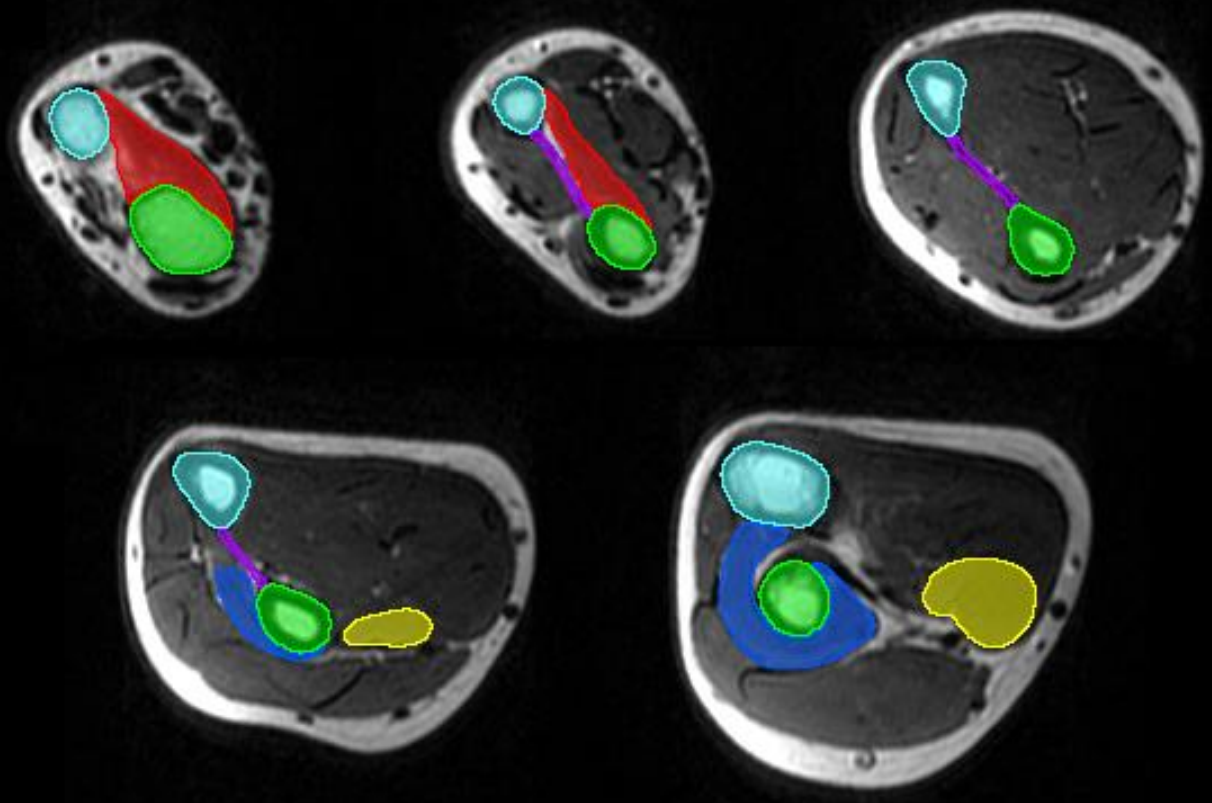
Enhancing Clinical Decision-Making in Posttraumatic Rotational Limitations of the Forearm Through Deep Learning-Based Segmentation Techniques

Wesley de Reus

MSc Thesis



Enhancing Clinical Decision-Making in Posttraumatic Rotational Limitations of the Forearm Through Deep Learning-based Segmentation Techniques

W.D.M. (Wesley) de Reus

Student number : 4651898

26-05-2025

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Biomechanical Engineering, TUDELFT

22-04-2024 – 07-03-2025

Supervisor(s):

Dr. J.W. (Joost) Colaris

Dr. J. (Jukka) Hirvasniemi

E.M. (Eline) van Es, MSc

Thesis committee members:

Dr. J.W. (Joost) Colaris (chair)

Dr. J. (Jukka) Hirvasniemi

Dr. E.H.G. (Edwin) Oei

E.M. (Eline) van Es, MSc

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus University Rotterdam

Preface

This thesis represents the final product of a challenging yet deeply rewarding seven-year journey at TU Delft, Erasmus MC and LUMC, in pursuit of becoming a Technical Physician. It signifies not only the completion of my academic path but also the conclusion of an important personal chapter. Over the past seven years, I have gained invaluable knowledge and experiences that have shaped me far beyond my initial expectations when I first embarked on this program. The journey has been one of remarkable educational growth and personal development, and I am grateful for every step along the way.

I would like to express my sincere gratitude to my supervisors, without whom this thesis would not have been possible. First, I wish to thank Joost, Eline, and Jukka, who took time out of their busy schedules to supervise and support me throughout my graduation year. Joost, your passion for advancing medical care within your field has been truly inspiring. I am grateful for the insights you shared into the clinical context of this research during the outpatient clinic sessions, surgeries, and meetings. Your approachable nature and consistent willingness to make time for questions and discussions greatly enriched my learning experience. Eline, as my daily supervisor, you were an essential source of support throughout this journey. Your availability and willingness to discuss ideas and challenges were invaluable to the successful completion of this thesis. I particularly appreciated your enthusiasm, your fresh perspectives, and your support with the organizational aspects of the project, all of which made the process smoother and more enjoyable. Jukka, your deep expertise in medical imaging and deep learning played a significant role in shaping the direction and quality of this work. I am sincerely thankful for your assistance in addressing technical challenges and for your thoughtful advice on research design, both of which were crucial in the success of this thesis. I would also like to thank the radiologist who generously dedicated time to meet with me for the validation of my manual annotations, providing thoughtful and constructive feedback. Additionally, I would like to extend my gratitude to all the staff members and residents of the Orthopedics & Sports Medicine department at Erasmus MC for their time, dedication, and invaluable contributions to enhancing my medical knowledge.

Finally, I would like to express my deepest appreciation for my family, friends and girlfriend. Your belief in me has pushed me to exceed my limits and kept me grounded, even during challenging times. Thank you for patiently listening to the unsolicited monologues about my research, knowing full well that not everyone could follow, yet always offering your unwavering support and encouragement.

Wesley de Reus

Rotterdam, March 2025

Abstract

Introduction: Fractures in the forearm are common and sometimes result in limitations of pronation/supination. Besides malunion as a possible cause, soft tissue involvement may play a more significant role as well. More insight in both causes of impaired forearm rotation could help to treat patients in the least invasive way as possible, potentially avoiding invasive corrective osteotomies in some patients.

Objectives: This study aimed to provide a deep-learning based framework for automated segmentation of anatomical structures involved in pronation/supination of the forearm on magnetic resonance (MR) images. This approach allows for visualization and quantitative analysis of the patient-specific anatomy, enabling efficient identification of soft tissue structures that may contribute to impaired forearm rotation.

Methods and materials: Manual ground truth annotations of six anatomical structures (radius, ulna, interosseous membrane, m. pronator quadratus, m. pronator teres, m. supinator) were performed on 24 fast-recovery fast spin-echo T2-weighted (FRFSE T2) in-phase Dixon images of the forearm. The dataset contained an equal distribution between affected and unaffected, and left and right forearms. Two nnU-Net configurations (2D and 3D) were trained on 20 manually segmented forearms using 5-fold cross-validation for segmentation of the six structures. An ensemble was created by combining predictions from both fully-trained models. A hold-out test set of 4 forearms was used to evaluate segmentation performance using the Dice similarity coefficient (DSC) and the average symmetric surface distance (ASSD) metrics. Additionally, relative volume difference (Δ_{rel}) between ground truth and predicted segmentations were computed to assess under- or oversegmentation.

Results: The 3D model achieved the best segmentation performance, with a median DSC score of 0.894 (IQR=0.094) and a median ASSD of 0.324 (IQR=0.386) mm. It slightly undersegmented the anatomy, with a median relative volume difference of -2.7% (IQR=7.1%). Qualitative results revealed that the 3D model produced segmentation masks that contained fewer and less severe segmentation errors compared to the 2D model and ensemble. Minor segmentation errors were observed in the interosseous membrane, the proximal part of the m. pronator quadratus and the insertion of the m. pronator teres in some cases.

Conclusion: The 3D nnU-Net model has proven its suitability for clinical use, enabling fast, reproducible and precise segmentation of structures involved in pronation/supination of the forearm. This approach facilitates bilateral comparisons of soft tissue structures through visual assessment and quantitative analysis, supporting patient-specific and minimally invasive decision-making.

Keywords

forearm, limited pronation/supination, soft tissues, deep learning, nnU-Net, MRI, automated segmentation

Contents

1. Introduction

1.1 Background

Fractures of the radius and/or ulna are common in children aged 5 to 14 years, accounting for approximately 34% of all pediatric fractures. [1] In this group, fractures are most frequently caused by axial loading applied to the forearm, typically resulting from a fall onto an outstretched hand. Although more prevalent in children, forearm fractures also occur in adults, primarily due to motor vehicle accidents, athletic injuries, and falls from height. Forearm rotation makes it possible to position the palm upward (i.e. supination) or downward (i.e. pronation) and is commonly used for carrying out tasks associated with daily life activities such as eating, writing, typing or accepting monetary change [2]. Pronation and supination involve a combination of rotation and translation of the radius, ulna and interosseous membrane (IOM) [3]. Forearm fractures can result in impaired rotation, hindering daily life activities. Movement is considered limited when the range of motion (ROM) for pronation/supination is reduced to less than the functional arc of 50 degrees in each direction [4], which patients often compensate for by using the ipsilateral shoulder.

Although malunion is frequently reported as a cause of limited pronation/supination following forearm fractures, its exact role remains a topic of debate in the literature. Several studies suggest that malalignment in angulation, translation and/or rotation in the forearm bones can lead to restricted pronation/supination due to mechanisms such as bone impingement [5]-[10]. However, others have claimed that there is no clear association between malunion and rotational limitations [11]-[15]. This perspective is supported by studies showing that corrective osteotomies for malunited forearm fractures did not consistently resolve limitations [14], [15]. Over time, there has been growing emphasis on the role of soft tissues in posttraumatic restriction of pronation/supination. Scarring of the injured soft tissue could result in contractures, which likely play a significant role in these limitations [16]. Furthermore, contractures of certain structures may lead to disuse and atrophy of other structures, potentially restricting motion even further. Several studies have demonstrated that surgical interventions targeting these constraints showed potential for improving function [17]-[20]. Currently, corrective osteotomies are often the therapy of choice for patients with malunion in case of rotational impairment, instability of the distal radioulnar joint (DRUJ) and/or pain, with the aim of restoring anatomy and improving function [21], [22]. This invasive procedure is generally considered only after conventional treatments such as physiotherapy and bracing have been attempted and proven ineffective. However, for patients with malunion causing only functional impairment, or in cases where functional impairment occurs without malunion, the focus could probably shift toward assessing the role of soft tissues in the limitation and addressing these issues, whether through surgical interventions or continued conventional therapies.

Magnetic resonance (MR) imaging is one of the most important imaging modalities in the field of orthopedics. This non-invasive technique can produce high-resolution images with excellent soft tissue contrast. While direct interpretation of these images is subjective, quantitative analysis may offer a more standardized and reproducible alternative. To date, no method has been published for performing quantitative analysis of soft tissue structures involved in pronation/supination of the forearm. Semantic segmentation offers a promising solution to address this challenge, involving the process of partitioning images into regions of interest (ROIs) by classifying each pixel or voxel into specific classes. While manual segmentation is time-consuming and subject to significant intra- and inter-observer variability, automating the process offers the potential for high speed and reproducibility. Deep learning, a subfield of machine learning, has emerged as a leading approach for automating segmentation tasks. Specifically, convolutional neural networks (CNNs) are designed to automatically identify patterns within images through convolution operations. These networks consist of multiple layers, including convolutional, pooling and fully connected layers, which work together to predict pixel or voxel labels and generate segmentation masks. Deep learning-based methods have significantly enhanced accuracy, efficiency and adaptability to diverse image modalities, providing a more reliable and reproducible alternative to manual approaches [23]. Automated segmentation enables the extraction of detailed quantitative metrics related to soft tissue morphology, such as volume, length and structural integrity. By comparing these metrics bilaterally, a deeper understanding of how soft tissues may contribute to the restrictions can be gained, facilitating patient-specific and minimally invasive decision-making.

1.2 Related work

Several studies have explored the application of deep learning-based techniques for automated segmentation of musculoskeletal structures on MR images [25]-[45]. Among these, U-Net has emerged as the most widely used architecture. Originally developed for biomedical image segmentation by Ronneberger et al., this CNN features an encoder-decoder structure [24]. The encoder captures hierarchical features from the input image through a series of convolutional layers, progressively downsampling the image. The decoder then reconstructs the segmentation mask by upsampling the encoded features, combining them with high-resolution features from the encoder through skip connections, which helps in precise localization and accurate segmentation.

The majority of research in this field has focused on segmentation of structures in the knee joint, with numerous studies using 2D U-Nets for segmentation of bones and cartilages in the knee on sagittal double echo steady state (DESS) images. Among these, Almajalid et al. reported a Dice similarity coefficient (DSC) score of 0.970 [25], Kim-Wang et al., achieved 0.984 [26], and Deng et al. reached 0.987 for segmentation of knee bones [27]. Latif et al. utilized an ensemble of 2D and 3D U-Nets for the same task, also attaining a DSC score of 0.987 [28]. Ambellan et al. combined 2D and 3D U-Nets with a statistical shape modeling (SSM) step and achieved a DSC score of 0.986 for segmentation of the femur and tibia [29]. Kemnitz et al. used a 2D U-Net for segmentation of both knee bones and muscles on axial T1-weighted MR images. They reported DSC scores of 0.970 and 0.956 for segmentation of knee bones and muscles, respectively [30]. Additionally, Flannery et al. conducted three studies targeting the anterior cruciate ligament (ACL). The first used a 2D U-Net for automated segmentation of the ACL on sagittal constructive interference in steady state (CISS) images. The second repurposed the same model for segmenting repaired ligaments and grafts, and the third adapted it for ACL segmentation on T2* images. DSC scores of 0.840, 0.800 and 0.760 were reported for these studies, respectively [31], [32], [33].

Substantial other work has been done in segmenting musculoskeletal structures in the shoulder region. Alipour et al. employed 2D U-Net models with different loss functions for automated segmentation of the rotator cuff muscles on oblique sagittal T1-weighted images of the shoulder. Their best-performing model employed the binary cross-entropy loss function, achieving a DSC score of 0.810 [34]. In a similar study, Medina et al. used sagittal T1-weighted images and achieved 0.963 [35]. Riem et al. also used sagittal T1-weighted images, adopting a 3D U-Net to segment rotator cuff muscles and shoulder bones, obtaining DSC scores of 0.905 and 0.932 for muscles and bones, respectively [36].

Beyond the knee and shoulder regions, another considerable amount of research has targeted the spine region. Three studies focusing on vertebrae segmentation on axial T2-weighted SPACE images using a 3D U-Net reported DSC scores of 0.925 by Zhu et al. [37], and 0.925 and 0.914 by Chen et al. [38] and Su et al. [39], respectively. Van der Graaf et al. conducted a similar study using sagittal T1- and T2-weighted images, achieving DSC scores of 0.930 and 0.920, respectively, for segmentation of vertebrae on both sequences [40]. Wesselink et al. adopted a 2D U-Net to segment the lumbar paraspinal muscles on axial T2-weighted images, achieving a DSC score of 0.921 [41].

While U-Net is widely regarded as state-of-the-art, achieving good results requires careful tuning of several network parameters, which can vary significantly depending on the segmentation task. To address the complexity and resource intensive nature of this tuning process, Isensee et al. introduced nnU-Net ('no-new-U-Net') as a self-configuring variant. [42] This framework automatically adapts to each dataset by configuring a U-Net based segmentation pipeline, determining data-dependent parameters –such as resampling strategy, image normalization scheme and batch size– that best match the data. This configuration process relies on a set of heuristic rules and a data fingerprint that contains data-specific properties, including image sizes, pixel spacings and intensity information. As a result, the need for expert knowledge and high computational costs is eliminated, providing nnU-Net with the user-friendly nature that it is renowned for.

Studies adopting nnU-Net for segmentation of musculoskeletal structures on MR images are still limited, likely attributable to its recent introduction. Li et al. adopted 2D and 3D nnU-Net for segmenting the mandibular condyle on sagittal PD-weighted images, reporting metric outcomes both in detected 2D slices and 3D volumes. The highest DSC value of 0.940 was reported for the 3D network when evaluated on 2D slices [43]. Hess et al. segmented rotator cuff muscles and shoulder bones using 2D and 3D nnU-Nets, as well as an ensemble, on axial, sagittal and coronal T1-weighted images. Their best performing model was the 3D network, which achieved DSC scores of 0.900 for muscles and 0.945 for bones [44]. A similar study by Kim et al. used 2D and 3D nnU-Net models for segmenting the same structures along with cuff tendons on axial T2-weighted images. Additionally, they compared the performance of both models with and without the use of a secondary labeling process, during which a secondary labeled dataset was created using the false-positive segmentation results and the manual annotations. The best performing model was the 3D nnU-

Net with secondary labelling, achieving DSC scores of 0.797, 0.978 and 0.801 for segmentation of muscles, bones and tendons, respectively [45]. Kamphuis et al. utilized an ensemble of 2D and 3D nnU-Nets for segmentation of femoral and acetabular bone in the hip joint using in-phase and water-only Dixon sequence images, and reported an average DSC score of 0.930 [46].

While most studies in this research field have focused on the lower extremity, spine or shoulder regions, similar research involving the forearm as anatomical region remains scarce. This explorative study serves as a pioneering effort to address this gap, aiming to provide a framework for the automated segmentation of anatomical structures involved in pronation/supination of the forearm on MR images using nnU-Net. This approach enables efficient identification of soft tissue structures that may be involved in posttraumatic restriction of pronation/supination through visualization and quantitative analysis of the patient-specific anatomy.

2. Methods and materials

2.1 Dataset and annotation procedures

The dataset used in this study consisted of MR scans of forearms both affected and unaffected by trauma from 28 patients, including children and adults (ages 7-35 years, 15 male, 13 female). Malunion was present in the affected forearms due to improperly healed fractures. These data were collected between 2019 and 2023 in a study approved by the medical ethics review committee. All images were acquired by 3.0T SIGNA scanners (GE Healthcare, Waukesha, WI, USA) via repeated scans in the axial planes of the forearm using a fast-recovery fast spin-echo T2-weighted (FRFSE T2) Dixon sequence (in-plane resolution: 0.3516 mm x 0.3516 mm, slice thickness: 4 mm, no interslice gap, flip angle: 111°, repetition time: 9480-13350 ms, echo time: 45-55 ms). Each patient had four different Dixon outputs available: in-phase, out-of-phase, water-only, and fat-only images. After exclusion of five patients due to poor scan quality or presence of severe artifacts, manual annotations were performed on 24 forearms from 23 patients. This set included 12 affected and 12 unaffected forearms, with an equal split between left and right to ensure diversity in the data.

Annotations of the radius, ulna, interosseous membrane, m. pronator quadratus, m, pronator teres and m. supinator were performed on axial in-phase images using 3D Slicer 5.2.1 [47]. Out-of-phase images were not eligible for this purpose due to chemical shift artifacts; however, they, along with water-only images, were occasionally referenced for guidance during the annotation process. Fat-only images were specifically used to minimize the inclusion of extramuscular fat during the annotation process. Given that all structures of interest are relatively elongated, segmentation was performed by annotating every other slice (i.e. annotating one slice and skipping the next), with the remaining slices filled in using interslice interpolation. The origin of the m. supinator on the distal humerus was not annotated as this region was not clearly visible on the scans. The interosseous membrane was annotated as a single continuous structure without differentiating its individual ligaments, as the relatively high slice thickness obscured finer details. Sagittal and coronal image representations were referenced to validate the axially annotated slices, ensuring proper alignment and accuracy of the annotations across all planes. A 2D median smoothing filter with a kernel size of 3.00 mm was applied to acquire smooth segment contours. To ensure quality and consistency, all 24 annotated forearms were reviewed and validated by an experienced musculoskeletal radiologist. A visual representation of the manual annotation procedure is provided in Figure 1.

The annotated dataset was converted to NIfTI format and divided into two subsets. The training set consisted of 20 manually segmented forearms, which was used in the training process of the networks. The test set, containing 4 forearms, was reserved exclusively for a final evaluation of the networks upon completion of the training process. Both sets maintained an equal distribution of affected and unaffected, as well as left and right forearms.

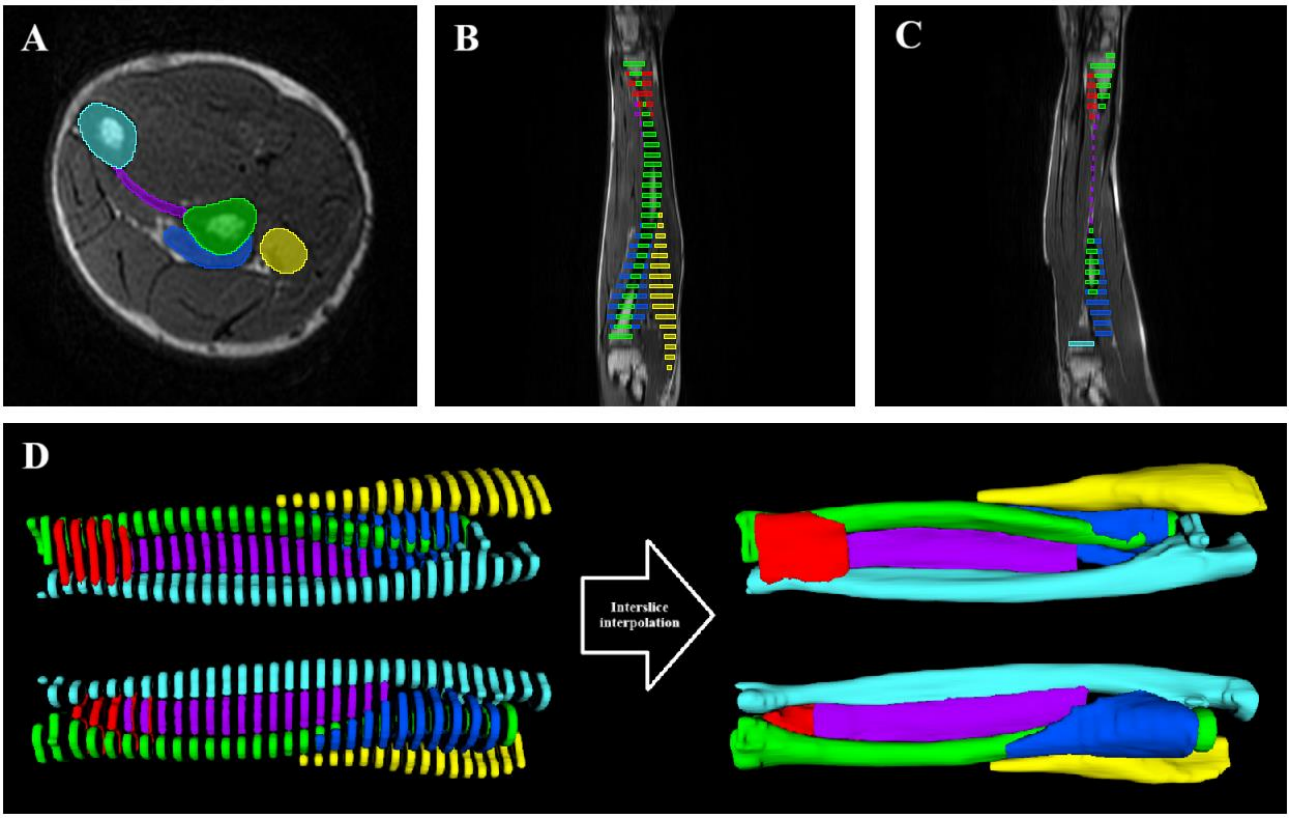


Figure 1. Manual annotation procedure. (A) Axial in-phase images were annotated at regular intervals. (B) Sagittal and (C) coronal image reconstructions were referenced to validate the axially annotated slices. (D) Interslice interpolation was used to fill in the remaining slices, resulting in the final segmentation mask. Green: radius, light blue: ulna, purple: interosseous membrane, red: m. pronator quadratus, yellow: m. pronator teres, dark blue: m. supinator.

2.2 Experiments and validation

To segment the forearm structures, two different configurations of nnU-Net were used. The first was a 2D network, where input and output are two-dimensional arrays, and convolution operations are performed in 2D. The second was a 3D full-resolution network, which processes volumetric data and performs convolution operations in 3D. As mentioned earlier, nnU-Net uses data-specific properties and a set of heuristic rules to infer data-dependent parameters. These, along with a set of fixed parameters, are used to create the pipeline fingerprints for both networks. A flowchart of the configuration pipeline is provided in Figure 2.

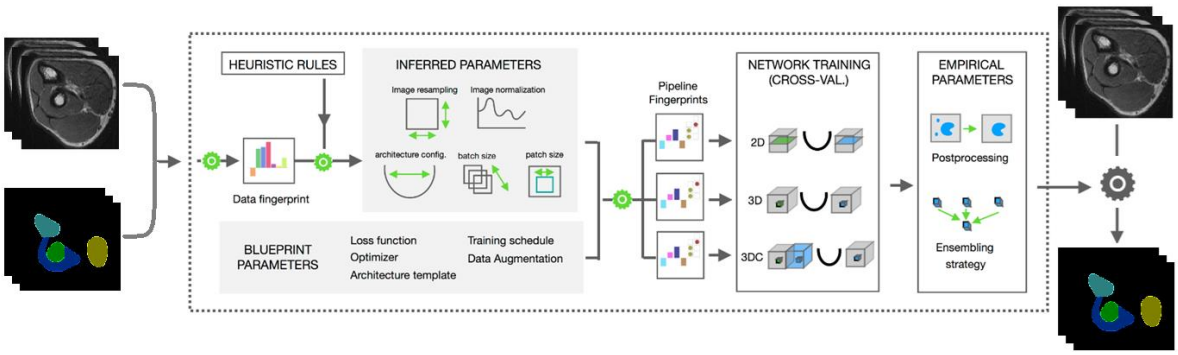


Figure 2. Flowchart of the automated nnU-Net configuration pipeline. A data fingerprint with data-specific properties is extracted from the image data. Based on this fingerprint and a set of heuristic rules, data-dependent parameters are automatically inferred to match the data. Along with a set of fixed parameters, the pipeline fingerprints are created. The 3D cascade network (3DC) was not considered in this study.

The 2D and 3D networks were trained for a fixed length of 1000 epochs, each consisting of 250 iterations. Patch sizes of 512x512 and 40x224x192 were used for 2D and 3D training, respectively, with batch sizes set to 12 and 2. Both models used the stochastic gradient descent (SGD) optimizer with Nesterov momentum of 0.99. The initial learning rate was 0.1, which was adjusted by a polynomial learning rate decay scheme (PolyLR). A combined cross-entropy and Dice loss function was used to optimize the models. To enhance generalization, data augmentation was used on the fly during training, including rotation, scaling, Gaussian noise and blur, brightness, contrast augmentation, simulation of low resolution, gamma correction, and mirroring. Intensity normalization was performed using a z-score normalization scheme. The models were trained for multi-class segmentation (six anatomical structures + background) using a 5-fold cross-validation scheme (i.e. five models were trained using 4/5 of the training set, with performance evaluated on the remaining 1/5). The best weights, minimizing the loss for each fold, were retained as the final weights for the model corresponding to that fold. An ensemble was created by combining predictions from both fully-trained models to determine whether the collective strengths of both models contributed to improved segmentation results when combined. Post-processing steps were applied to the 20 predicted segmentation masks resulting from the five validation sets (or ensembling). This included the process of removing all but the largest component for a specific ROI if it resulted in an improved DSC score. If this process improved the average DSC for a specific ROI, it was retained during inference.

The segmentation performance of both networks (2D and 3D) and their ensemble were evaluated on the test dataset to provide insight into how they performed on unseen data. For the 2D and 3D model inference, the five models resulting from training five different folds were used as a natural model ensemble for predicting test cases. The same pre- and post-processing steps that were applied during training were also applied during the inference process.

2.3 Evaluation metrics

To evaluate segmentation performance, the DSC was used as the primary metric [48]. The DSC is a widely used overlap-based metric that quantifies the proportion of overlap between the predicted and ground truth segmentations, and is calculated as shown in Equation 1. However, since single-pixel differences can significantly impact the metric's outcome, especially in smaller structures [49], the average symmetric surface distance (ASSD) metric was also used [50]. The ASSD is a distance-based metric that measures the average of shortest distances from each point on the surface of the predicted segmentation to the closest point on the surface of the ground truth segmentation, and vice versa. This symmetric approach ensures that the metric is not biased towards one surface. It is defined as shown in Equation 2. ASSD values were calculated in Python using the seg-metrics package developed by Jia et al [51].

Additionally, the volumes of the ground truth and predicted segmentation for each ROI were compared. The relative difference (Δ_{rel}) was computed to quantify fractional the difference between both volumes, providing insight into the extent and direction of potential under- or oversegmentation.

All quantitative results are reported as median values with interquartile ranges (IQR) to ensure robustness to potential outliers. This approach provides a more reliable summary of central tendency and variability in small sample settings, where individual deviations can disproportionately influence mean-based statistics.

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

Where sets A and B represent the ground truth and predicted segmentations, respectively, TP is the true positive count, FP is the false positive count, and FN is the false negative count. The DSC ranges from 0, indicating no overlap between the sets, to 1, indicating complete overlap.

$$ASSD(A,B) = \frac{1}{|A| + |B|}\left(\sum_{a \in A} \min_{b \in B} d(a,b) + \sum_{b \in B} \min_{a \in A} d(b,a)\right) \tag{2}$$

Where a and b are points from sets A and B, representing the ground truth and predicted segmentations, respectively, and d is the Euclidian distance between points a and b. A smaller ASSD value indicates more overlap between the sets. ASSD is measured in millimeters.

3. Results

3.1 Quantitative results

Table 1 summarizes the median DSC and ASSD scores on the unseen test dataset (n=4), along with IQR values, for all separate ROIs across the three networks. Detailed results are provided in Appendix II. On the test set, the 3D network achieved the highest overall median DSC score of 0.894 (IQR=0.094), with individual ROI scores of 0.949 (IQR=0.006) for the radius, 0.942 (IQR=0.002) for the ulna, 0.739 (IQR=0.140) for the interosseous membrane, 0.855 (IQR=0.014) for the m. pronator quadratus, 0.896 (IQR=0.013) for the m. pronator teres and 0.875 (IQR=0.048) for the m. supinator. The 3D network achieved the highest median DSC for all ROIs except the interosseous membrane, where the ensemble performed slightly better with a score of 0.744 (IQR=0.129). Cross-validation DSC results are provided in Appendix I.

Similar to the DSC, the 3D network also achieved the lowest overall median ASSD of 0.324 (IQR=0.386) mm, with individual ROI scores of 0.116 (IQR=0.038) mm for the radius, 0.254 (IQR=0.120) mm for the ulna, 0.426 (IQR=0.725) mm for the interosseous membrane, 0.585 (IQR=0.154) mm for the m. pronator quadratus, 0.367 (IQR=0.221) mm for the m. pronator teres and 0.441 (IQR=0.311) mm for the m. supinator. It outperformed the 2D network and the ensemble for all ROIs, except for the m. supinator, where the ensemble achieved a slightly lower ASSD of 0.436 (IQR=0.264) mm.

Table 1. Dice similarity coefficient and average symmetric surface distance results on the test dataset, quantifying the level of agreement between ground truth and predicted segmentations.

	2D		3D		Ensemble	
ROI	**DSC**	**ASSD (mm)**	**DSC**	**ASSD (mm)**	**DSC**	**ASSD (mm)**
Radius	0.941 (0.008)	0.202 (0.138)	**0.949 (0.006)**	**0.116 (0.038)**	0.946 (0.003)	0.147 (0.073)
Ulna	0.928 (0.017)	0.458 (0.240)	**0.942 (0.002)**	**0.254 (0.120)**	0.940 (0.009)	0.298 (0.077)
IOM	0.737 (0.122)	0.459 (0.690)	0.739 (0.140)	**0.426 (0.725)**	**0.744 (0.129)**	0.438 (0.702)
MPQ	0.841 (0.024)	0.614 (0.305)	**0.855 (0.014)**	**0.585 (0.154)**	0.849 (0.009)	0.587 (0.114)
MPT	0.877 (0.019)	0.619 (0.229)	**0.896 (0.013)**	**0.367 (0.221)**	0.884 (0.010)	0.443 (0.212)
MS	0.864 (0.038)	0.520 (0.214)	**0.875 (0.048)**	0.441 (0.311)	0.869 (0.047)	**0.436 (0.264)**
Total	0.873 (0.096)	0.502 (0.380)	**0.894 (0.094)**	**0.324 (0.386)**	0.884 (0.095)	0.371 (0.364)

Note: Results are presented as median (IQR). The best performance for each metric, per ROI and for the total, across the three networks is highlighted in bold. DSC=Dice similarity coefficient, ASSD=average symmetric surface distance, IQR=interquartile range, ROI=region of interest, IOM=interosseous membrane, MPQ=m. pronator quadratus, MPT=m. pronator teres, MS=m. supinator.

Based on the median relative volume difference values, the 2D, 3D and ensemble networks undersegmented the anatomy with -5.5% (IQR=11.4%), -2.7% (IQR=7.1%) and -3.9% (IQR=9.0%), respectively. The 3D network achieved the lowest overall absolute score, indicating the best performance related to volume estimation. It achieved the lowest absolute score for the radius (+0.5%, IQR=1.3%), the m. pronator teres (-5.1%, IQR=7.0%) and the m. supinator (-8.6%, IQR=16.0%). The ensemble showed the best performance for the ulna (-3.0%, IQR=2.4%) and the interosseous membrane (-2.6%, IQR=4.4%), while the 2D network did for the m. pronator quadratus (-7.2%, IQR=19.7%). These results are summarized in Table 2, with detailed results provided in Appendix III.

Table 2. Relative volume difference scores quantifying the fractional difference between ground truth and predicted segmentation volumes of the test dataset.

	2D	3D	Ensemble
ROI	**Δ_{rel}**	**Δ_{rel}**	**Δ_{rel}**
Radius	-0.014 (0.029)	**+0.005 (0.013)**	-0.007 (0.012)
Ulna	-0.044 (0.049)	-0.032 (0.031)	**-0.030 (0.024)**
IOM	-0.040 (0.067)	-0.028 (0.071)	**-0.026 (0.044)**
MPQ	**-0.072 (0.197)**	-0.122 (0.134)	-0.107 (0.144)
MPT	-0.130 (0.031)	**-0.051 (0.070)**	-0.105 (0.052)
MS	-0.138 (0.101)	**-0.086 (0.160)**	-0.125 (0.126)
Total	-0.055 (0.114)	**-0.027 (0.071)**	-0.039 (0.090)

Note: Results are presented as median (IQR). The lowest absolute value, per ROI and for the total, across the three networks is highlighted in bold. Δ_{rel}=relative volume difference, IQR=interquartile range, ROI=region of interest, IOM=interosseous membrane, MPQ=m. pronator quadratus, MPT=m. pronator teres, MS=m. supinator.

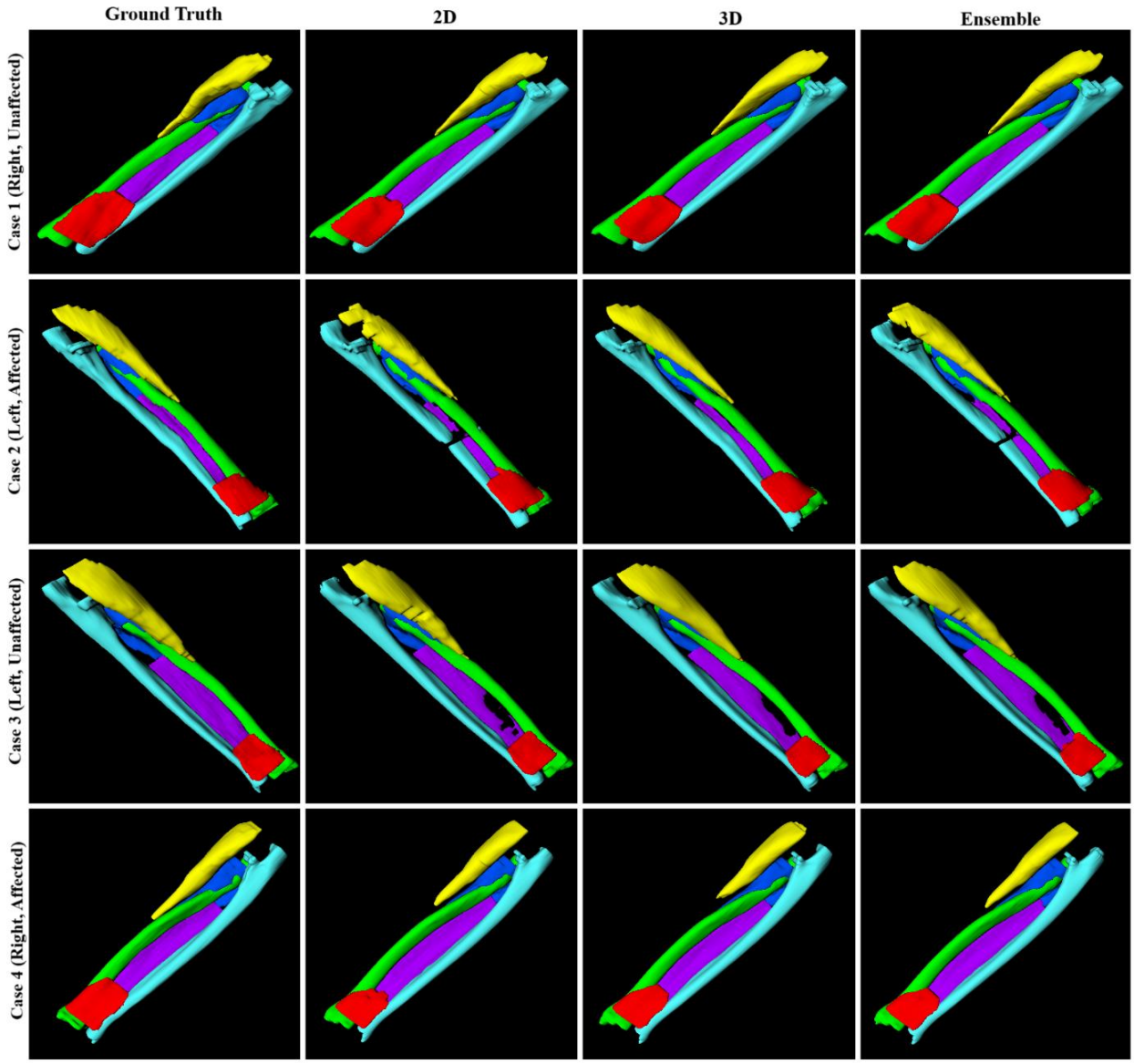


Figure 3. Segmented cases from the test dataset. The first column shows the manually annotated ground truth segmentation masks, while the second, third and fourth columns display the segmentation masks predicted by the 2D model, 3D model, and ensemble, respectively. Green: radius, light blue: ulna, purple: interosseous membrane, red: m. pronator quadratus, yellow: m. pronator teres, dark blue: m. supinator.

3.2 Qualitative results

Figure 3 displays the segmented test cases predicted by the three different networks, along with the corresponding ground truth segmentations. Overall, the masks predicted by the 3D model showed the fewest and least severe segmentation errors compared to those generated by the 2D model and ensemble, which is consistent with the quantitative results. This is particularly evident in case 2 (left, affected forearm), where the 2D model and ensemble contained significant segmentation errors in multiple structures. While all three models made notable errors in segmenting the interosseous membrane in two cases, the mistakes in the 3D model's predictions were less pronounced and more confined, containing fewer irregularities. Additionally, the three models exhibited minor difficulties in accurately segmenting the proximal part of the m. pronator quadratus and the insertion of the m. pronator teres in some cases, with some inconsistencies across the predictions. The models were able to effectively exclude extramuscular fat tissue during the segmentation process. Figure 4 shows example slices from a test case segmented by the 3D model, shown alongside the corresponding ground truth slices.

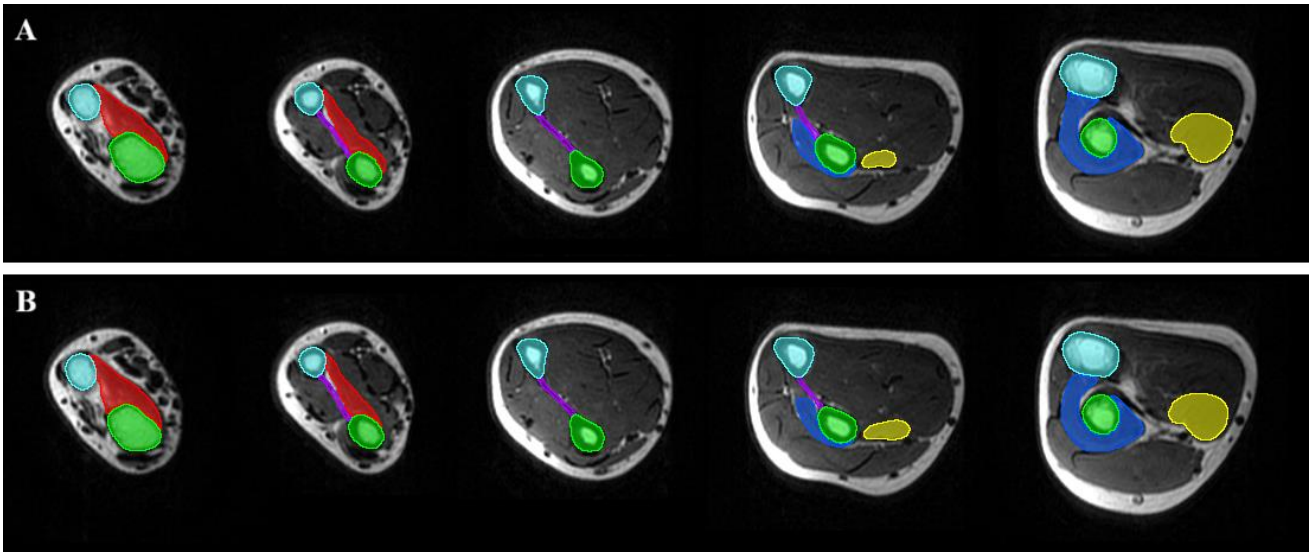


Figure 4. Example segmentation from the test dataset (case 1: right, unaffected forearm). (A) Five representative cross-sectional slices shown from distal to proximal, illustrating the manually segmented ground truth. (B) Corresponding slices segmented by the 3D nnU-Net model, shown in the same order. Green: radius, light blue: ulna, purple: interosseous membrane, red: m. pronator quadratus, yellow: m. pronator teres, dark blue: m. supinator.

4. Discussion

With its robust capabilities, nnU-Net has proven highly effective in segmenting the forearm anatomy both affected and unaffected by trauma. To the best of our knowledge, this study marks the first successful application of these methods to the forearm, representing a significant breakthrough in the literature for this specific anatomical region. Compared to its 2D counterpart, which offers lower computational and memory demands but suffers from discontinuity along the z-axis, the 3D model captures spatial patterns across the full volume of the image, providing a more comprehensive representation of anatomical structures [52]. Although ensembling of the two networks was explored, it did not improve either quantitative or qualitative performance. As a result, the 3D nnU-Net model was identified as the most suitable configuration for clinical implementation. Compared to the time-consuming and labor-intensive manual approach, which requires at least six hours, this automated method demonstrated sufficiently accurate segmentation within seconds. This enables bilateral comparisons of soft tissue structures through visual assessment and quantitative analysis, based on properties such as volume, length and texture-based features. Based on these analyses, it may be possible to identify whether impairment of forearm rotation is potentially caused by soft tissue alterations rather than bony malunion. In such cases, invasive corrective osteotomies could probably be avoided if treatment is focused primarily on soft tissue pathology, such as through physiotherapy and/or bracing.

The 3D model achieved median DSC scores of 0.949 (IQR=0.006) and 0.942 (IQR=0.002) for segmentation of the radius and ulna, respectively. These were the highest DSC values among all segmented structures, indicating the model's particularly strong performance on the bones. These results are in line with previously reported values in the literature. For example, Kamphuis et al. used an ensemble of 2D and 3D nnU-Nets to segment femoral and acetabular bone in the hip joint, along with cartilage, using Dixon sequence images [46]. Their work compared the performance on different image combinations, including water-only, in-phase plus water-only, and fat-only plus water-only images. They reported DSC scores of 0.961, 0.967 and 0.950 for segmentation of femoral bone, respectively, and 0.886, 0.893 and 0.896 for segmentation of acetabular bone. Hess et al. applied 3D nnU-Net to segment shoulder bones on T1-weighted images, achieving a DSC of 0.945 [44], while Kim et al. reported 0.987 using T2-weighted images [45]. These results are comparable to the DSC scores for segmentation of the forearm bones achieved in the present study, reinforcing the effectiveness of nnU-Net in segmenting osseous structures.

For segmentation of the three muscles, the 3D model achieved the following median DSC scores: 0.855 (IQR=0.014) for the m. pronator quadratus, 0.896 (IQR=0.013) for the m. pronator teres and 0.875 (IQR=0.048) for the m. supinator. Reported DSC values for muscle segmentation vary considerably across the literature. In the previously mentioned studies, Hess et al. achieved a DSC of 0.900 for segmentation of the shoulder muscles [44], while Kim et al. reported 0.797 [45]. Alipour et al. and Medina et al. used conventional 2D U-Net models and T1-weighted images for this purpose, and obtained DSC scores of 0.810 [34] and 0.963 [35], respectively. The results achieved in the present study

fall within this reported range, highlighting the comparability and robustness of the 3D nnU-Net approach for muscle segmentation. The relatively lower score of the m. pronator quadratus may be attributed to two possible factors. First, the proximal part of this muscle was often difficult to visualize, which led to the manual annotation of this area requiring some degree of estimation. As a result, this may have introduced inconsistencies into the training data. Second, as mentioned earlier, the DSC metric is sensitive to small discrepancies, and even minor deviations can significantly affect scores in smaller structures. However, the ASSD and relative volume difference results provide further insight, with this muscle achieving the poorest scores across both, consistent with qualitative observations where it appeared slightly undersized in some cases. Another issue, not clearly reflected in quantitative results, involved the insertion of the m. pronator teres. Like the proximal m. pronator quadratus, this region was difficult to annotate accurately and frequently showed incomplete or imprecise segmentation in qualitative assessments, despite minimal impact on overall scores.

Segmentation of the interosseous membrane yielded a median DSC score of 0.739 (IQR=0.140), the lowest among the six structures. Although lower than the scores obtained for segmentation of bones and muscles, it still exceeds the commonly cited threshold of 0.700 for acceptable segmentation overlap [53]. This relatively lower performance is also consistent with previous studies reporting lower DSC scores for fibrous structures compared to osseous and muscular structures. For example, Kim et al. achieved a DSC of 0.801 for segmentation of rotator cuff tendons in the study mentioned earlier [45], while Flannery et al. achieved 0.760 for segmentation of the ACL on T2* images using a 2D U-Net [33]. Similar to the proximal m. pronator quadratus and the insertion region of the m. pronator teres, the structure itself was more challenging to manually segment, as it was occasionally difficult to visualize on individual slices. This may again have affected the quality of the training data. Additionally, the main limitation of the DSC metric discussed earlier may also help explain the lower score. Although the ASSD and relative volume difference scores did not rank the interosseous membrane as the lowest, the qualitative results revealed that the model faced challenges in accurately segmenting this structure, as evidenced by incomplete segments with visible gaps in two cases. It is worth noting, however, that these gaps appeared in regions where the structure was also difficult to annotate manually. Despite the inaccuracies, the overall volumes were still reasonably well matched, explaining the low relative volume difference. As for the ASSD, similar to the DSC, this structure had an IQR that was significantly greater than that of the others, indicating greater variability in segmentation performance, which helps explain the difference in its ranking across both metrics.

While the 3D network consistently achieved the highest DSC results for segmentation of most structures, its median score of 0.739 (IQR=0.140) for the interosseous membrane was slightly lower than that of the ensemble, which reached 0.744 (IQR=0.129). However, this structure exhibited the largest DSC IQR across all six structures in both configurations, suggesting that the segmentation results were less consistent. Similarly, for the ASSD metric, the 3D network achieved the lowest scores for all structures except one. It reached an ASSD of 0.441 (IQR=0.311) for the m. supinator, while the ensemble obtained a slightly lower 0.436 (IQR=0.264). This structure had the second largest IQR for ASSD among all six structures in both networks. Although these differences are worth noting, they are minimal and may be coincidental. As no formal statistical test was performed, their significance remains uncertain. This also applies to the relative volume difference results, where the 3D model did not achieve the lowest absolute scores for the ulna, interosseous membrane and m. pronator quadratus.

The second case in the test set involved an affected left forearm. Both the quantitative and qualitative results demonstrated that all models faced difficulties in accurately segmenting the anatomy in this case. This was particularly evident in the 2D and ensemble networks, where large holes were present in the interosseous membrane and m. pronator teres, and even the radius and ulna contained visible segmentation errors. This case involved an arm which was severely damaged by trauma, resulting in anatomical changes that may not have been well-represented in the training set. The scan also contained some motion artifacts, though not severe enough to warrant exclusion. While the 3D model achieved better results for this case than the 2D and ensemble approaches, a noticeable gap remains in the interosseous membrane. This observation may suggest that this model could potentially underperform for segmentation of this structure in cases with more extensive anatomical changes due to trauma. However, a visible hole was also observed in the third case, which involved an unaffected left arm. As a result, it remains unclear whether segmentation inaccuracies are due to anatomical alterations caused by trauma.

This work has a number of limitations. Although internal validation yielded satisfactory results, no external validation was performed using an independent test dataset. External validation would have provided further insight into the model's ability to generalize beyond the specific dataset used for training, ensuring its applicability in other clinical contexts. Two additional limitations arise from the small size of the (test) dataset. First, no statistical analyses were performed to determine whether the segmentation performances of the three networks significantly differed from each

other. While the 3D model consistently achieved the best median metric scores across most ROIs, it did not outperform the other two in some cases. Although the differences are minimal and may be coincidental, it is difficult to draw definitive conclusions regarding their significance without statistical testing. The small size of the test dataset also limited the ability to compare segmentation performance between affected and unaffected forearms, as noted earlier. This distinction would have been particularly valuable for clinical implementation, as it could offer insight into the potential inaccuracies specific to each side. Future work should focus on expanding the dataset, particularly by increasing the number of test cases. Additionally, incorporating a broader range of anatomical presentations, such as cases with congenital abnormalities or post-surgical changes, could enhance the model's generalizability and clinical applicability. To further strengthen the findings, external validation and statistical analyses should be considered.

5. Conclusion

In this study, nnU-Net was employed as a deep-learning based method for automated segmentation of anatomical structures involved in pronation/supination of the forearm on MR images. Among the three different configurations compared, the 3D network achieved the most accurate results. The model demonstrated the ability to perform fast, reproducible and precise segmentation, offering significant potential for efficient visualization and quantitative analysis of the forearm anatomy. This approach enables patient-specific and minimally invasive decision-making by differentiating soft tissue from osseous pathology as (main) cause for impaired forearm rotation. Given the limited focus on the forearm in existing literature, this work serves as a pioneering effort, laying the foundation for future advancements in both clinical practice and research.

References

1. Rafi, B. M., Tiwari, V. (2023). Forearm Fractures. In *StatPearls*. StatPearls Publishing.
2. Colaris, J., van der Linden, M., Selles, R., Coene, N., Allema, J. H., Verhaar, J. (2010). Pronation and supination after forearm fractures in children: Reliability of visual estimation and conventional goniometry measurement. *Injury*, *41*(6), 643–646.
3. Nakamura, T., Yabe, Y., Horiuchi, Y., Yamazaki, N. (1999). In vivo motion analysis of forearm rotation utilizing magnetic resonance imaging. *Clinical biomechanics (Bristol, Avon)*, *14*(5), 315–320.
4. Valone, L. C., Waites, C., Tartarilla, A. B., Whited, A., Sugimoto, D., Bae, D. S., Bauer, A. S. (2020). Functional Elbow Range of Motion in Children and Adolescents. *Journal of pediatric orthopedics*, *40*(6), 304–309.
5. Bronstein, A., Heaton, D., Tencer, A. F., Trumble, T. E. (2014). Distal radius malunion and forearm rotation: a cadaveric study. *Journal of wrist surgery*, *3*(1), 7–11.
6. Dumont, C. E., Thalmann, R., Macy, J. C. (2002). The effect of rotational malunion of the radius and the ulna on supination and pronation. *The Journal of bone and joint surgery. British volume*, *84*(7), 1070–1074.
7. Matthews, L. S., Kaufer, H., Garver, D. F., Sonstegard, D. A. (1982). The effect on supination-pronation of angular malalignment of fractures of both bones of the forearm. *The Journal of bone and joint surgery. American volume*, *64*(1), 14–17.
8. Price, C. T., Scott, D. S., Kurzner, M. E., Flynn, J. C. (1990). Malunited forearm fractures in children. *Journal of pediatric orthopedics*, *10*(6), 705–712.
9. Roberts J. A. (1986). Angulation of the radius in children's fractures. *The Journal of bone and joint surgery. British volume*, *68*(5), 751–754.
10. Trousdale, R. T., Linscheid, R. L. (1995). Operative treatment of malunited fractures of the forearm. *The Journal of bone and joint surgery. American volume*, *77*(6), 894–902.
11. Tarr, R. R., Garfinkel, A. I., Sarmiento, A. (1984). The effects of angular and rotational deformities of both bones of the forearm. An in vitro study. *The Journal of bone and joint surgery. American volume*, *66*(1), 65–70.
12. Högström, H., Nilsson, B. E., Willner, S. (1976). Correction with growth following diaphyseal forearm fracture. *Acta orthopaedica Scandinavica*, *47*(3), 299–303.
13. Nilsson, B. E., Obrant, K. (1977). The range of motion following fracture of the shaft of the forearm in children. *Acta orthopaedica Scandinavica*, *48*(6), 600–602.
14. Krukhaug, Y., Hove, L. M. (2007). Corrective osteotomy for malunited extra-articular fractures of the distal radius: a follow-up study of 33 patients. *Scandinavian journal of plastic and reconstructive surgery and hand surgery*, *41*(6), 303–309.
15. Nagy, L., Jankauskas, L., Dumont, C. E. (2008). Correction of forearm malunion guided by the preoperative complaint. *Clinical orthopaedics and related research*, *466*(6), 1419–1428.
16. Colaris, J. W., Allema, J. H., Reijman, M., de Vries, M. R., Ulas Biter, L., Bloem, R. M., van de Ven, C. P., Verhaar, J. A. (2014). Which factors affect limitation of pronation/supination after forearm fractures in children? A prospective multicentre study. *Injury*, *45*(4), 696–700.
17. Nakamura, T., Yabe, Y., Horiuchi, Y. (1994). A biomechanical analysis of pronation-supination of the forearm using magnetic resonance imaging: dynamic changes of the interosseous membrane of the forearm during pronation–supination. *Nihon Seikeigeka Gakkai zasshi*, *68*(1), 14–25.
18. Kleinman, W. B., Graham, T. J. (1998). The distal radioulnar joint capsule: clinical anatomy and role in posttraumatic limitation of forearm rotation. *The Journal of hand surgery*, *23*(4), 588–599.
19. Philips, T., Duerinckx, J., Van Melkebeke, L., van Riet, R., Caekebeke, P. (2024). The pronator contracture syndrome: A new entity in supination restriction. *Shoulder & Elbow. 0(0).*
20. Bert, J. M., Linscheid, R. L., McElfresh, E. C. (1980). Rotatory contracture of the forearm. *The Journal of bone and joint surgery. American volume*, *62*(7), 1163–1168.
21. Haines, S. C., Bott, A. (2023). Current Concepts: Corrective Osteotomy for Extra-Articular Deformity Following a Distal Radius Fracture. *Cureus*, *15*(10), e47019.
22. van Es, E. M., Dijkhof, M., Souer, J. S., van Ewijk, F. J., Hoogendam, L., Slijper, H. P., Selles, R. W., Colaris, J. W.; Hand-Wrist Study Group (2024). Forearm rotation improves after corrective osteotomy in patients with symptomatic distal radius malunion. *Heliyon*, *10*(9), e29570.
23. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G. (2018). Recent Advances in Convolutional Neural Networks. *Pattern Recognition, 77*, 354-377.
24. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *LNCS, 9351*, 234-241.
25. Almajalid, R., Zhang, M., Shan, J. (2022). Fully Automatic Knee Bone Detection and Segmentation on Three-Dimensional MRI. *Diagnostics (Basel, Switzerland)*, *12*(1), 123.
26. Kim-Wang, S. Y., Bradley, P. X., Cutcliffe, H. C., Collins, A. T., Crook, B. S., Paranjape, C. S., Spritzer, C. E., DeFrate, L. E. (2023). Auto-segmentation of the tibia and femur from knee MR images via deep learning and its application to cartilage strain and recovery. *Journal of biomechanics*, *149*, 111473.
27. Deng, Y., You, L., Wang, Y., Zhou, X. (2021). A Coarse-to-Fine Framework for Automated Knee Bone and Cartilage Segmentation Data from the Osteoarthritis Initiative. *Journal of digital imaging*, *34*(4), 833–840.
28. Latif, M. H. A., Faye, I. (2021). Automated tibiofemoral joint segmentation based on deeply supervised 2D-3D ensemble U-Net: Data from the Osteoarthritis Initiative. *Artificial intelligence in medicine*, *122*, 102213.
29. Ambellan, F., Tack, A., Ehlke, M., Zachow, S. (2019). Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative. *Medical image analysis*, *52*, 109–118.

30. Kemnitz, J., Baumgartner, C. F., Eckstein, F., Chaudhari, A., Ruhdorfer, A., Wirth, W., Eder, S. K., Konukoglu, E. (2020). Clinical evaluation of fully automated thigh muscle and adipose tissue segmentation using a U-Net deep learning architecture in context of osteoarthritic knee pain. *Magma (New York, N.Y.)*, *33*(4), 483–493.
31. Flannery, S. W., Kiapour, A. M., Edgar, D. J., Murray, M. M., Fleming, B. C. (2021). Automated magnetic resonance image segmentation of the anterior cruciate ligament. *Journal of orthopaedic research: official publication of the Orthopaedic Research Society*, *39*(4), 831–840.
32. Flannery, S. W., Kiapour, A. M., Edgar, D. J., Murray, M. M., Beveridge, J. E., Fleming, B. C. (2022). A transfer learning approach for automatic segmentation of the surgically treated anterior cruciate ligament. *Journal of orthopaedic research: official publication of the Orthopaedic Research Society*, *40*(1), 277–284.
33. Flannery, S. W., Barnes, D. A., Costa, M. Q., Menghini, D., Kiapour, A. M., Walsh, E. G., Bear Trial Team, Kramer, D. E., Murray, M. M., Fleming, B. C. (2023). Automated segmentation of the healed anterior cruciate ligament from T2* relaxometry MRI scans. *Journal of orthopaedic research: official publication of the Orthopaedic Research Society*, *41*(3), 649–656.
34. Alipour, E., Chalian, M., Pooyan, A., Azhideh, A., Shomal Zadeh, F., Jahanian, H. (2024). Automatic MRI-based rotator cuff muscle segmentation using U-Nets. *Skeletal radiology*, *53*(3), 537–545.
35. Medina, G., Buckless, C. G., Thomasson, E., Oh, L. S., Torriani, M. (2021). Deep learning method for segmentation of rotator cuff muscles on MR images. *Skeletal radiology*, *50*(4), 683–692.
36. Riem, L., Feng, X., Cousins, M., DuCharme, O., Leitch, E. B., Werner, B. C., Sheean, A. J., Hart, J., Antosh, I. J., Blemker, S. S. (2023). A Deep Learning Algorithm for Automatic 3D Segmentation of Rotator Cuff Muscle and Fat from Clinical MRI Scans. *Radiology. Artificial intelligence*, *5*(2), e220132.
37. Zhu, Z., Liu, E., Su, Z., Chen, W., Liu, Z., Chen, T., Lu, H., Zhou, J., Li, Q., Pang, S. (2024). Three-Dimensional Lumbosacral Reconstruction by An Artificial Intelligence-Based Automated MR Image Segmentation for Selecting the Approach of Percutaneous Endoscopic Lumbar Discectomy. *Pain physician*, *27*(2), E245–E254.
38. Chen, T., Su, Z. H., Liu, Z., Wang, M., Cui, Z. F., Zhao, L., Yang, L. J., Zhang, W. C., Liu, X., Liu, J., Tan, S. Y., Li, S. L., Feng, Q. J., Pang, S. M., Lu, H. (2022). Automated Magnetic Resonance Image Segmentation of Spinal Structures at the L4-5 Level with Deep Learning: 3D Reconstruction of Lumbar Intervertebral Foramen. *Orthopaedic surgery*, *14*(9), 2256–2264.
39. Su, Z., Liu, Z., Wang, M., Li, S., Lin, L., Yuan, Z., Pang, S., Feng, Q., Chen, T., Lu, H. (2022). Three-dimensional reconstruction of Kambin's triangle based on automated magnetic resonance image segmentation. *Journal of orthopaedic research: official publication of the Orthopaedic Research Society*, *40*(12), 2914–2923.
40. van der Graaf, J. W., van Hooff, M. L., Buckens, C. F. M., Rutten, M., van Susante, J. L. C., Kroeze, R. J., de Kleuver, M., van Ginneken, B., Lessmann, N. (2024). Lumbar spine segmentation in MR images: a dataset and a public benchmark. *Scientific data*, *11*(1), 264.
41. Wesselink, E. O., Elliott, J. M., Coppieters, M. W., Hancock, M. J., Cronin, B., Pool-Goudzwaard, A., Weber Ii, K. A. (2022). Convolutional neural networks for the automatic segmentation of lumbar paraspinal muscles in people with low back pain. *Scientific reports*, *12*(1), 13485.
42. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, *18*(2), 203–211.
43. Li, M., Punithakumar, K., Major, P. W., Le, L. H., Nguyen, K. T., Pacheco-Pereira, C., Kaipatur, N. R., Nebbe, B., Jaremko, J. L., Almeida, F. T. (2022). Temporomandibular joint segmentation in MRI images using deep learning. *Journal of dentistry*, *127*, 104345.
44. Hess, H., Ruckli, A. C., Bürki, F., Gerber, N., Menzemer, J., Burger, J., Schär, M., Zumstein, M. A., Gerber, K. (2023). Deep-Learning-Based Segmentation of the Shoulder from MRI with Inference Accuracy Prediction. *Diagnostics (Basel, Switzerland)*, *13*(10), 1668.
45. Kim, H., Shin, K., Kim, H., Lee, E. S., Chung, S. W., Koh, K. H., Kim, N. (2022). Can deep learning reduce the time and effort required for manual segmentation in 3D reconstruction of MRI in rotator cuff tears?. *PloS one*, *17*(10), e0274075.
46. Kamphuis, M.A., Oei, E.H.G., Runhaar, J., Hanff, D., Bierma-Zeinstra, S.M.A., Klein, S., Hirvasniemi, J. (2024). Enhancing Model Performance in Hip Joint Segmentation by Leveraging Multiple Image Outputs from Dixon MRI. *Osteoarthritis Imaging, 4*(1), 2772-6541.
47. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J. C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J. V., Pieper, S., Kikinis, R. (2012). 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic resonance imaging*, *30*(9), 1323–1341.
48. Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells, W. M., 3rd, Jolesz, F. A., Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index. *Academic radiology*, *11*(2), 178–189.
49. Reinke, A., Tizabi, MD., Sudre, C., Eisenmann, M., Rädsch, T., Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Bankhead, P., Benis, A., Cardoso, MJ., Cheplygina, V., Cimini, B., Collins, G., Farahani, K., Glocker, B., Godau, P., Noyan, A. (2022). Common Limitations of Image Processing Metrics: A Picture Story. *arXiv*. https://doi.org/10.48550/arXiv.2104.05642.
50. Yeghiazaryan, V., Voiculescu, I. (2018). Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of medical imaging (Bellingham, Wash.)*, *5*(1), 015006.
51. Jia, J., Staring, M., Stoel, B.C. (2024). Seg-metrics: a Python package to compute segmentation metrics. *arRxiv*. https://doi.org/10.48550/arXiv.2403.07884.
52. Hesamian, M. H., Jia, W., He, X., Kennedy, P. (2019). Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of digital imaging*, *32*(4), 582–596.
53. Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE transactions on medical imaging*, *13*(4), 716–724.

Appendix I. DSC results (cross-validation)

Supplemental Table 1. Mean Dice similarity coefficient results for each cross-validation fold, quantifying the average level of agreement between ground truth and predicted segmentations in each fold's validation set. Averaged results across all folds are included, both before and after post-processing.

		2D	3D	Ensemble
Fold	**ROI**	**DSC**	**DSC**	**DSC**
0	Radius	0.924	0.932	-
	Ulna	0.917	0.926	-
	IOM	0.678	0.666	-
	MPQ	0.813	0.845	-
	MPT	0.880	0.884	-
	MS	0.865	0.878	-
	Mean	0.846	0.855	-
1	Radius	0.921	0.928	-
	Ulna	0.917	0.924	-
	IOM	0.680	0.694	-
	MPQ	0.842	0.866	-
	MPT	0.868	0.888	-
	MS	0.841	0.858	-
	Mean	0.845	0.859	-
2	Radius	0.941	0.941	-
	Ulna	0.919	0.928	-
	IOM	0.737	0.736	-
	MPQ	0.874	0.864	-
	MPT	0.888	0.904	-
	MS	0.858	0.881	-
	Mean	0.870	0.876	-
3	Radius	0.915	0.931	-
	Ulna	0.914	0.920	-
	IOM	0.733	0.760	-
	MPQ	0.759	0.829	-
	MPT	0.855	0.882	-
	MS	0.855	0.873	-
	Mean	0.839	0.866	-
4	Radius	0.933	0.920	-
	Ulna	0.925	0.926	-
	IOM	0.709	0.745	-
	MPQ	0.832	0.857	-
	MPT	0.837	0.887	-
	MS	0.856	0.872	-
	Mean	0.849	0.868	-
All	Radius	0.927	0.930	0.932
	Ulna	0.918	0.924	0.925
	IOM	0.707	0.720	0.718
	MPQ	0.824	0.852	0.840
	MPT	0.866	0.889	0.876
	MS	0.855	0.872	0.866
	Mean	0.850	0.865	0.859
All (PP)	Radius	0.927	**0.934**	0.932
	Ulna	0.918	0.924	0.925
	IOM	0.707	**0.720**	0.718
	MPQ	0.824	0.852	**0.840**
	MPT	**0.866**	0.889	0.876
	MS	**0.855**	**0.872**	**0.866**
	Mean	**0.850**	**0.865**	**0.859**

Note: Values highlighted in bold indicate cases where post-processing improved the DSC score by removing all but the largest component. DSC=Dice similarity coefficient, ROI=region of interest, IOM=interosseous membrane, MPQ=m. pronator quadratus, MPT=m. pronator teres, MS=m. supinator, PP=post-processing.

Appendix II. DSC/ASSD results (inference)

Supplemental Table 2. Dice similarity coefficient and average symmetric surface distance results for each case in the test dataset, quantifying the level of agreement between ground truth and predicted segmentations.

				2D		3D		Ensemble	
Case	**A/U**	**L/R**	**ROI**	**DSC**	**ASSD (mm)**	**DSC**	**ASSD (mm)**	**DSC**	**ASSD (mm)**
1	U	R	Radius	0.946	0.132	0.951	0.110	0.950	0.115
			Ulna	0.944	0.166	0.942	0.175	0.945	0.163
			IOM	0.783	0.173	0.794	0.129	0.787	0.162
			MPQ	0.855	0.456	0.849	0.724	0.855	0.608
			MPT	0.872	0.436	0.897	0.308	0.884	0.369
			MS	0.853	0.549	0.842	0.583	0.851	0.534
			Median (IQR)	0.864 (0.073)	0.305 (0.283)	0.873 (0.087)	0.242 (0.374)	0.870 (0.078)	0.266 (0.331)
2	A	L	Radius	0.932	0.325	0.954	0.113	0.945	0.178
			Ulna	0.923	0.539	0.938	0.390	0.937	0.314
			IOM	0.508	1.865	0.567	1.675	0.533	1.714
			MPQ	0.833	0.714	0.864	0.519	0.851	0.566
			MPT	0.847	0.707	0.894	0.340	0.873	0.511
			MS	0.819	0.920	0.855	0.582	0.834	0.653
			Median (IQR)	0.840 (0.082)	0.711 (0.288)	0.879 (0.070)	0.455 (0.214)	0.862 (0.083)	0.539 (0.268)
3	U	L	Radius	0.944	0.142	0.947	0.118	0.947	0.116
			Ulna	0.933	0.377	0.943	0.298	0.942	0.281
			IOM	0.701	0.637	0.684	0.651	0.700	0.647
			MPQ	0.848	0.513	0.843	0.651	0.846	0.526
			MPT	0.895	0.530	0.921	0.394	0.914	0.374
			MS	0.874	0.491	0.894	0.300	0.886	0.337
			Median (IQR)	0.885 (0.069)	0.502 (0.120)	0.908 (0.082)	0.347 (0.288)	0.900 (0.079)	0.356 (0.193)
4	A	R	Radius	0.938	0.262	0.942	0.247	0.943	0.221
			Ulna	0.906	0.639	0.941	0.210	0.922	0.373
			IOM	0.772	0.281	0.795	0.200	0.787	0.228
			MPQ	0.805	1.072	0.860	0.506	0.833	0.855
			MPT	0.881	0.820	0.877	1.030	0.884	0.804
			MS	0.909	0.240	0.918	0.186	0.917	0.189
			Median (IQR)	0.894 (0.084)	0.460 (0.508)	0.898 (0.071)	0.229 (0.239)	0.901 (0.075)	0.301 (0.474)
			Total median (IQR)	0.873 (0.096)	0.502 (0.380)	0.894 (0.094)	0.324 (0.386)	0.884 (0.095)	0.371 (0.364)

Note: A/U=affected/unaffected, L/R=left/right, DSC=Dice similarity coefficient, ASSD=average symmetric surface distance, IQR=interquartile range, ROI=region of interest, IOM=interosseous membrane, MPQ=m. pronator quadratus, MPT=m. pronator teres, MS=m. supinator.

Appendix III. Relative volume difference results (inference)

Supplemental Table 3. Relative volume difference scores for each case in the test dataset, quantifying the fractional difference between ground truth and predicted volumes.

						2D		3D		Ensemble
Case	**A/U**	**L/R**	**ROI**	**GTV (cm^3)**	**PRV (cm^3)**	**Δ_{rel}**	**PRV (cm^3)**	**Δ_{rel}**	**PRV (cm^3)**	**Δ_{rel}**
1	U	R	Radius	26.503	26.388	-0.004	26.228	-0.010	26.311	-0.007
			Ulna	32.067	31.055	-0.032	31.505	-0.018	31.235	-0.026
			IOM	3.236	3.059	-0.055	3.001	-0.073	3.029	-0.064
			MPQ	6.913	5.839	-0.155	5.669	-0.180	5.790	-0.163
			MPT	15.411	14.568	-0.055	15.576	+0.011	14.884	-0.034
			MS	14.174	11.691	-0.175	11.609	-0.181	11.630	-0.180
			Median (IQR)	-	-	-0.055 (0.092)	-	-0.046 (0.141)	-	-0.049 (0.110)
2	A	L	Radius	37.511	36.461	-0.028	37.885	+0.010	37.264	-0.007
			Ulna	41.070	40.761	-0.008	39.162	-0.046	39.947	-0.027
			IOM	2.470	2.253	-0.088	2.899	+0.174	2.435	-0.014
			MPQ	7.293	7.375	+0.011	7.300	+0.001	7.327	+0.005
			MPT	19.308	16.872	-0.126	18.698	-0.032	17.633	-0.087
			MS	19.797	15.164	-0.234	16.599	-0.162	15.566	-0.214
			Median (IQR)	-	-	-0.058 (0.104)	-	-0.016 (0.050)	-	-0.021 (0.063)
3	U	L	Radius	32.686	33.664	+0.030	33.053	+0.011	33.505	+0.025
			Ulna	39.964	37.772	-0.055	39.228	-0.018	38.637	-0.033
			IOM	4.226	4.123	-0.024	4.081	-0.034	4.064	-0.038
			MPQ	6.271	6.367	+0.015	5.867	-0.064	5.949	-0.051
			MPT	25.971	21.856	-0.158	24.155	-0.070	22.802	-0.122
			MS	18.574	16.717	-0.100	18.645	+0.004	17.276	-0.070
			Median (IQR)	-	-	-0.040 (0.094)	-	-0.026 (0.055)	-	-0.045 (0.031)
4	A	R	Radius	36.036	35.215	-0.023	36.045	0.000	35.281	-0.021
			Ulna	40.410	34.998	-0.134	38.015	-0.059	36.179	-0.105
			IOM	4.723	5.129	+0.086	4.618	-0.022	4.913	+0.040
			MPQ	6.883	4.988	-0.275	5.597	-0.187	5.260	-0.236
			MPT	23.835	20.671	-0.133	20.176	-0.154	20.553	-0.138
			MS	18.583	17.548	-0.056	18.400	-0.010	17.859	-0.039
			Median (IQR)	-	-	-0.095 (0.103)	-	-0.041 (0.117)	-	-0.072 (0.104)
			Total median (IQR)	-	-	-0.055 (0.114)	-	-0.027 (0.071)	-	-0.039 (0.090)

Note: A/U=affected/unaffected, L/R=left/right, GTV=ground truth volume, PRV=predicted volume, Δ_{rel}=relative volume difference, IQR=interquartile range, ROI=region of interest, IOM=interosseous membrane, MPQ=m. pronator quadratus, MPT=m. pronator teres, MS=m. supinator.