EPiC Series in Health Sciences
Volume 8, 2026, Pages 134–138
Proceedings of The 25th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

# Expert Validation of CT-Based Machine Learning Model for Segmentation and Quantification of Deltoid Muscles for Shoulder Arthroplasty

Hamidreza Rajabzadeh-Oghaz[1], Josie Elwell[1], Francois Boux de Casson[2], Sandrine Polakovic[2], Ashish Singh[1], Rakesh Raushan[1], Likitha Shetty[1], Bradley Schoch[3], William Aibinder [4], Bruno Gobbato[5], Christopher P Roche[1], Vikas Kumar[1]*

[1] Exactech, Inc. Gainesville, FL, USA
[2] Blue-Ortho, Florida. Meylan, France
[3] Mayo Clinic, Florida. Jacksonville, FL, USA
[4] University of Michigan. Ann Arbor, MI, USA
[5] R. José Emmendoerfer, Nova Brasília, Jaraguá do Sul – SC, Brazil

hamid.oghaz@exac.com, josie.elwell@exac.com, francois.bouxdecasson@blue-ortho.com, sandrine.polakovic@blue-ortho.com, ashish.singh@exac.com, rakesh.Raushan@exac.com, likitha.shetty@exac.com, brad.schoch@gmail.com, williamaibindermd@gmail.com, bgobbato@gmail.com, chris.roche@exac.com, vikas.kumar@exac.com

**Abstract**

The deltoid muscles play a crucial role in maintaining balanced arm function and enabling abduction following shoulder arthroplasty. Currently, pre-operative assessments of deltoid integrity rely primarily on visual inspection of medical images and subjective ratings. A recent work has shown accuracy of machine learning based pipeline to correctly segment and quantify characteristics of deltoid muscle in shoulder CT scans. In this paper, with the inputs from medical experts, we evaluated clinical acceptance and non-inferiority of the ML-based segmentations compared to the corrections provided by expert surgeons. The non-inferiority of the ML model was assessed by comparing model-generated masks to surgeons' and inter-surgeon variations in metrics such as volume and fatty infiltration percentage. Expert validation showed 97% of masks to be clinically acceptable, with only 6% of ML generated masks requiring any major corrections. The

median error in the volume and fatty infiltration measurements was <1% between the ML-generated masks and the masks corrected by surgeons. The non-inferiority analysis demonstrated no significant difference between the generated masks to surgeons' and inter-surgeon variations ($p<0.05$).

# 1 Introduction

Treatment planning of patients who need shoulder arthroplasty relies on various pre-operative factors, including the integrity and condition of patient's joint, bone, and muscle. In the native shoulder, the range of motion and stability are largely dependent on function of the deltoid and the rotator cuff muscles. Deltoid muscles are the primary elevator and the rotator cuff, while also contributing to motion, dynamically stabilizes the joint. Degenerative changes to muscles, such as excessive fat infiltration or loss of muscle mass due to atrophy, can impact muscle function, joint stability, and range of motion. Degenerative changes to muscles, such as excessive fat infiltration (FI) or loss of muscle mass due to atrophy, can impact muscle function, joint stability, and range of motion [1, 2]. We recently developed an CT-based pipeline to segment and quantify shape and texture of deltoid muscle [3]. Then segmentation was based on fine-tuning of pretrained SwinUNETR [4], using manually labeled deltoid mask of 97 randomly selected patients [3]. Applying the pipeline on 1,057 patients revealed that the shape of the deltoid muscle, particularly its flatness, plays a significant role in predicting arthroplasty success [5]. Herein, we aim to: a) assess the clinical acceptance rate of the ML-generated deltoid masks, b) quantify segmentation and error between ML and surgeon-generated masks, and c) test the non-inferiority of ML to surgeon compared to inter-surgeon variations.

# 2 Methods

The population for the validation study was randomly selected from a multi-center clinical outcome database [6]. The population was chosen to represent at least three samples of patients from different demographics (age, gender, diagnosis, and treatment) and image-specific variables (image kernels, CT scan manufacturer). The selected cases underwent review to ensure they were not part of the development process. A total of 32 patients, 47% female, were selected for expert validation. Most patients were diagnosed with osteoarthritis (78%), followed by rotator cuff arthropathy (16%) and rotator cuff tear (19%). Patients received imaging with various CT scanners (50% GE, 28% SIEMENS, and 22% Toshiba).

Three qualified surgeons (fellowship-trained shoulder) and three technicians with experience in the manual segmentation of medical images participated in this study. The masks and the respective CT scans were randomly distributed among surgeons such that each surgeon reviewed about 20 cases, and each case was reviewed by at least two surgeons. Each mask evaluation consists of answering two questions: (a) Is the quality of the segmented mask clinically acceptable? (b) Does the mask benefit from minor or major corrections? For masks that required correction, a ground-truth was generated by technicians based on surgeon's comments with verification of the final mask by the surgeon. The differences between the ML and expert-curated masks were quantified using Dice coefficient, distance map, percentage of surface mesh with a gap more than 0.5 mm, correction ratio, percentage of corrected volume to ground-truth volume, and percentage error in volume and Fatty infiltration (FI).

A non-inferiority analysis was used to test whether the error in ML segmentation and quantification is substantially worse than the variation between the two surgeons reviewing a common set of masks.

For non-inferiority analysis, we followed a framework proposed by Ostmeier et al., [7] non-inferiority margin (Δ) was assumed as follows: for Dice coefficient as (1 – the minimum inter-surgeon Dice coefficient), for other error metrics as maximum of inter-surgeon error. The following summarizes our null hypothesis for non-inferiority analysis.

$$Dice_{inter-surgeon} \geq Dice_{ml-surgeon} + \Delta_{Dice} \quad (1)$$

$$Error_{inter-surgeon} \leq Error_{ml-surgeon} - \Delta_{Error} \quad (2)$$

A non-parametric one-sided Wilcoxon rank-sum was used to test the hypothesis, with a significance level of ($p < 0.05$).

# 3 Results

During the clinical validation, one mask was removed because delineation of the deltoid was not possible due to the presence of hematoma. The evaluation durations for surgeons A, B, and C were 90 minutes, 120 minutes, and 180 minutes, respectively. There was 100% agreement among surgeons on clinical acceptance/rejection ratings for all cases. The acceptance rate was 95% for Surgeons A and C, and 100% for Surgeon B, indicating a total acceptance rate of 97%. Most masks (81%) were suggested for minor correction. Only two masks (6%) were suggested for major correction, and one was clinically rejected. Figure 1 shows the ML and surgeon-generated masks for three samples: two clinically accepted with minor correction, and one rejected with major corrections. In total there were two cases suggested for major correction, where deltoid muscle presented with high degree of fattiness, resulting in darker area as shown in Figure 1 sample 3. Table 1 summarizes the non-inferiority results. The non-inferiority margin for Dice coefficient was 0.08. For the error metrics, the margins for the distance map, correction ratio, volume difference, and fat difference were 44%, 17%, 10%, and 3%, respectively. In summary, the model to surgeon error was non-inferior compared to the inter-surgeon variation for all metrics and surgeons. For surgeons A and B, the error between the model to surgeon was smaller than the inter-surgeon variations. For surgeon C, model to surgeon error was higher, but still found to be non-inferior to the inter-surgeon error.

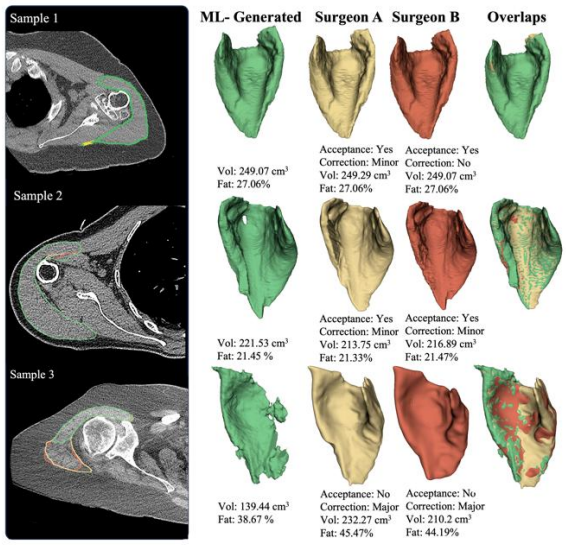


**Figure 1: Comparing masks generated by ML and surgeons.**

**Table 1.** Summary of Non-Inferiority Analysis:
Values represent the median followed by the 5th and 95th percentiles of the metrics.

| | Surgeon A | | | Surgeon B | | | Surgeon C | | |
|---|---|---|---|---|---|---|---|---|---|
| Errors | ML to Surgeon | Inter-Surgeon | Non Inferior p-value | ML to Surgeon | Inter-Surgeon | Non Inferior p-value | ML to Surgeon | Inter-Surgeon | Non Inferior p-value |
| Dice coefficient | 1.00 [0.97, 1.00] | 1.00 [0.98, 1.00] | P=.001 | 1.00 [0.98, 1.00] | 1.00 [0.98-1.00] | <.001 | 1.00 [0.98, 1.00] | 1.00 [0.98, 1.00] | P<.001 |
| Distance Map Error | 0.56% [0.00%, 9.01%] | 2.84% [0.00%, 8.45%] | P<.001 | 1.025% [0.00%, 7.24%] | 3.00% [0.00%, 8.88%] | P<.001 | 2.84% [0.00%, 6.25%] | 2.84% [0.00%, 10.3%] | P<.001 |
| Correction Ratio | 0.16% [0.00%, 6.26%] | 0.96% [0.00%-4.46%] | P<.001 | 0.30 [0, 3.345] | 0.725 [0.00, 3.58] | P<.001 | 0.8 [0.00%, 3.31%] | 0.65 [0.00%, 3.51%] | P<.001 |
| Volume Diff | 0.16% [0.00%, 3.51%] | 0.53% [0.00%, 3.77%] | P<.001 | 0.20% [0.00%, 3.44%] | 0.39% [0.00%, 1.98%] | P<.001 | 0.79% [0.00%, 3.31%] | 0.30% [0.00%, 1.85%] | P<.001 |
| FI Diff | 0.04% [0.00%, 3.52%] | 0.12 [0.00, 2.38%] | P<.001 | 0.09% [0.00%, 1.42%] | 0.105% [0.00%, 2.47%] | P<.001 | 0.06% [0.00%, 2.19%] | 0.08% [0.00%, 2.49%] | P<.001 |

# 4 Discussion

Commonly, ML studies are being validated using internal datasets and by quantifying mathematical metrics such as the Dice coefficient, which may not necessarily reflect the clinical acceptance of the ML models [8] [9]. Validation of ML models with the users for whom the model is intended and with external datasets that are generated outside of the development process can help assess clinical readiness and acceptability. In this study, we conducted an expert validation to evaluate clinical acceptability and tested the non-inferiority of the ML model compared to experienced orthopedic surgeons. Our findings confirmed high clinical acceptance of generated deltoid mask and demonstrated its non-inferiority in the measurement of volume or fatty infiltration.

Our study has some limitations. Due to inherent manual and costly process of expert validation the results are based on relatively small sample of patients who were already determined to undergo surgery. We also did not evaluate non-inferiority to the case where surgeons generate ground-truth from scratch likely exhibiting greater variability. Another limitation of the current study is lacking normal subjects as all scans were collected from patients who were selected for surgery. Also, further test is required to evaluate model performance on patients with high fatty infiltration or atrophy.

# 5 Conclusions

In this study, we evaluated a ML model for segmentation of deltoid muscles by demonstrating its high clinical acceptance rate and showing the non-inferiority of ML error compared to the variation between the expert surgeons. These findings contribute to the implementation of image-based ML models in clinical settings, leading to more effective treatment planning and patient satisfaction.

# References

[1] Goutallier, D., Postel, J.-M., Bernageau, J., Lavau, L., Voisin, M.-C.: Fatty muscle degeneration in cuff ruptures. Pre- and postoperative evaluation by CT scan. Clin. Orthop. (1994).

[2] Yoon, J.P., Seo, A., Kim, J.J., Lee, C.-H., Baek, S.-H., Kim, S.Y., Jeong, E.T., Oh, K.-S., Chung, S.W.: Deltoid muscle volume affects clinical outcome of reverse total shoulder arthroplasty in patients with cuff tear arthropathy or irreparable cuff tears. PLOS ONE. 12, (2017).

[3] Rajabzadeh-Oghaz, H., Elwell, J., Kumar, V., Mabrouk, L., Daviller, C., Berry, D., Singh, A., Polakovic, S., Schoch, B., Roche, C.: Machine-Learning Model for Quantification of Deltoid Characteristics. Proc. 2024 Orthop. Res. Soc. (2024)

[4] Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V. and Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022).

[5] Rajabzadeh-Oghaz, H., Kumar, V., Berry, D., Singh, A., Schoch, B., Aibinder, W., Gobbato, B., Polakovic, S., Elwell, J., Roche, C.P.: Impact of Deltoid CT Image Data on the Accuracy of Machine Learning Predictions of Clinical Outcomes After Anatomic and Reverse Total Shoulder Arthroplasty. J. Clin. Med. (2024)

[6] Roche, C.P., Jones, R.B., Routman, H., Marczuk, Y., Flurin, P.-H., Wright, T.W., Zuckerman, J.D.: Longitudinal analysis of shoulder arthroplasty utilization, clinical outcomes, and value: a comparative assessment of changes in improvement over 15 years with a single platform shoulder prosthesis. J. Shoulder Elbow Surg. (2023).

[7] Ostmeier, S., Axelrod, B., Verhaaren, B.F.J., Christensen, H., Mahammedi, A., Liu, Y., Pulli, B., Li, L.-J., Zaharchuk, G., Heit, J.J.: Non-inferiority of deep learning ischemic stroke segmentation on non-contrast CT within 16-hours compared to expert neuroradiologists. Sci. Rep. 13, (2023).

[8] Boman, M.: Human-Curated Validation of Machine Learning Algorithms for Health Data. Digit. Soc. 2, (2023).

[9] Cabitza, F., Campagner, A., Soares, F., de Guadiana-Romualdo, L.G., Challa, F., Sulejmani, A., Seghezzi, M., Carobene, A.: The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput. Methods Programs Biomed. 208, 106288–106288 (2021)