10.1148/ryai.2020190116

Appendix E1

MRI Scan Acquisition

For the 100 patients included in the supervised learning group, scans were performed with one of our 1.5--T or 3T scanners (Vida, Avanto, Skyra, Verio or Prisma Fit; Siemens, Erlangen, Germany). Patients were placed in a supine position with arm in mild external rotation. Axial imaging extended from the acromioclavicular joint to below the axillary pouch. Dedicated shoulder array coils were used for imaging. For the 50 patients included in the transfer learning and the 35 patients included in the measurement group, scans were performed on one of our 1.5-T or 3-T scanners (Aera, Skyra or Prisma Fit, Siemens).

Segmentation Model

Padding in convolutions were incorporated to obtain the same size of input image and segmentation mask. Random horizontal flipping of the images was used for data augmentation to generalize on left- and right-sided images. As the number of slices changes between acquisitions, we selected either the center 32 slices from volumes that have more than 32 slices or mirrored the slices at the volume borders to obtain 32 slices from volumes that have less than 32 slices. This way we were able to train a single three-dimensional (3D) convolutional neural network (CNN) accepting a volume with 32 slices. Weights of the loss function for the specific class were calculated using the training set by dividing the total number of voxels with the number of voxels for that particular class. By random sampling, the sample was partitioned into four groups of 25 patients. Each group served as a validation set to assess the accuracy obtained from the other training set groups.

Transfer Learning Model

Two-dimensional (2D) slices from Dixon-based images and segmentation masks were interpolated using bicubic splines of third order to match the voxel size of proton density images, and they were center cropped to a size of 320 × 320 prior to model training. During transfer learning from 3D CNN, from each MR image we extracted four interleaved volumes of 32 slices after mirroring the slices at the volume borders to obtain 128 slices. Postprocessing details defined in the previous section were employed. Additional image postprocessing was used to bring the voxel size and dimension of the segmentation masks to the original water-only Dixon-based images.

Hardware Specifications

Experiments for both 2D and 3D CNNs based on U-Net architecture were performed on a server using an NVIDIA (Santa Clara, Calif) 16GB Tesla P100 GPU card.

Source Code

The source code for this study is available at:
https://github.com/denizlab/Shoulder_Segmentation

3D Models Production and Measurements

Using the software 3DSlicer (v4.11.0) we created 3D volume surface rendered models with the volume intensity: *(a)* For the glenoid, the model was obtained with a range of 1.00 to 1.01 of volume intensity, and *(b)* for the humeral head, the model was obtained with a range of 2.99 to 3.00 of volume intensity.

Two musculoskeletal radiologists assessed normal anatomy by measuring the humeral head width and diameter and the glenoid width, diameter, and height on the fully automated and the semiautomated 3D models at the first measurement session. They also estimated the glenoid bone loss (GBL) percentage on patients with an anterior shoulder instability history. These measurements were repeated by reader 1 during a second session 2 weeks after the first session.

Statistical Analysis

The level of agreement was assessed in terms of the absolute value of the differences. Reader agreement was assessed using the data from the first reading session for reader 1 to avoid confounding the difference between readers with any change in the performance of reader 1 over time (due to learning that often occurs when a task is performed repeatedly over a short period of time). We also report 95% confidence intervals (CIs) for the differences between the fully and semiautomated models in terms of the mean error between the diameter measurements performed by the two readers. The 95% CIs in these cases imply that there is a 95% confidence that the true mean error of the fully automated diameter measurements is captured within the interval.

CNNs Training Time

For the 2D CNN, the mean time for training each epoch was 5 minutes and 13 seconds, while for fully automated segmentation mask generation it was 1.28 seconds. For the 3D CNN, the mean time for training each epoch was 7 minutes and 13 seconds, while for fully automated segmentation mask generation was 1.12 seconds.

Outliers Overview

Four cases were outliers in both 2D and 3D CNNs. They presented with different causal factors, including field inhomogeneity, bone marrow heterogeneity, and a combination of motion artifact and field inhomogeneity.

Treatment Selection Agreement

Comparing how the GBL percentage measurements performed by reader 1 during the first and second sessions would impact treatment selection, the intrareader agreement to select the same treatment was 95.8% (23 of 24) for the fully automated models and 100% (24 of 24) for the semiautomated models. When comparing the GBL measurement treatment impact for readers 1 and 2 during the first session, the interreader agreement to select the same treatment was 95.8% (23 of 24) for the fully-automated models, and 95.8% (23 of 24) for the semiautomated models.

Table E1

Outliers

Causal factors	2D CNN (*n* = 100)*	3D CNN (*n* = 100)*
Total	8 (8%)	9 (9%)
Bone marrow edema	2 (2%)	1 (1%)
Bone marrow heterogeneity	3 (3%)	3 (3%)
Image artifacts		
Motion	1 (1%)	…
Field inhomogeneity	1 (1%)	1 (1%)
Motion + field inhomogeneity	1 (1%)	1 (1%)
Partial volume effects	…	2 (2%)

Note.— CNN = convolutional neural network, 3D = three-dimensional, 2D = two-dimensional.

*For each dataset, *n* is the number of patients enrolled in each task.