UCSF
UC San Francisco Previously Published Works

Title

Towards Automatic Cartilage Quantification in Clinical Trials - Continuing from the 2019 IWOAI Knee Segmentation Challenge.

Permalink

https://escholarship.org/uc/item/4k55c7nc

Journal

Osteoarthritis Imaging, 3(1)

Authors

Dam, Erik
Desai, Arjun
Deniz, Cem
et al.

Publication Date

2023-03-01

DOI

10.1016/j.ostima.2023.100087

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

HHS Public Access
Author manuscript
Osteoarthr Imaging. Author manuscript; available in PMC 2024 July 19.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Published in final edited form as:
Osteoarthr Imaging. 2023 March ; 3(1): . doi:10.1016/j.ostima.2023.100087.

Towards Automatic Cartilage Quantification in Clinical Trials – Continuing from the 2019 IWOAI Knee Segmentation Challenge

Erik B Dam[1,*], **Arjun D Desai**[2], **Cem M Deniz**[3], **Haresh R Rajamohan**[4], **Ravinder Regatte**[3], **Claudia Iriondo**[5], **Valentina Pedoia**[5], **Sharmila Majumdar**[5], **Mathias Perslev**[1], **Christian Igel**[1], **Akshay Pai**[6], **Sibaji Gaj**[7], **Mingrui Yang**[7], **Kunio Nakamura**[7], **Xiaojuan Li**[7], **Hasan Maqbool**[8], **Ismail Irmakci**[9], **Sang-Eun Song**[8], **Ulas Bagci**[9], **Brian Hargreaves**[2], **Garry Gold**[2], **Akshay Chaudhari**[2]

[1]University of Copenhagen, Copenhagen, Denmark

[2]Stanford University, Stanford, CA USA

[3]New York University, Langone Health, New York, NY USA

[4]New York University, New York, NY USA

[5]University of California, San Francisco, CA USA

[6]Cerebriu A/S, Copenhagen, Denmark

[7]Cleveland Clinic, Cleveland, OH USA

[8]University of Central Florida, Orlando, FL USA

[9]Northwestern University, Evanston, IL USA

Abstract

Objective—To evaluate whether the deep learning (DL) segmentation methods from the six teams that participated in the IWOAI 2019 Knee Cartilage Segmentation Challenge are appropriate for quantifying cartilage loss in longitudinal clinical trials.

Design—We included 556 subjects from the Osteoarthritis Initiative study with manually read cartilage volume scores for the baseline and 1-year visits. The teams used their methods originally trained for the IWOAI 2019 challenge to segment the 1130 knee MRIs. These scans were anonymized and the teams were blinded to any subject or visit identifiers. Two teams also submitted updated methods. The resulting 9,040 segmentations are available online.

The segmentations included tibial, femoral, and patellar compartments. In post-processing, we extracted medial and lateral tibial compartments and geometrically defined central medial and lateral femoral sub-compartments. The primary study outcome was the sensitivity to measure cartilage loss as defined by the standardized response mean (SRM).

Results—For the tibial compartments, several of the DL segmentation methods had SRMs similar to the gold standard manual method. The highest DL SRM was for the lateral tibial compartment at 0.38 (the gold standard had 0.34). For the femoral compartments, the

[*]Corresponding author: erikdam@di.ku.dk.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

gold standard had higher SRMs than the automatic methods at 0.31/0.30 for medial/lateral compartments.

Conclusion—The lower SRMs for the DL methods in the femoral compartments at 0.2 were possibly due to the simple sub-compartment extraction done during post-processing. The study demonstrated that state-of-the-art DL segmentation methods may be used in standardized longitudinal single-scanner clinical trials for well-defined cartilage compartments.

Keywords

MRI; knee; deep learning; clinical trial; cartilage

Introduction

Over 60 years ago it was realized that radiographs could be used to stage osteoarthritis[1], and later it was demonstrated that magnetic resonance imaging (MRI) allowed cartilage quantification[2]. Computer-based quantification was introduced using semi-automatic methods[3] and then a fully automatic method in 2005[4]. Gradually, methods for quantification of cartilage, bone, meniscus, and synovium compartments have been proposed[5].

Currently, computer-based quantification of disease-related effects has proven effective for epidemiological research into OA pathogenesis using large cohort studies, recently targeting cartilage loss[6] and bone shape[7], and previously also cartilage composition[8], bone structure[9], and cartilage surface integrity[10]. Many of the recently proposed methods leverage advances in machine learning, particularly in deep learning (DL) [11].

However, these methods have limited impact on the patients and the clinicians. The reasons include:

- OA is a multi-faceted disease involving many tissues. Therefore, even if a specific grading like Kellgren Lawrence (KL) can be automated[12], this does not remove the need for expert radiologists.
- MRI is a complicated qualitative modality with no direct physical interpretation of the intensity levels like CT. MRI sequences visualizing clinically important information are implemented differently across scanner vendors and scanner models. This challenges development and validation of software methods.
- The deep learning methods are data-hungry during training. For a clinical setting where multiple, vaguely-defined outcomes involving several tissues need to be extracted from multiple MRI sequences acquired from any given scanner model, the required training data may currently be prohibitive.
- Even if many DL methods for knee MRI segmentation have been published, they have only to a limited degree been validated for biomarker quantification in large, longitudinal cohorts.

We investigated whether state-of-the-art computer-based methods are suitable for use in clinical trials. This could support treatment development by allowing cheaper clinical trials and potentially more sophisticated analysis of the treatment effect beyond volumetry (e.g.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

related to shape[7] or surface integrity[10]). Clinical trials have a more controlled environment than the general clinical setting. Specifically for MRI analysis, clinical trials will have pre-defined, standardized sequences. However, the acquisitions will be subject to inter-site differences and to changes over time as the scanners age and receive hardware replacements (e.g. coil or magnet) and software upgrades. Due to these potentially interacting changes and the complexity of the MRI acquisition and reconstruction process, it is challenging to predict the impact on the resulting MRI scans. These potential changes can be grouped into changes occurring gradually over time, *Drift*, and abrupt changes occurring due to a specific event, *Shift*.

A previous, preliminary study[13] demonstrated that even for a well-organized study like the Osteoarthritis Initiative (OAI)[14], there were substantial Drift and Shift effects, as illustrated in Figure 1. For the two illustrated sites, the mean scan intensities gradually increased, similarly by 1.7% and 1.8% per year. However, this Drift was interrupted by Shift events causing the mean intensity to abruptly increase or decrease. These Shifts caused large intensity changes up to 50%. It is unclear whether the Drift is related to overall scan effects or to more tissue-specific, non-linear effects. The OAI setup adheres to the OARSI recommendations[15] using the same Siemens 3T scanners on all MRI sites and the same set of predefined sequences. However, given the duration of the OAI, the setup is dynamic due to replacement of spare parts and scanner software updates, challenging the initial standardization.

The motivations for this study were:

- Clinical trials could potentially benefit from computer-based quantification.
- The 2019 IWOAI Knee Segmentation Challenge demonstrated that several methods had good segmentation accuracy performance[16]. However, the conclusions were unclear for assessing cartilage thickness changes in the small cohort.
- Most state-of-the-art knee MRI quantification methods are Deep Learning methods. Deep Learning methods have been criticized for being sensitive to changes in the input distributions[17], potentially requiring dedicated architecture design to address this.

The objectives were:

- To focus on the clinical trial use case where an MRI biomarker is used as efficacy biomarker for a chondroprotective treatment.
- To evaluate the robustness of state-of-the-art DL methods with respect to Drift and Shift effects.
- To provide a large collection of segmentation masks as publicly available data for future research.

We pursued this by evaluating the methods from the original 2019 IWOAI Knee Segmentation Challenge[16] on a large sub-cohort from the OAI.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Methods

Study Cohort

The OAI study[14] provides a large cohort with observations from multiple visits, including knee MRI. We used an OAI sub-cohort with publicly available gold standard estimates of cartilage quantity at different visits. This allowed evaluation of the ability to quantify changes in cartilage quantity.

Specifically, we used the cartilage volume scores from OAI project 9B derived from manual segmentations produced by Chondrometrics[18] for the medial/lateral tibial and the load-bearing part of the medial/lateral femoral cartilage compartments at visits 00 and 01 for 565 knees coming from 556 subjects. Project 9B included the index knees from a subset of the OAI Progression cohort and is representative of a typical clinical trial population with KL 2 and 3 knees, some pain, and substantial joint space width remaining. The study population characteristics are shown in Table 1.

The 88 knees with publicly available semi-manual segmentation masks for the DESS sequence provided by iMorphics[14] were used for training here and in the original 2019 IWOAI segmentation challenge[16]. This set includes semi-automated masks for the tibial, femoral, and patellar cartilage compartments.

Anonymization of knee MRI

All scans that were segmented in this study were anonymized before being shared with the participating teams. Thereby, the teams were blinded to any clinical, radiology readings, or visit information.

Teams and Segmentation Methods

The six teams from the 2019 IWOAI segmentation challenge were invited and all teams accepted. The teams were instructed to use the same method as in the original challenge, but were invited to also submit segmentations from an updated version. Teams 3 and 4 did this. Training of the 2019 challenge methods was therefore done previously[16], while training of updated methods, all segmentations, and further analysis were done for this study.

The methods are summarized in Table 2, highlighting pre-processing or intensity normalization steps that could likely affect the robustness against drift and shift effects.

Post-processing of Segmentations

The gold standard compartment definitions differed from the compartment masks used for training the DL methods. Therefore, we performed post-processing of the segmentations from the DL methods.

The tibial cartilage compartment was split into a medial and a lateral compartment using a simple k-Means split based on the scan coordinates for the voxels included in the tibial compartment.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The process for defining femoral cartilage sub-compartments is illustrated in Figure 2. First, a medial/lateral femoral compartment was defined using the sagittal extent of the tibial medial/lateral compartments. The gap between the tibial compartments was split in three. The center-most coordinate range was excluded from the femoral mask, defining medial/lateral sub-compartments. For each medial/lateral femoral compartment, a load-bearing sub-region was defined using the axial and coronal coordinates for all voxels included in the segmentation. Mean axial and coronal coordinates defined a center. The direction of most variation was computed using principal component analysis. From the center and in this direction, a "ceiling" was defined. Voxels below this ceiling in the medial/lateral femoral segmentation were included to define load-bearing femoral sub-regions.

These resulting medial/lateral femoral sub-regions were intended to approximate load-bearing sub-compartments similar to those used by Chondrometrics for the gold standard readings that were defined as "using 75% of the distance between the trochlear notch and the posterior of the femoral condyle"[14].

The medial and lateral tibial compartments are denoted MT and LT. The central medial and lateral femoral compartments are denoted cMF and cLF.

Efficacy Biomarker

The segmentation methods were primarily evaluated by their ability to quantify cartilage loss. We used cartilage volume as the imaging biomarker. The analysis included the medial and lateral tibial cartilage compartments, medial and lateral load-bearing femoral compartments, and the patellar cartilage compartment. However, no gold standard was available for the patellar cartilage compartment.

In addition, we computed statistics for the median volume (Med) across the teams for each cartilage compartment to allow a simple summary evaluation.

Performance Metrics

Following BIPED[19], a clinical trial efficacy biomarker should be evaluated for the sensitivity to measure treatment effects. However, the lack of available treatment studies challenged this. Following the OARSI recommendations[15], we therefore evaluated the sensitivity to change in the efficacy biomarker. Specifically, we evaluated the compartment-wise standardized response mean (SRM) for the cartilage volume: SRM = mean(Δvol)/std(Δvol), where Δvol is the signed change in volume from visit 00 to visit 01. The mean and standard deviation (std) were computed across the cohort for each compartment.

Robustness to Intensity Drift and Shift

The performance metrics above evaluate suitability for use in multi-center clinical trials. However, they do not directly reveal how sensitive the methods are to Drift and Shift events.

Therefore, we evaluated whether there were significant jumps in estimated cartilage volume across potential Shift events. Secondly, focused on Drift, we investigated whether there were trends in estimated cartilage volumes (after normalization for any jumps across Shift

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

events). Finally, we explored if any such effects would possibly be explained by differences in population across Shift events.

Specifically, we fitted a piece-wise linear model to total central tibio-femoral cartilage volume as a function of scan date. Here, discontinuities were allowed at known, potential intensity Shift events at each site (illustrated in Figure 1). The line slope was constrained to be equal for all line pieces for each site (assuming constant scanner drift). Given the complexity of Drift/Shift sources, these models were mainly discussed qualitatively.

Statistical Analysis

For evaluation of the accuracy of the cartilage volume measurements, we computed the mean relative signed difference (Offset) and the linear correlation coefficient (r) between the DL and the gold standard estimates for the baseline scans. For further investigation of potential bias in the segmentations, we constructed Bland-Altman plots.

For evaluation of the accuracy of the longitudinal volume changes, we computed the mean volume difference for each method. To test whether the volume change estimates were statistically different from the gold standard, we used a paired t-test on the quantified volume differences between baseline and follow-up.

As primary performance measure, we computed the standardized response mean (SRM) for each method and compartment. Confidence intervals for the SRMs were computed using bootstrapping with 95% Cis estimated using the bias corrected and accelerated percentile method[20].

For detection of a Shift effects, we included Shift events with more than 10 scans in the intervals directly before and after and used an unpaired t-test between the before/after samples to statistically significant differences. For detection of Drift effects, we quantified the Pearson linear correlation coefficient for the samples after correction for Shift effects (in the spirit of the piece-wise linear model).

The extraction of the cartilage volumes from segmentation masks was done using JupyterLab notebooks that are available at the study repository (see below). The statistical analysis was done using Matlab.

Data and Open Access

The original MRI scans are available at the OAI web site at https://nda.nih.gov/oai/.

The segmentations from all teams and the JupyterLab notebook that extracts all cartilage compartment volumes are available at the University of Copenhagen Electronic Research Data Archive (ERDA) with DOI https://doi.org/10.17894/UCPH.14A5084C-4618-4A8F-9A59-867654EC060B at https://erda.ku.dk/archives/1518a9c6b1db56269ef6ef62badd9d31/published-archive.html. The list of included scans and the unblinding codes are available upon request to the corresponding author.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Some teams have shared the source code and, in some cases, also the trained models. The links are included in Table 2.

Results

All scans from the 556 subjects in the cohort (see Table 1 for demographics) were successfully analyzed by all 6 participating teams.

Volume Biomarker Accuracy

The agreement between the volume measurements from the gold standard segmentations and DL methods are in the two top sections of Table 3. For the baseline (BL) measurements, there were high correlations for the tibial MT and LT compartments with r between 0.90 and 0.95. For the cMF and in particular for the cLF compartments, the correlations were lower and less consistent between the methods, ranging from 0.75 to 0.9. The DL methods typically estimated higher cartilage volumes for the tibial compartments with between 6 and 10 % over-estimation. For the femoral compartments, the methods were less consistent in over/under-estimation.

For the change measurements from BL to follow-up (FU), the methods typically estimated cartilage loss similar to the gold standard. The median volume from the teams (Median) was around 0.2 % from the gold standard for all four tibial/femoral compartments. The paired t-tests showed that for 5 of the 36 compartment/method comparisons, p-values at 0.01 or 0.02 indicated that these estimates were statistically different from the gold standard. No correction for multiple comparisons was done.

Sensitivity to Change

The primary performance evaluation was the sensitivity to measure changes in cartilage volumes estimated by the SRM (bottom section of Table 3). The gold standard had SRMs between 0.21 and 0.34 for the four tibial/femoral compartments. The median of the DL methods had SRMs between 0.17 and 0.33 for these compartments and 0.22 for the patellar compartment. The highest SRMs were found for the LT compartment with 0.34 for the gold standard and 0.38 for the best DL method.

Drift and Shift Robustness

The impact of Shift events on the resulting total central tibio-femoral (cTF) cartilage volume measurements is shown in Table 4. For perspective, the differences in BMI, Age, KL score across the Shift events are included in Table 4. These differences are generally not statistically significant.

There were clear, statistically significant differences in observed cartilage volumes for two of the three Shift events. For one of these, from site 20576, these differences are similar between the gold standard and the DL methods. However, for the event at site 20575 involving both a change of scanner station and a scanner software update, the cartilage volume differences across the event are higher for the DL methods than for the gold standard. Here, the Median of the DL estimates showed mean cartilage volume difference at −10% whereas the gold standard showed −4%.

The impact of potential Drift effects is shown in Table 5. Correlation between scan day and cTF cartilage volume could indicate a scanner drift effect. Also in Table 5, BMI, Age, and KL grade are included for perspective. For three of the sites, 20575/20576/20579, there was positive or no significant correlation between scan day and BMI/Age/KL. A positive correlation corresponds with the population becoming older, heavier, and more affected by OA over time. For these sites, there was no significant correlation between scan day and total tibio-femoral cartilage volumes.

However, for site 20574, there was a negative correlation between Age and scan day and a clear, positive correlation between cartilage volume and scan day. These observations are treated in the Discussion.

Discussion

Primary Results

Sensitivity to change was similar for the gold standard and the computer-based methods for the compartments with matching definitions, MT and LT. Here, from Table 3, the gold standard had SRMs at 0.21 and 0.34, respectively, and the median of the DL methods had 0.17 and 0.33. The highest SRM was achieved by method T3+ for the LT compartment with SRM 0.38.

For the femoral compartments with differing region definitions, SRMs were higher for the gold standard at 0.31 and 0.30 for cMF and cLF compared to 0.19 and 0.21 for the median of the DL methods. This does not reveal whether the simple sub-compartment definition was inappropriate or whether the segmentations were less accurate. However, given the results of the IWOAI 2019 segmentations challenge[16], the femoral segmentations were likely accurate and the culprit is most likely a too naïve sub-compartment definition.

We further investigated whether the sensitivity to change was affected by the degree of pathology since automated segmentation methods are often suspected of being less robust for knees with advanced OA. Since the cohort almost exclusively contained KL 2 and KL 3 knees (see Table 1), we split the cohort into KL<=2 and KL>2 sub-groups. The SRMs for cTF are shown in the supplementary material. The results show that SRM was very similar for these two sub-groups for the DL methods (0.32 for KL <=2 vs 0.31 for KL>2 for the median of the DL methods). However, for the gold standard comparison the SRM was higher for the KL>2 group (0.42 vs 0.36). This may indicate that the DL segmentations could be less robust than the manual for progressed OA. This would be very relevant to investigate further on cohorts with a wider range of KL scores.

Unlike the gold standard, the DL segmentations included the patellar cartilage. Here, the SRM for the median of the methods was 0.22, somewhat lower than for the tibial compartments. This may be a consequence of the sub-cohort being originally defined as likely progressors in the TF compartments.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

The extended version of Table 3 in the supplementary material includes 95% confidence intervals for the SRM scores. These reveal that the differences in SRM between methods are generally not statistically significant. *For instance, for the LT compartment where the SRMs are highest, the gold standard method had 0.34 [0.26;0.41] and the median of the DL methods had 0.33 [0.26;0.40]. For the cLF compartment where the gold standard was better defined anatomically, the gold standard had 0.30 [0.22;0.37] and the median of the DL methods had 0.21 [0.12;0.28].*

Secondary Results

The segmentation accuracy was evaluated by the correlation between the cartilage volumes for the gold standard and computer-based methods for the baseline scans. For the tibial compartments these were very high at 0.94 and 0.95, but lower for the central femoral compartments at 0.86 and 0.82, as seen in Table 3. This is consistent with the lower performance in the femoral compartments mentioned above.

The Offset between DL methods and gold standard were between −4% and +11% for the median of the DL methods for the four compartments at baseline. However, the estimations of % Loss is more consistent between the DL methods and the gold standard. This indicates that the DL methods are likely consistently over- or under-segmenting in each compartment compared to the gold standard. This is confirmed by the Bland-Altman plots in the supplementary material.

Robustness Against Sanner Drift and Shift Events

The OAI was not designed to investigate scanner drift and shift effects and it is challenging to conclude whether the methods handled these effects robustly.

Focusing on the Shift effects, it is clear from Figure 1 that the events can have drastic effects on the scan intensity level. Table 4 reveals the computer-based methods appeared to be sensitive to some of these. In particular, one event at site 20575 was associated with large differences in cTF volume before and after the event. The gold standard method reported −4% difference whereas the DL methods reported between −8% and −12%. This would indicate that even if several of the DL methods are overall comparable to the gold standard in terms of robustly quantifying changes (as evaluated by SRM in Table 3), they may still be more sensitive to some events. Here, a change of scanner station including a change of scanner software version appeared to cause the DL methods to detect less cartilage in their segmentations. However, since the subjects scanned before and after the event are not controlled, some of the volume difference across the event is likely due to actual differences between subjects.

This study design challenge becomes even more clear for the investigation of the Drift effects in Table 5. For three sites, there were no statistically significant trend in the cTF cartilage volumes over time. However, for site 20574, there were strong linear correlations between scan day and measured volumes with coefficients around 0.3 for both gold standard and DL methods. The positive correlation even suggests the subjects were growing cartilage. However, there was also a correlation between age and scan day at −0.17, suggesting the subjects became younger. Figure 3 shows that the ages were simply lower for subjects

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

scanned later in the study, so even if all subjects aged one year between the visits, the overall trend was actually a decline of 3.6% in subject age per year. This makes it hard to conclude whether the trend in cTF volume at an increase of 11.3% per year (not shown) was due to a scanner drift effect or an age confounding effect.

Potentially, a different cohort design would allow for analysis of more subtle drift/shift effects. In the present cohort, the subjects generally experience cartilage loss. A cohort designed to include structurally stable knees (possibly defined based on available semi-quantitative readings on the OAI) would potentially make it simpler to isolate measured cartilage volume changes as a consequence of drift/shift effects.

Further, more sophisticated statistical analysis could potentially incorporate the known potential confounders (e.g. age, BMI, and KL) to reveal if the cartilage volume measurements are independently affected by Shift effects. However, the uncontrolled study design (regarding these effects) makes this challenging.

Deep Learning Architecture Choices

The methods were quite similar in performance for the secondary performance measures, while there appeared to be qualitatively larger differences in the primary outcome. However, there was no clear relationship between architectural differences and performances.

For both the teams with updated methods, the updates resulted in slight performance improvements (Table 3). This indicates that ELU is indeed slightly better than RELU. And that adding more elaborate optimization (added data augmentation) may provide improvement. Both of these choices are consistent with general advances in Deep Learning.

Limitations of the Study

Mean cartilage thickness including denuded areas may be a more sensitive and suitable biomarker for cartilage loss than volume, although studies are conflicting[21–23]. However, this requires a model of the underlying bone to determine the total area of bone, which is not included in this study. We therefore focused on cartilage volume to directly validate the segmentation methods rather than some elaborate thickness estimation step.

In the present study, we performed a post-processing step to extract a central load-bearing sub-region in the femoral compartments. It is known that cartilage loss is inhomogeneous across the femoral compartment, likely even including areas with thickening[24,25]. However, our simple post-processing step was based on the cartilage segmentation and not the bone anatomy. Therefore, segmentation variability will also lead to variability in the sub-region definition. However, in line with the above decision regarding volume vs thickness, we chose a crude post-processing step to keep the analysis simple.

Precision is often highlighted in evaluation of biomarkers[5]. This was not possible here since scan-rescan MRIs are not publicly available for the OAI. However, precision is reflected in the SRM since high measurement variation contributes to higher variation on change measurements and thereby lower SRM. However, low SRM may not imply poor precision.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Readings done by trained experts in clinical trials are often performed paired, inspecting baseline and follow-up together (ideally blinded to order). This is to ensure consistent readings and high precision. In this study, the automatic methods segmented the scans individually. Thereby, there is room for improvement since automatic methods could also benefit from paired analysis.

The statistical analysis assumed that the knees are independent. However, for 9 subjects both knees are included, violating this assumption. Since this subset is relatively very small, we did not use more sophisticated statistical analysis correcting for this. However, in the table in the supplementary materials, we added statistics for the primary SRM outcome for a subset only including a single knee for each subject. This analysis demonstrated qualitatively identical SRM values for the 565 and the 556 knee cohorts.

Finally, the OAI data used for this study is more standardized than most clinical trials. Specifically, all sites used the same scanner model and the exact same MRI sequences. The validated DL methods may be challenged by a less standardized, multi-vendor setup. Therefore, the crucial next step is to validate the DL methods on clinical trial data such as the cohorts included in phase II of the FNIH Biomarkers Consortium ("Progress OA")[26]. One straightforward strategy for handling the multi-vendor setup is to expand the collection of manually segmented training cases with examples representing all the scanner models and then retraining the DL models. This strategy is simple but requires expert annotator resources.

Conclusion

The aim of this study was – in line with the original IWOAI 2019 challenge – not to select a winner. Rather, this study intended to learn how robust different state-of-the-art DL methods were for quantifying longitudinal cartilage loss possibly confounded by scanner Drift and Shift effects. Since the OAI study was not designed to investigate these Drift/Shift effects, it is hard to firmly conclude how sensitive the DL methods were. However, it would appear that at least for some Shift effects, the DL methods were more affected than the gold standard method, causing an apparent cartilage loss.

Even so, the DL methods performed well for the primary study outcome measure, the sensitivity to detect cartilage volume changes. For the well-defined tibial compartments, the best DL methods were similar to the gold standard. The highest observed SRM at 0.38 was from a DL method in the lateral tibial compartment (compared to 0.34 for the gold standard).

Given the possible impact of Drift/Shift effects on computer-based methods, two simple trial design recommendations could be:

- Re-scan a subset of recently scanned subjects directly following a major system update to reveal potential Shift effects. A power analysis could reveal a suitable size for this subset.
- Similar to the randomization of subjects for treatment/control groups, visit dates should be designed to avoid bias (as seen in Figure 3).

Author Manuscript

Further, when adopting a DL method for a study, we recommend considering the evaluated methods:

- Select a subset of the methods based on the performance metrics reported here.
- Consider if these methods provide open-source implementations that you can use.
- Investigate if these methods have been validated on other cohorts, anatomies, and modalities.
- Realize that you may need to add quantification steps if you need more advanced imaging biomarkers than volume.
- Realize that you will need to retrain the methods using manually segmented training scans if you are not investigating scans acquired using the OAI DESS sequence.

These considerations imply that automatic cartilage quantification in clinical studies is still not available from a simple off-the-shelf software package. Adaptation and expert assistance are still needed. However, it should be noted that radiologist-based reading in clinical trials is also neither free nor trivial – requiring expert readers and elaborate protocols.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The OAI is a public-private partnership funded by the National Institutes of Health and conducted by the OAI Study Investigators. Private funding partners include Merck; Novartis, GlaxoSmithKline; and Pfizer. 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 partners.

The suggestion to define a structurally stable cohort (defined using semi-quantitative readings in the OAI) for further analysis of drift/shift effects was due to a reviewer. Thank you for the constructive reviews!

Competing Interest Statement

One of the participating teams was supported by NIH funding (NIH/NIAMS R01 AR075422 and AR077452).

The authors have no competing interests related to this study.

References

1. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957;16(4):494–501. [PubMed: 13498604]
2. Eckstein F, Sittek H, Milz S, Putz R, Reiser M. The morphology of articular cartilage assessed by magnetic resonance imaging (MRI). Surg Radiol Anat. 1994;16:429–438. [PubMed: 7725201]
3. Solloway S, Hutchinson CE, Waterton JC, Taylor CJ. The use of active shape models for making thickness measurements of articular cartilage from MR images. Magn Reson Med. 1997;37(6):943–952. [PubMed: 9178247]
4. Folkesson J, Dam E, Olsen OF, Pettersen P, Christiansen C. Automatic Segmentation of the Articular Cartilage in Knee MRI Using a Hierarchical Multi-class Classification Scheme. In: Duncan JS, Gerig G, eds. Medical Image Computing and Computer-Assisted Intervention –

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

MICCAI 2005. Vol 3749. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2005:327–334. Accessed February 27, 2013. http://link.springer.com/chapter/10.1007/11566465_41

5. Eckstein F, Guermazi A, Gold G, et al. Imaging of cartilage and bone: promises and pitfalls in clinical trials of osteoarthritis. Osteoarthr Cartil OARS Osteoarthr Res Soc. 2014;22(10):1516–1532. doi:10.1016/j.joca.2014.06.023

6. Iriondo C, Liu F, Calivà F, Kamat S, Majumdar S, Pedoia V. Towards understanding mechanistic subgroups of osteoarthritis: 8-year cartilage thickness trajectory analysis. J Orthop Res. 2021;39(6):1305–1317. doi:10.1002/jor.24849 [PubMed: 32897602]

7. Bowes MA, Kacena K, Alabas OA, et al. Machine-learning, MRI bone shape and important clinical outcomes in osteoarthritis: data from the Osteoarthritis Initiative. Ann Rheum Dis. 2021;80(4):502–508. doi:10.1136/annrheumdis-2020-217160 [PubMed: 33188042]

8. Linka K, Thüring J, Rieppo L, et al. Machine learning-augmented and microspectroscopy-informed multiparametric MRI for the non-invasive prediction of articular cartilage composition. Osteoarthritis Cartilage. 2021;29(4):592–602. doi:10.1016/j.joca.2020.12.022 [PubMed: 33545330]

9. Marques J, Genant HK, Lillholm M, Dam EB. Diagnosis of osteoarthritis and prognosis of tibial cartilage loss by quantification of tibia trabecular bone from MRI. Magn Reson Med. 2013;70(2):568–575. doi:10.1002/mrm.24477 [PubMed: 22941674]

10. Dam EB, Runhaar J, Bierma-Zienstra S, Karsdal M. Cartilage cavity-an MRI marker of cartilage lesions in knee OA with Data from CCBR, OAI, and PROOF. Magn Reson Med. 2018;80(3):1219–1232. doi:10.1002/mrm.27130 [PubMed: 29493000]

11. Chaudhari AS, Kogan F, Pedoia V, Majumdar S, Gold GE, Hargreaves BA. Rapid Knee MRI Acquisition and Analysis Techniques for Imaging Osteoarthritis. J Magn Reson Imaging. 2020;52(5):1321–1339. doi:10.1002/jmri.26991 [PubMed: 31755191]

12. Thomas KA, Kidziński Ł, Halilaj E, et al. Automated Classification of Radiographic Knee Osteoarthritis Severity Using Deep Neural Networks. Radiol Artif Intell. 2020;2(2):e190065. doi:10.1148/ryai.2020190065

13. Dam EB. Simple Methods for Scanner Drift Normalization Validated for Automatic Segmentation of Knee Magnetic Resonance Imaging - with data from the Osteoarthritis Initiative. ArXiv171208425 Cs. Published online December 22, 2017. Accessed December 26, 2017. http://arxiv.org/abs/1712.08425

14. Nevitt M. The Osteoarthritis Initiative (OAI). The Osteoarthritis Initiative. https://nda.nih.gov/oai/

15. Hunter DJ, Altman RD, Cicuttini F, et al. OARSI Clinical Trials Recommendations: Knee imaging in clinical trials in osteoarthritis. Osteoarthr Cartil OARS Osteoarthr Res Soc. 2015;23(5):698–715. doi:10.1016/j.joca.2015.03.012

16. Desai AD, Caliva F, Iriondo C, et al. The International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge: A Multi-Institute Evaluation and Analysis Framework on a Standardized Dataset. Radiol Artif Intell. Published online February 10, 2021:e200078. doi:10.1148/ryai.2021200078

17. Jog A, Fischl B. Pulse Sequence Resilient Fast Brain Segmentation. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Lecture Notes in Computer Science. Springer International Publishing; 2018:654–662. doi:10.1007/978-3-030-00931-1_75

18. Eckstein F, Maschek S, Wirth W, 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(5):674–679. doi:10.1136/ard.2008.089904 [PubMed: 18519425]

19. Bauer DC, Hunter DJ, Abramson SB, et al. Classification of osteoarthritis biomarkers: a proposed approach. Osteoarthritis Cartilage. 2006;14(8):723–727. [PubMed: 16733093]

20. DiCiccio TJ, Efron B. Bootstrap confidence intervals. Stat Sci. 1996;11(3):189–228. doi:10.1214/ss/1032280214

21. Eckstein F, Wirth W. Quantitative Cartilage Imaging in Knee Osteoarthritis. Arthritis. 2011;2011:1–19. doi:10.1155/2011/475684

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

22. Hunter DJ, Niu J, Zhang Y, et al. Change in cartilage morphometry: a sample of the progression cohort of the Osteoarthritis Initiative. Ann Rheum Dis. 2009;68(3):349–356. doi:10.1136/ard.2007.082107 [PubMed: 18408248]

23. Cromer MS, Bourne RM, Fransen M, Fulton R, Wang SC. Responsiveness of quantitative cartilage measures over one year in knee osteoarthritis: Comparison of radiography and MRI assessments. J Magn Reson Imaging. 2014;39(1):103–109. doi:10.1002/jmri.24141 [PubMed: 23580461]

24. Jørgensen DR, Lillholm M, Genant HK, Dam EB. On Subregional Analysis of Cartilage Loss from Knee MRI. Cartilage. 2013;4(2):121–130. doi:10.1177/1947603512474265 [PubMed: 26069655]

25. Buck RJ, Wyman BT, Hellio Le Graverand MP, Hudelmaier M, Wirth W, Eckstein F. Osteoarthritis may not be a one-way-road of cartilage loss – comparison of spatial patterns of cartilage change between osteoarthritic and healthy knees. Osteoarthritis Cartilage. 2010;18(3):329–335. doi:10.1016/j.joca.2009.11.009 [PubMed: 19948267]

26. Hunter DJ, Deveza LA, Collins JE, et al. Multivariable Modeling of Biomarker Data From the Phase I Foundation for the National Institutes of Health Osteoarthritis Biomarkers Consortium. Arthritis Care Res. 2022;74(7):1142–1153. doi:10.1002/acr.24557

27. Dam EB, Lillholm M, Marques J, Nielsen M. Automatic segmentation of high- and low-field knee MRIs using knee image quantification with data from the osteoarthritis initiative. J Med Imaging. 2015;2(2):024001–024001. doi:10.1117/1.JMI.2.2.024001

28. Deniz CM, Xiang S, Hallyburton RS, et al. Segmentation of the Proximal Femur from MR Images using Deep Convolutional Neural Networks. Sci Rep. 2018;8(1):16485. doi:10.1038/s41598-018-34817-6 [PubMed: 30405145]

29. Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). ; 2016:565–571. doi:10.1109/3DV.2016.79

30. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2018;40(4):834–848. doi:10.1109/TPAMI.2017.2699184 [PubMed: 28463186]

31. Mortazi A, Karim R, Rhode K, Burt J, Bagci U. CardiacNET: Segmentation of Left Atrium and Proximal Pulmonary Veins from MRI Using Multi-view CNN. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, eds. Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017. Lecture Notes in Computer Science. Springer International Publishing; 2017:377–385. doi:10.1007/978-3-319-66185-8_43

32. de Brébisson A, Vincent P. The Z-loss: a shift and scale invariant classification loss belonging to the Spherical Family. ArXiv160408859 Cs Stat. Published online May 27, 2016. Accessed August 12, 2021. http://arxiv.org/abs/1604.08859

33. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 Bias Correction. IEEE Trans Med Imaging. 2010;29(6):1310–1320. doi:10.1109/TMI.2010.2046908 [PubMed: 20378467]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

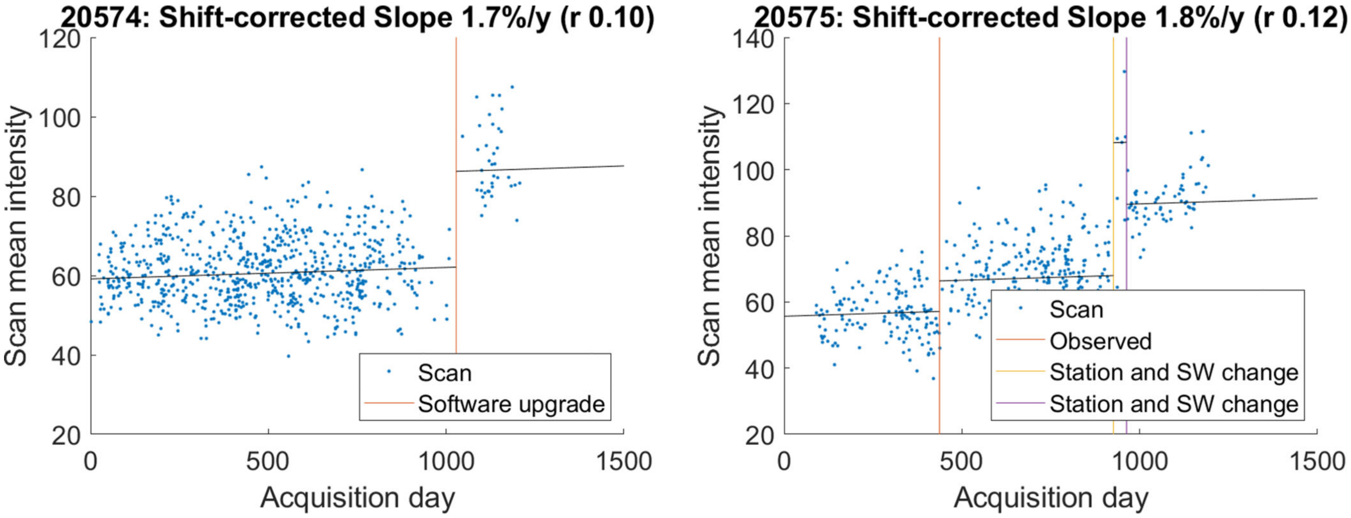

Figure 1:

Drift and Shift effects as observed from two OAI sites. The MRI scan mean intensity is plotted against the scan acquisition day for two of the four OAI sites. For both sites, the scan intensities increased gradually by 1.7% and 1.8% per year. However, these drifts were interrupted due to changes in scanner software and hardware, resulting in relatively large, abrupt shifts in intensity. The shift events were derived from the DICOM headers (e.g. attributes for software version and station name), except for the “Observed” event that appeared highly plausible from the data but had unknown origin. Note that these figures include all OAI scans at these sites at the 0 and 1 year visits and not only the Project 9B scans included in this study.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

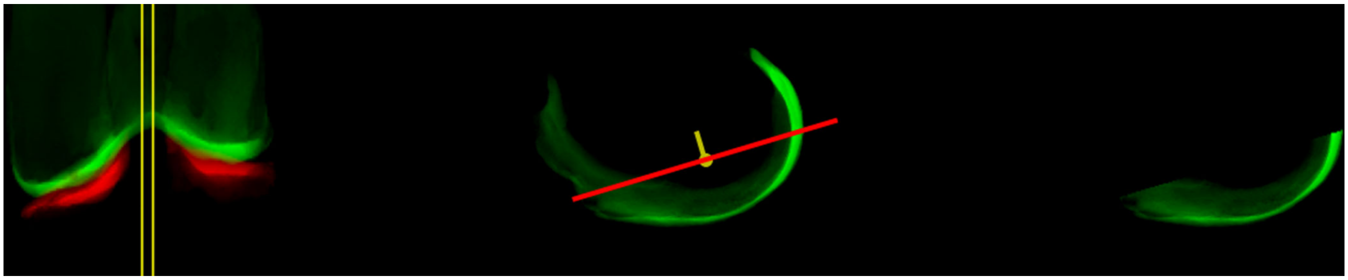

Figure 2:

The segmentations included the tibial and femoral compartments that were split during postprocessing. The process is illustrated for the first scan segmented by Team 1. The visualizations project the 3D segmented compartments onto a scan axis and sum up the segmented voxels for each position, giving a radiograph-like impression.

Left: During postprocessing, the tibial compartments were split into a medial and a lateral compartment by the k-means algorithm and the gap between these tibial compartments was split in three. The dividing planes defined medial and lateral femoral sub-compartments. The tibial and femoral cartilages are visualized in red and green, respectively, seen from the front with the medial compartment to the right. The yellow lines show the gap between medial and lateral tibial cartilage.

Center: For medial and lateral femoral compartments, the coordinates of the included voxels were used to define a sub-compartment approximating a load-bearing region. The center figure shows the lateral femoral cartilage with segmented voxels accumulated medio-laterally with anterior left and posterior to the right. The center-of-mass is computed to define a center for the compartment. Principal component analysis is used to compute a primary and a secondary mode of orientation. The red line is the primary orientation going through the center-of-mass. Everything below the red line is included in the selected sub-compartment.

Right: The final lateral femoral sub-compartment, approximating a load-bearing sub-region.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

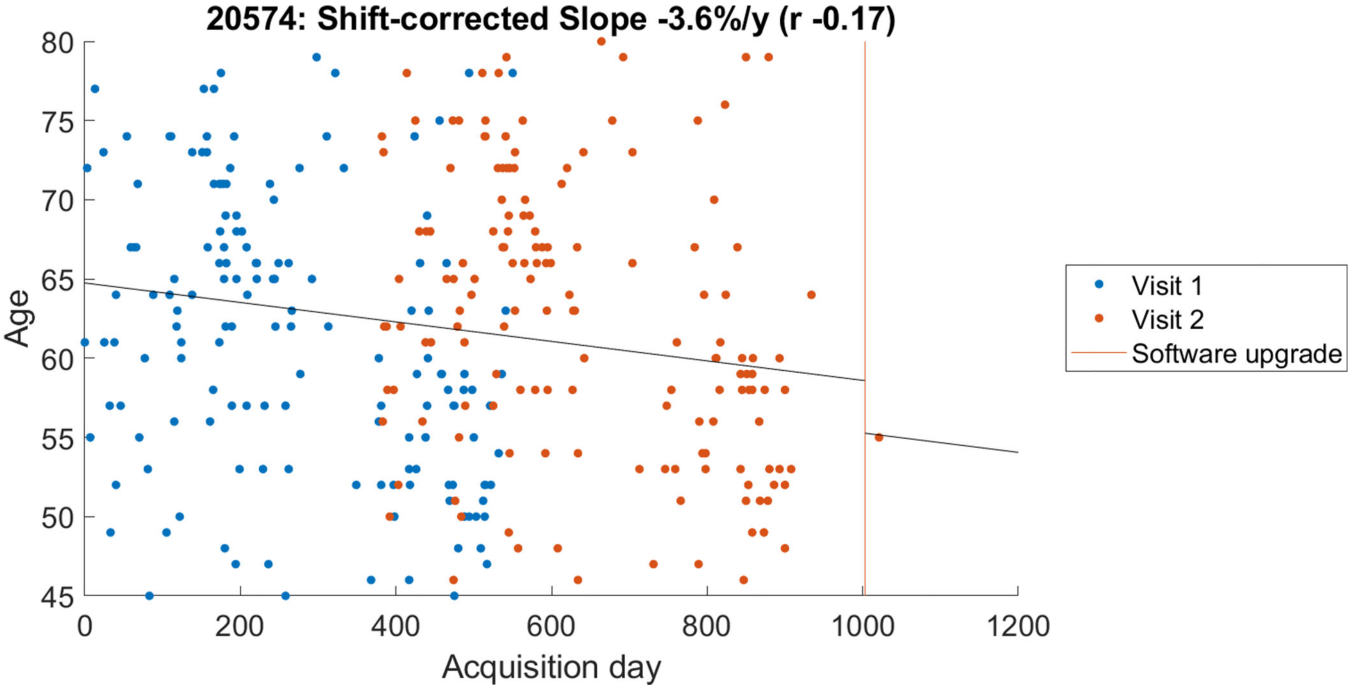

Figure 3:

Apparent rejuvenation at site 20574. The subjects in the later part of each visit are younger. This means that even if each individual was approximately one year older at the second visit, the overall trend suggests that the subjects were 3.6% younger per study year.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

The study population of 556 subjects at baseline summarized by number, age, BMI, KL score, WOMAC pain score, and minimal medial joint space width (mmJSW) for the four sites. For Age, BMI, pain, and mmJSW the table shows mean ± standard deviation at baseline. For KL, the first line shows baseline grade and the second line shows number of progressors at 1 year follow-up. The population is evenly distributed and fairly homogeneous across sites, resembling a clinical trial cohort with mostly KL 2/3 subjects, some pain, and preserved JSW.

Site	N (%female)	Age [years]	BMI [kg/m^2]	KL (N at 0,1,2,3,4)	WOMAC Pain	mmJSW [mm]
20574	153 (61)	61.3 ± 8.8	29.6 ± 4.6	00 03 78 71 01 00 00 10 10 00	4.1 ± 3.5	4.0 ± 1.5
20575	128 (61)	61.7 ± 8.9	30.9 ± 5.3	02 02 69 53 02 01 01 05 04 00	5.5 ± 4.0	4.0 ± 1.4
20576	146 (51)	60.9 ± 8.6	29.6 ± 4.3	01 06 63 75 01 00 03 03 02 00	4.1 ± 3.4	3.9 ± 1.4
20579	129 (58)	60.7 ± 9.1	30.7 ± 5.5	01 07 54 66 01 00 00 04 11 00	5.4 ± 4.2	3.7 ± 1.4

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2:

Segmentation Methods. The manual expert segmentation (Gold) is used as gold standard and a non-DL method (KIQ[27]) is used for comparison. The methods from the 2019 IWOAI knee segmentation challenge are named T1-T6 (as originally[16]).

Team	Method	Optimization Loss	Intensity Normalization
Gold	The gold standard segmentations were done using slice-wise manual outlining done by training readers in a quality-assured setup[18].	Visual inspection	Manual intensity windowing
KIQ	kNN voxel classification using feature selection among Gaussian derivative features[27]. This non-Deep Learning method from the preliminary study[13] is used as baseline method.	Dice	Affine global intensity correction from multi-atlas rigid registration
T1	Multi-class 3D U-Net architecture with dilated convolutions at the bottleneck layer to increase the effective receptive field[28]github.com/denizlab/2019_IWOAI_Challenge	First cross-entropy and then fine-tuning with soft Dice	Volumes were zero-mean whitened (zero-mean, unit variance).
T2	Cascaded ensemble of 3D and 2D variants of the V-Net[29] using dropout and intensity/geometric data augmentation[6]	Dice	Data augmentation using randomly sampled intensity transformations
T3	Multi-planar sampling of volume into 2D U-Net with batch normalization and geometric data augmentation github.com/perslev/MultiPlanarUNet	Unweighted cross-entropy	Intensity normalized to median zero and inter-quartile range one
T3^{+}	As T3 where activation is ELU instead of RELU github.com/perslev/MultiPlanarUNet	Unweighted cross-entropy	As T3
T4	Modified DeeplabV3 with dense connections at bottleneck block and dilated multi-scale features[30]	Soft multiclass Dice	No Augmentation. Volumes were normalized to zero-mean and unit variance.
T4^{+}	Added dropout layer to T4 architecture and extended data augmentation	Soft multiclass Dice	Geometric, Intensity and Noise addition based data augmentation. Volumes were normalized to zero-mean and unit variance.
T5	Encoder-decoder CNN architecture using Dense Blocks with tri-planar fusion of 2D models with geometric data augmentation[31]github.com/ali-mor/IWOAI_challenge	Z-loss[32]	Bias field and intensity normalization using N4ITK[33] followed by edge-preserving non-linear diffusion for noise reduction.
T6	2D U-Net applied slice-wise in the sagittal plane. U-Net had 5 pooling steps. http://github.com/ad12/DOSMA	Soft Dice	Volumes were zero-mean whitened (zero-mean, unit variance).

“+” The updated methods have a added. For each method, the optimization target and any intensity normalization steps are summarized. Links are included for methods with available open-source implementation and trained model weights.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3:

Accuracy and Sensitivity to Change. The agreement for volume measurements between the gold standard and each team are given both at baseline (BL) and for changes from BL to follow-up (FU). For BL, this is shown by Pearson correlation coefficient r and signed relative difference in mean value (Offset in %). For changes FU-BL, this is shown as the mean signed loss and the p value from a paired t-test indicating whether each method is significantly different from the gold standard. The sensitivity to measure change given by the standardized response mean (SRM) for each compartment for each method is at the bottom section of the table.

Team			**Gold**	**KIQ**	**T1**	**T2**	**T3**	**T3+**	**T4**	**T4+**	**T5**	**T6**	**Median**
BL	MT	r		0.91	0.93	0.94	0.94	0.94	0.93	0.94	0.91	0.90	0.94
		Offset		5.3	4.4	6.9	4.1	2.1	10.6	14.8	−20.3	6.7	5.8
	LT	r		0.88	0.92	0.92	0.95	0.95	0.95	0.95	0.94	0.91	0.95
		Offset		11.2	10.8	14.1	9.5	7.6	14.1	19.9	−4.7	10.9	11.2
	cMF	r		0.81	0.84	0.84	0.86	0.87	0.83	0.85	0.86	0.78	0.86
		Offset		2.5	−2.7	0.3	−4.7	−5.6	0.3	−6.5	−17.8	−1.8	−4.0
	cLF	r		0.75	0.79	0.79	0.81	0.80	0.82	0.80	0.81	0.78	0.82
		Offset		4.6	9.0	9.4	4.5	3.5	10.7	6.7	−4.9	7.6	6.5
FU-BL	MT	% Loss	1.0	0.9	0.7	0.5	0.8	0.7	0.3	0.5	1.3	0.6	0.8
		p		0.75	0.37	0.07	0.57	0.31	0.01	0.08	0.75	0.49	0.48
	LT	% Loss	1.4	1.5	1.8	1.4	1.8	1.8	1.9	1.6	1.3	2.0	1.7
		p		0.42	0.08	0.34	0.01	0.01	0.01	0.02	0.62	0.12	0.04
	cMF	% Loss	1.4	1.1	1.3	1.1	1.1	1.0	0.9	1.7	1.4	0.3	1.2
		p		0.34	0.64	0.35	0.21	0.09	0.18	0.57	0.30	0.09	0.35
	cLF	% Loss	1.1	1.1	1.2	1.5	1.6	1.5	1.1	0.7	1.1	1.9	1.4
		p		0.74	0.50	0.13	0.08	0.07	0.62	0.41	0.83	0.10	0.19
	P	% Loss		1.5	2.1	2.1	2.3	2.2	2.6	1.8	0.6	1.5	2.0
FU-BL	**MT**	**SRM**	**0.21**	**0.14**	**0.12**	**0.11**	**0.15**	**0.16**	**0.05**	**0.10**	**0.16**	**0.05**	**0.17**
	LT		**0.34**	**0.25**	**0.25**	**0.30**	**0.33**	**0.38**	**0.29**	**0.33**	**0.19**	**0.17**	**0.33**
	cMF		**0.31**	**0.16**	**0.16**	**0.15**	**0.16**	**0.15**	**0.11**	**0.21**	**0.18**	**0.02**	**0.19**
	cLF		**0.30**	**0.18**	**0.15**	**0.19**	**0.22**	**0.24**	**0.13**	**0.07**	**0.16**	**0.14**	**0.21**
	P		**NaN**	**0.12**	**0.21**	**0.23**	**0.23**	**0.23**	**0.24**	**0.23**	**0.06**	**0.13**	**0.22**

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4:

Shift Effects. Change in total central load-bearing tibio-femoral (cTF) cartilage volume across shift events causing intensity distribution shift.

Shift Events			
Site	20575	20575	20576
Event	Observed	Station change and SW update	Observed
Day	337	826	322
Scans before/after	39/152	152/65	86/207
Difference in BMI (%)	−2.8	−1.8	−1.1
Difference in Age (%)	1.5	1.7	*−3.8
Difference in KL (%)	−1.9	2.8	−3.6
Speculative expectation on OA state across event	Marginally better?	Marginally worse?	Likely better?
Difference in cTF Volume			
Gold (%)	−0.4	−3.9	**8.5
KIQ (%)	−6.0	***−12.2	*5.7
T1 (%)	−3.8	**−10.5	*7.0
T2 (%)	−3.9	**−10.3	*6.5
T3	−4.0	*−9.3	*7.8
T3+	−4.2	*−9.2	**8.0
T4	−5.1	**−10.1	*6.7
T4+	−4.8	*−9.3	*7.0
T5	−1.9	*−8.0	*7.8
T6	−4.6	*−9.4	*7.8
Med	−4.6	**−10.0	*7.3

Only events with at least 10 scans both before and after are included.

*, **, or *** All differences across shift events are shown as mean relative difference between before/after scans in % and marked by for statistical significance given by the p-value from an unpaired t-test with p<0.05, p<0.01, or p<0.001, respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5:

Drift Effects. Trend for cross-sectional drift in for central load-bearing tibio-femoral (cTF) cartilage volume where any changes across shift events are normalized away.

Site	20574	20575	20576	20579
BMI	0.01	*0.15	0.03	−0.10
Age	**−0.17	−0.04	***0.23	0.07
KL	−0.03	−0.04	**0.16	*0.15
Gold	***0.30	−0.01	−0.07	0.06
KIQ	***0.30	0.10	0.01	0.07
T1	***0.30	0.03	−0.02	0.08
T2	***0.29	0.04	−0.01	0.07
T3	***0.29	0.05	−0.05	0.06
T3+	***0.29	0.05	−0.05	0.06
T4	***0.29	0.07	−0.03	0.07
T4+	***0.30	0.05	−0.02	0.07
T5	***0.31	−0.01	−0.04	0.08
T6	***0.29	0.05	−0.04	0.07
Med	***0.29	0.05	−0.03	0.07

*, **, and *** The trends are shown as the Pearson correlation coefficient with statistical significance marked by for p<0.05, p<0.01, and p<0.001, respectively