E-35
Pain	**Manual**	-0.025 [-0.006,-0.045]	-0.26 [-0.06,-0.45]	9.860E-03	-0.03 [0.009,-0.069]	-0.16 [0.07,-0.26]	1.23E-01
Pain	**Automated**	-0.037 [-0.018,-0.057]	-0.38 [-0.2,-0.54]	2.170E-04	-0.035 [-0.008,-0.062]	-0.26 [-0.06,-0.42]	1.05E-02
Non-Progressors	**Manual**	-0.016 [-0.002,-0.031]	-0.16 [-0.03,-0.31]	2.792E-02	-0.03 [-0.015,-0.045]	-0.29 [-0.15,-0.44]	9.56E-05
Non-Progressors	**Automated**	-0.039 [-0.022,-0.056]	-0.32 [-0.19,-0.44]	1.056E-05	-0.067 [-0.044,-0.089]	-0.42 [-0.3,-0.52]	2.01E-08

Results are shown for all 582 knees ("All"), together with the 4 subgroups; joint space narrowing progressors ("JSN", n=102), both joint space narrowing and pain progressors ("JSN and Pain", n=183), pain progressors ("Pain", n=101), and non-progressors ("Non-Progressors", n=196). SRM 95% confidence limits were estimated using a bootstrap method