Radiology Advances, 2026, 3, umag005
https://doi.org/10.1093/radadv/umag005
Published: January 25, 2026
Original Research

Optimizing MRI annotation workflows for high-accuracy deep learning thigh muscle segmentation in athletes

Alexis B. Slutsky-Ganesh, PhD*[1,2,3,4], Salomé Baup, MS[5], Upasana U. Bharadwaj, MD, PhD[1,2,3], Jake A. Slaton, MS[1,2,3,6], Melanie Valencia, BS[1,2,3], Jed A. Diekfuss, PhD[1,2,3,7], Taylor M. Zuleger, PhD[1,2,3,7], Shayla M. Warren, BS[1,2,3], Kim D. Barber Foss, PhD[1,2,3], Kyle Hammond, MD[2], John W. Xerogeanes, MD[2], Ruud B. van Heeswijk, PhD[5], Gregory D. Myer, PhD[1,2,3,8], Augustin C. Ogier, PhD[5]

[1]Emory Sports Performance And Research Center (SPARC), Flowery Branch, GA, 30542, United States
[2]Emory Sports Medicine Center, Emory Healthcare, Atlanta, GA, 30329, United States
[3]Department of Orthopaedics, Emory University School of Medicine, Atlanta, GA, 30322, United States
[4]Department of Kinesiology, North Carolina Agricultural and Technical State University, Greensboro, NC, 27411, United States
[5]Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, 1011, Switzerland
[6]School of Health, Robert Gordon University, Aberdeen AB10 7QB, United Kingdom
[7]Department of Veterans Affairs, Atlanta VA Medical Center, Atlanta, GA, 30033, United States
[8]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, 30332, United States
*Corresponding author: Alexis B. Slutsky-Ganesh, PhD, North Carolina Agricultural and Technical State University, Emory Sport Performance and Research Center, 203 Corbett Sports Center, 405 N Benbow Road, Greensboro, NC 27411, United States (abganesh@ncat.edu)

Abstract

Background **Accurate thigh muscle segmentation from magnetic resonance imaging (MRI) enables quantitative assessment of muscle health. Although manual segmentation is the gold standard, it is labor-intensive and variable, and existing automated/semi-automatic approaches remain limited by segmentation errors/user dependence, restricting scalability. Defining data requirements for robust automated segmentation therefore remains a critical unmet need.**

Purpose **To determine the number of annotated lower extremity MRI studies needed to train an accurate deep learning (DL) model for thigh muscle segmentation and to assess the effect of training size on agreement of downstream quantitative measures.**

Materials and methods **Lower extremity MR images were obtained from competitive athletes with anterior cruciate ligament injuries and professional-level football athletes scanned at a single site on a 3 T GE Premier system. Fourteen thigh muscles were segmented using semi-automatic propagation followed by manual correction to generate high-quality ground-truth assisted manual segmentations (Seg_M). Thirteen DL models (nnU-Net) were trained with Seg_M on increasing numbers of training subjects (N_{train}) ranging from N_{train} = 5 up to N_{train} = 120, each evaluated on a fixed independent test set of 41 subjects. Automated segmentation (Seg_A) performance was evaluated using standard geometric accuracy metrics (Dice similarity coefficient [DSC], relative volume difference [RVD], Hausdorff Distance [HD], HD95, and average symmetric surface distance [ASSD]). To determine whether Seg_A would lead to *meaningful* quantitative MRI results, we also compared fat fraction and diffusion-tensor imaging measures extracted from Seg_A to those derived from Seg_M.**

Results **DL model training on N_{train} = 20 subjects achieved high accuracy on the fixed test set (mean ± SD: DSC 0.94 ± 0.02; RVD 4.9% ± 5.2%; ASSD 0.8 ± 0.4 mm; HD95 3.2 ± 2.8 mm), with modest improvement at 50 subjects.**

Conclusion **Twenty annotated images were sufficient for clinically acceptable performance, supporting streamlined segmentation and quantitative reporting in athlete care and research.**

Keywords segmentation, qMRI, quantitative MRI, muscle, artificial intelligence, neural networks, propagation, musculoskeletal

Received: September 26, 2025. **Revised:** December 17, 2025. **Accepted:** January 9, 2026

Abbreviations
ACL = anterior cruciate ligament; ASSD = average symmetric surface distance; DL = deep learning; DSC = Dice Similarity Coefficient; DTI = diffusion tensor imaging; FF = fat fraction; ICC = intraclass correlation coefficient; qMRI = quantitative MRI; RVD = relative volume difference

Summary
Deep learning-based thigh muscle segmentation on MRI achieved high accuracy with as few as 20 subjects and reduced annotation effort while preserving agreement for volume, fat fraction, and diffusion metrics.

Key Results

- Annotating 9 fixed axial levels per thigh and propagating to the full stack yields assisted manual labels suitable for training.
- Thirteen deep learning models for thigh muscle segmentation were trained on IDEAL MR water-only images from 5 to 120 subjects.
- A training size of 20 subjects met predefined accuracy targets for major muscle groups, with modest improvements when increased to 50.
- The code, propagation tool, and trained model are available publicly for reuse and validation.

Introduction

Accurate segmentation of thigh muscles from magnetic resonance imaging (MRI) is essential in both clinical and research settings to quantitatively assess muscle health and monitor recovery trajectories following injury.[1] Injuries such as anterior cruciate ligament (ACL) ruptures often result in significant changes in muscle morphology and composition, including atrophy and fatty infiltration, complicating the evaluation of muscle recovery during the rehabilitation process.[2] Further, in elite athletes, precise muscle characterization and identification of muscle imbalances can inform neuromuscular training optimization and injury prevention strategies.[3] Quantitative MRI (qMRI) techniques, including fat fraction (FF) mapping and diffusion tensor imaging (DTI), provide non-invasive biomarkers of muscle composition and microstructure that are relevant to muscle recovery and health. However, extracting meaningful, region-specific measurements from such images requires accurate segmentation of individual muscles. Thus, the utility of qMRI in both clinical and research practice depends on the availability of reliable segmentation tools that can accommodate anatomical variability and pathological changes.

Manual muscle segmentation, considered the gold standard, is labor-intensive, time-consuming,[4] and subject to inter-rater variability,[5] limiting its research utility and clinical scalability.[6] Automated segmentation presents a promising solution to overcome challenges with manual segmentation but is particularly vulnerable to error given the poor contrast between select adjacent muscles and the large variability of muscle shapes. Semi-automatic segmentation approaches reduce manual effort[7] but still require substantial user input and expertise, making them unsuitable for large-scale studies. Applications of deep learning (DL)-based methods for medical image segmentation, particularly convolutional neural networks, have demonstrated the potential to fully automate segmentation workflows with high accuracy and consistency.[7,8] However, the success of these fully automatic methods relies on established ground truth databases for accurate training. Semi-automatic segmentation, combined with manual correction, offers a practical approach to generate high-quality ground truth annotations to further train DL models. These annotations can be generated by manually segmenting a subset of 2D slices for propagation across the full 3D volume, or by refining initial annotations produced by artificial intelligence.[9]

However, it remains unclear how many images need to be semi-automatically segmented to obtain an adequate training set to confidently transition to fully automatic DL-based segmentation for routine utilization. We used the adaptable nnU-Net framework[10] to train 3D muscle-segmentation models on lower extremity MR images using incrementally larger training sets. We evaluated performance using geometric agreement, clinical relevance, and biomarker consistency (FF and DTI metrics) between manual and automated segmentations. This study aimed to define dataset-size requirements for reliable automatic thigh muscle segmentation to guide future clinical and research use.

Materials and methods

Dataset

This study included MRIs from 2 cohorts: athletes without acute injuries (Athlete) and athletes with an acute ACL rupture (ACL). Inclusion/exclusion criteria, enrollment, and final included samples are presented in Figure 1 (additional information in Supplementary Material). These studies were approved by the institutional review board of Emory University (American Professional Football Study [STUDY00003840]; ACL Registry Study [STUDY00002682]).

MRI acquisition

Scans were performed on a 3.0 T GE SIGNA Premier scanner (GE Healthcare) using a 30-channel Adaptive Image Receive Anterior Array coil. Each subject underwent 2 bilateral lower extremity sequences: (1) IDEAL-Spoiled Gradient-Echo imaging and (2) MUSE-DTI. Both sequences were acquired axially in 3 overlapping 40-slice stacks (5-slice overlap), spanning from the

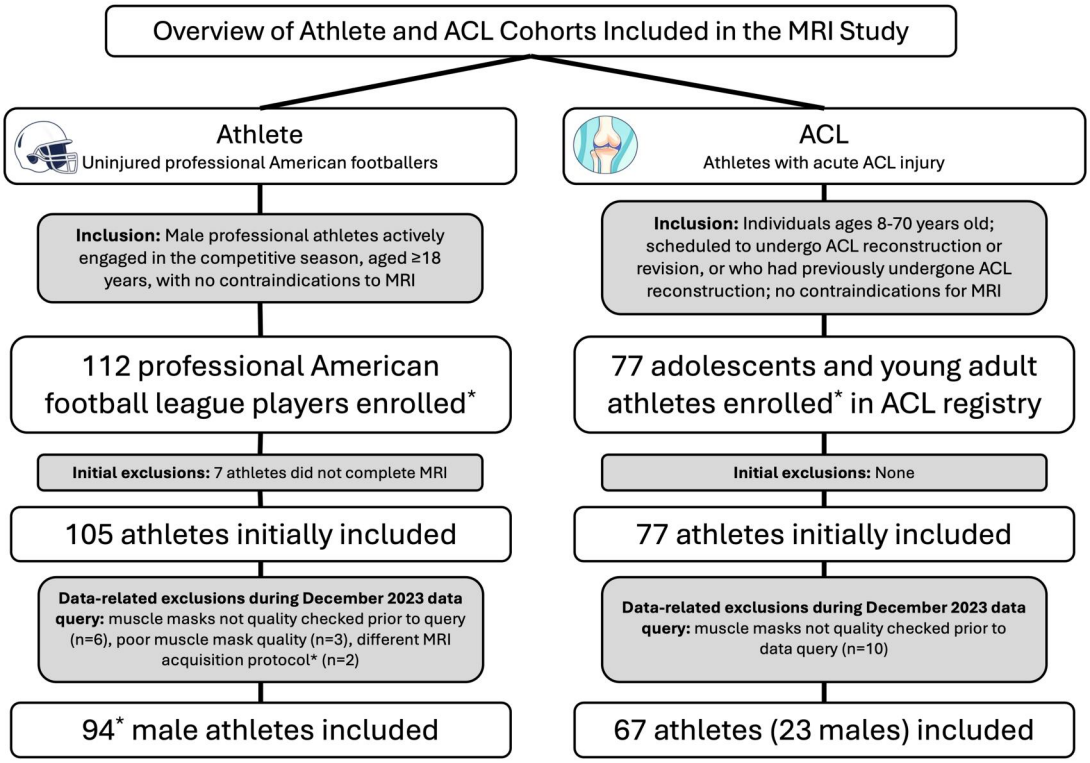

Figure 1 Study cohort overview and inclusion flow. Diagram illustrating the 2 study cohorts and participant inclusion for MRI analyses, including enrollment, exclusions, and final analytic samples for athletes with and without acute anterior cruciate ligament (ACL) injury. *Asterisks denote cohort-specific considerations detailed in the Supplementary Material, including enrollment timing, MRI acquisition differences, and partial muscle-level data inclusion.*

iliac crest to approximately the patella, with a resolution of 1.76 mm × 1.76 mm × 5 mm. Additional acquisition parameters, DTI preprocessing details, and descriptives are provided in Supplementary Material.

IDEAL imaging generates separate water and fat images; the water images, offering high muscle contrast, were used for segmentation, and fat fraction (FF) was calculated using both water and fat images as the percent fat within each voxel.[11] MUSE-DTI was used to assess diffusion-based muscle quality metrics, including fractional anisotropy (FA), mean diffusivity (MD), and radial diffusivity (RD; descriptives in Table S1).

Semi-automatic segmentation for assisted manual ground truth annotations

The muscles planned for segmentation included the left and right long head of *biceps femoris* (BFL), short head of *biceps femoris* (BFS), *rectus femoris* (RF), *semimembranosus* (SM), *semitendinosus* (ST), *vastus lateralis* (VL), and *vastus medialis* (VM). These muscles were selected for their clinical relevance to sports medicine, as they are key muscles in injuries (eg, hamstring strain injuries) and are often affected by ACL injuries (eg, quadriceps atrophy).

A previously validated semi-automatic segmentation approach[12,13] followed by manual correction was chosen to allow the building of reliable assisted manual annotations as ground truth for DL model training and testing (full description in Supplementary Material).

Nine slices were annotated for each muscle and then the segmentations were 3D-propagated through the remaining MRI volume (Figure 2) using a combination of diffeomorphic registration techniques,[12,13] enabling smooth, non-rigid deformation of the manually segmented masks across adjacent slices, ensuring the preservation of anatomical integrity. The propagated segmentations were then manually refined by expert researchers to correct any errors and adjust boundaries as needed. These refined results, referred to as assisted manual segmentations (Seg_M), were used as ground truth for DL model training. Two experts completed all the annotation refinements, and they were confirmed by a third expert prior to being used as ground truth.

DL-based fully automatic segmentation

For the DL model, we used the nnU-Net framework,[10] which has proven to be highly effective for volumetric medical image segmentation tasks. We employed the 3D full-resolution configuration, as it provided the best results in preliminary testing for muscle segmentation. The 3D configuration is particularly advantageous as it preserves the spatial relationships between slices and accurately captures the muscle's 3-dimensional structure. Training (80%) and validation (20%) sets were created, stratified equally by group (Athlete and ACL).

A primary aim was to determine the minimum number pre-annotated images that enabled accurate automatic segmentation using DL. To address this aim, we trained models on datasets of varying sizes (N_{train}, defined as the number of subjects in the training set), ranging from 5 to 120 subjects (Figure 3). These datasets were

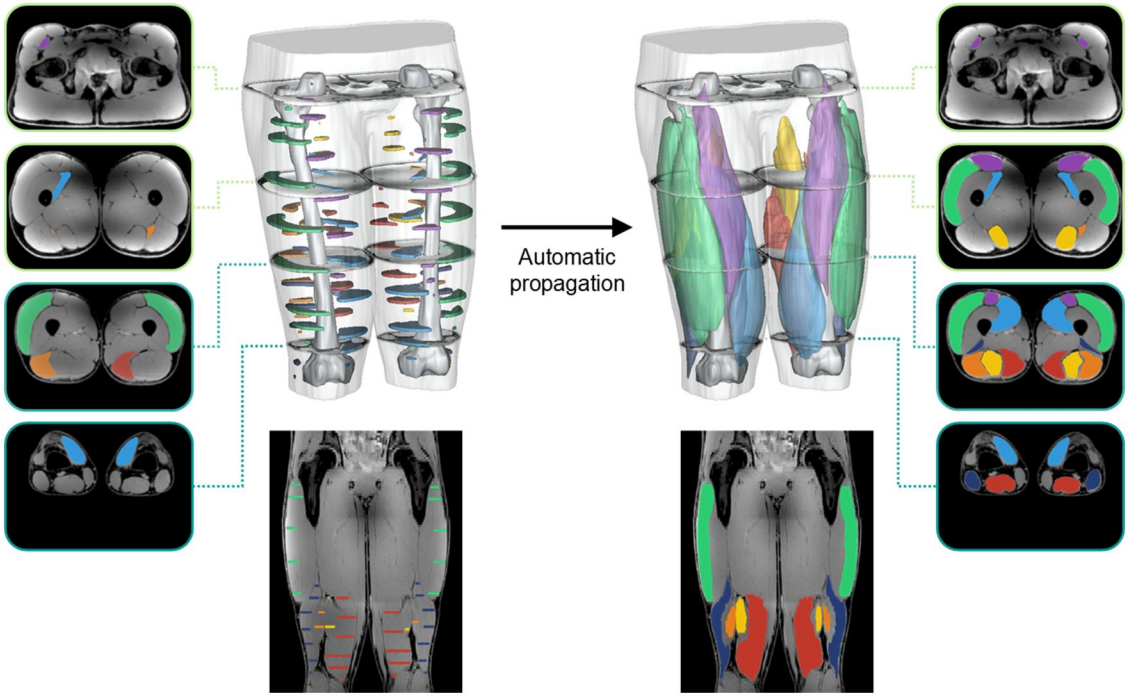

Figure 2 Illustration of the muscle segmentation process in the thighs, where 3D visualizations and corresponding axial and coronal slices demonstrate the transition from sparse manual annotations to full volumetric segmentation. (Left) Initial segmentation, where each muscle was manually segmented with 9 slices. The muscles were segmented independently, and the chosen slices varied across muscles based on proximal and distal slices. (Right) Automatic propagation of the initial segmentations to generate a complete 3D segmentation across the entire region of interest. The segmented muscles include *biceps femoris* long head (orange), *biceps femoris* short head (dark blue), *rectus femoris* (purple), *semimembranosus* (red), *semitendinosus* (yellow), *vastus lateralis* (green), and *vastus medialis* (light blue).

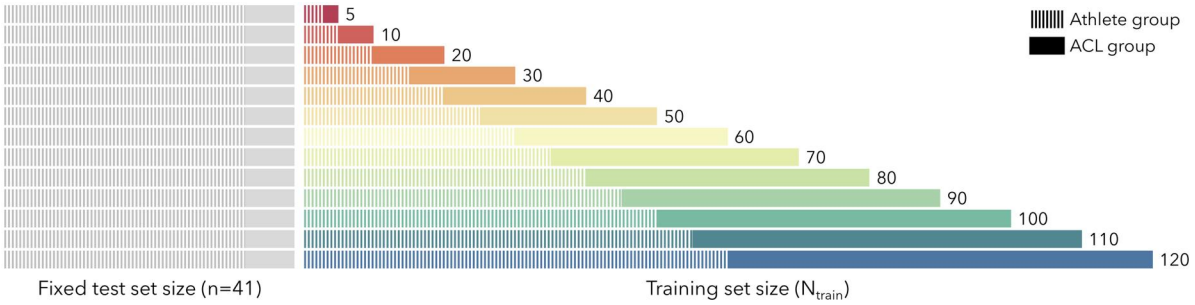

Figure 3 Training-set composition and evaluation strategy. Training-set size increased across experiments, with balanced contributions from athlete subjects (male only) and anterior cruciate ligament (ACL) subjects (male and female), while the test set of 41 subjects (34 Athlete, 7 ACL) remained fixed to ensure fair comparison.

constructed by selecting subjects in a sequential manner, sorted by acquisition date within each population group while keeping the contribution from each group similar. This sequential approach mirrored the typical enrollment process in clinical or research studies, allowing us to assess how the size of the training dataset influences model performance and identify the point at which assisted manual segmentation of additional subjects no longer significantly improves the DL-based fully automatic segmentation quality.

For consistency, the model's performance was evaluated using a fixed test set composed of 41 subjects—34 Athlete and 7 ACL subjects—selected as the last acquired cases within each respective group. This test set remained constant throughout the experiments, ensuring a fair comparison of the model's performance across different N_{train}. Demographic characteristics for each training-set configuration N_{train} and for the fixed test set are provided in Table S2.

The automatic segmentations generated by the model were referred to as Seg_A, and their accuracy was compared against the ground truth assisted manual segmentations Seg_M. Five of 41 test subjects were excluded from DTI analyses due to missing ($n = 4$) or poor-quality ($n = 1$) diffusion data.

The model was trained using Dice loss, which is well-suited for medical image segmentation[14] due to its direct optimization of spatial overlap between predicted and ground truth masks. The architecture, optimization parameters, and training schedule were fixed according to the default nnU-Net configuration,[10] and trainings

were run on a Linux Xeon Gold Workstation (6248rcpu@3.00 GHz, 96 GB RAM) with an Nvidia RTX A6000 GPU. Due to the potential confusion of the DL model in distinguishing left from right muscles in our 2-legged thigh MR images, a fully automatic post-processing step was implemented to correct misassigned labels on Seg_A. The method applies K-means clustering on non-zero voxel coordinates to identify the 2 legs, computes an x-axis midpoint between the clusters, and then reassigns muscle labels based on their spatial position relative to this threshold. After label reassignment, an additional post-processing step was applied to retain only the largest connected component for each segmented muscle.

For reference, the open-source MuscleMap DL model[15] was also applied to the same fixed test set, using the same subsequent post-processing steps, to provide an external baseline for comparison.

Validation metrics

To evaluate the performance of the trained models, we used several metrics that are standard in medical image segmentation tasks.[7] The Dice similarity coefficient (DSC) was used to measure the overlap between Seg_A and Seg_M. The Hausdorff distance (HD) was employed to quantify the spatial accuracy of the segmentation by measuring the maximum of the minimum distance between Seg_A and Seg_M boundaries. To reduce sensitivity to outliers, the 95th percentile HD (HD95) was also computed. The average symmetric surface distance (ASSD) was used to calculate the average distance between the surfaces of Seg_A and Seg_M segmentations, providing a more comprehensive measure of segmentation accuracy. Lastly, relative volume difference (RVD) was used to assess how well the

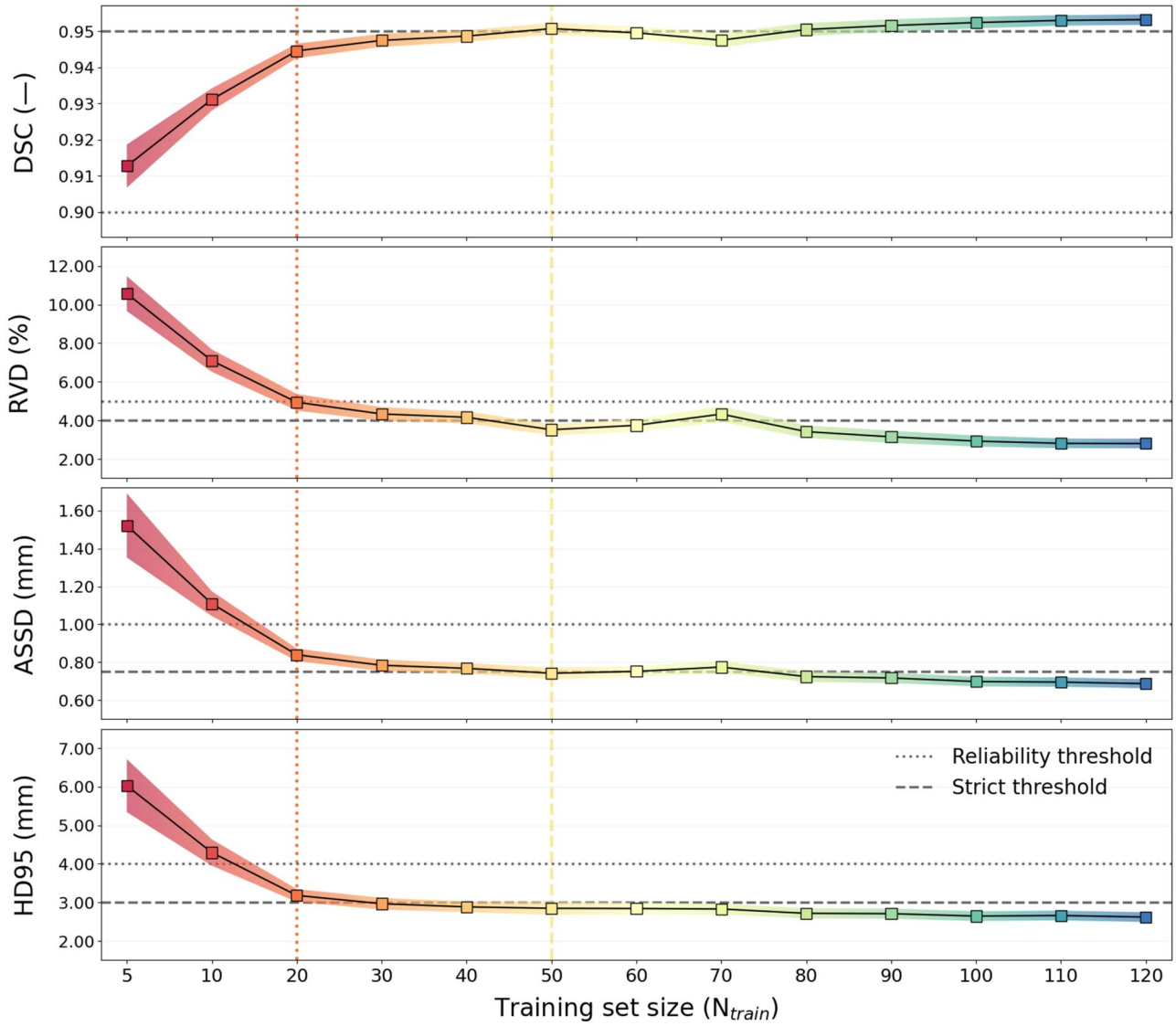

Figure 4 Learning curves for geometric metrics as a function of training set size (N_{train}) on the fixed test set ($n = 41$). Points show the mean value across all muscles and all test subjects, with shaded bands indicating 95% confidence intervals. Dotted horizontal lines denote the *a priori* reliability thresholds (DSC > 0.90, RVD < 5%, ASSD < 1.0 mm, HD95 < 4.0 mm), and dashed horizontal lines denote the *a priori* strict thresholds (DSC > 0.95, RVD < 4%, ASSD < 0.75 mm, HD95 < 3.0 mm). The dotted vertical line at $N_{train} = 20$ marks the N_{train} meeting all reliability criteria, and the dashed vertical line at $N_{train} = 50$ marks the smallest N_{train} meeting all strict criteria. Abbreviations: ASSD, average symmetric surface distance; DSC, dice similarity coefficient; HD95, 95% Hausdorff distance; RVD, relative volume difference.

model predicted the muscle volume, which is important for clinical applications involving quantitative analysis of muscle size. The same set of geometric metrics was used to evaluate the segmentation performance of the MuscleMap[15] baseline model.

To evaluate the consistency of the segmentations in extracting qMRI biomarkers, Seg_M and Seg_A masks were resampled and superimposed onto the FF and DTI maps. FF values were used to quantify fatty infiltration, while DTI-derived biomarkers, including FA, MD, and RD, were used to characterize muscle integrity. Agreement between the segmentation approaches was assessed by comparing the biomarker values extracted from Seg_A and Seg_M. The intraclass correlation coefficient (ICC) with a 2, 1

Table 1 Results from nnU-Net deep learning model N_{train} datasets validated on 41 test subjects.

Parameter	N_{train} = 5	N_{train} = 20	N_{train} = 50	N_{train} = 120
DSC (—)				
Biceps femoris-LH	0.89 ± 0.11	0.94 ± 0.03	0.95 ± 0.02	0.95 ± 0.02
Biceps femoris-SH	0.89 ± 0.06	0.93 ± 0.03	0.94 ± 0.02	0.94 ± 0.02
Rectus femoris	0.95 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.97 ± 0.01
Semimembranosus	0.89 ± 0.05	0.94 ± 0.02	0.94 ± 0.02	0.95 ± 0.02
Semitendinosus	0.92 ± 0.04	0.95 ± 0.02	0.95 ± 0.02	0.96 ± 0.01
Vastus lateralis	0.89 ± 0.11	0.94 ± 0.02	0.95 ± 0.02	0.95 ± 0.02
Vastus medialis	0.95 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01
Grand total	0.91 ± 0.07	0.94 ± 0.02	0.95 ± 0.02	0.95 ± 0.02
RVD (%)				
Biceps femoris-LH	14.8 ± 14.7	6.0 ± 6.2	3.6 ± 3.1	2.9 ± 3.1
Biceps femoris-SH	13.7 ± 11.2	7.0 ± 7.0	5.2 ± 4.8	4.2 ± 4.3
Rectus femoris	4.5 ± 4.7	3.4 ± 3.8	2.5 ± 2.2	2.1 ± 1.3
Semimembranosus	12.8 ± 8.9	6.0 ± 5.4	3.6 ± 3.5	3.0 ± 3.3
Semitendinosus	8.3 ± 7.3	4.5 ± 4.6	3.4 ± 3.2	2.5 ± 2.6
Vastus lateralis	14.5 ± 15.1	4.4 ± 4.5	3.9 ± 4.0	3.0 ± 2.8
Vastus medialis	5.5 ± 4.2	3.3 ± 2.9	2.3 ± 1.9	1.8 ± 1.8
Grand total	10.6 ± 11.0	4.9 ± 5.2	3.5 ± 3.5	2.8 ± 3.0
ASSD (mm)				
Biceps femoris-LH	2.0 ± 2.1	1.0 ± 0.5	0.7 ± 0.3	0.7 ± 0.3
Biceps femoris-SH	1.2 ± 1.0	0.7 ± 0.3	0.6 ± 0.2	0.6 ± 0.2
Rectus femoris	0.7 ± 0.3	0.6 ± 0.2	0.5 ± 0.2	0.5 ± 0.2
Semimembranosus	1.9 ± 1.4	1.0 ± 0.5	0.8 ± 0.4	0.7 ± 0.3
Semitendinosus	1.1 ± 0.7	0.7 ± 0.3	0.7 ± 0.4	0.6 ± 0.2
Vastus lateralis	2.7 ± 4.4	1.1 ± 0.4	1.0 ± 0.4	0.9 ± 0.3
Vastus medialis	1.1 ± 0.4	0.8 ± 0.3	0.8 ± 0.3	0.7 ± 0.3
Grand total	1.5 ± 2.1	0.8 ± 0.4	0.7 ± 0.4	0.7 ± 0.3
HD (mm)				
Biceps femoris-LH	32.2 ± 33.4	23.7 ± 32.2	18.9 ± 31.2	17.4 ± 31.1
Biceps femoris-SH	19.8 ± 11.7	13.6 ± 6.9	12.6 ± 6.5	11.2 ± 6.1
Rectus femoris	13.1 ± 19.1	12.0 ± 19.1	11.8 ± 19.1	11.0 ± 19.0
Semimembranosus	28.5 ± 18.7	13.4 ± 8.7	11.8 ± 8.2	11.0 ± 7.5
Semitendinosus	16.9 ± 11.1	13.4 ± 7.3	11.6 ± 7.5	10.3 ± 6.6
Vastus lateralis	32.4 ± 23.1	20.8 ± 15.9	20.6 ± 15.8	18.7 ± 15.8
Vastus medialis	21.6 ± 10.9	18.3 ± 10.3	17.8 ± 10.3	17.3 ± 10.2
Grand total	23.5 ± 20.9	16.5 ± 17.1	15.0 ± 16.6	13.8 ± 16.4
HD95 (mm)				
Biceps femoris-LH	7.9 ± 7.9	3.9 ± 2.6	2.9 ± 1.3	2.8 ± 1.5
Biceps femoris-SH	4.6 ± 5.2	2.3 ± 1.4	2.0 ± 1.0	1.9 ± 1.0
Rectus femoris	2.3 ± 1.3	1.9 ± 0.6	1.7 ± 0.7	1.7 ± 0.6
Semimembranosus	8.5 ± 9.5	3.8 ± 2.1	3.2 ± 1.8	3.0 ± 1.6
Semitendinosus	3.8 ± 3.1	2.5 ± 1.0	2.6 ± 3.2	2.1 ± 0.7
Vastus lateralis	11.3 ± 15.4	4.8 ± 2.1	4.6 ± 2.3	4.1 ± 1.7
Vastus medialis	3.8 ± 1.8	3.1 ± 1.4	2.9 ± 1.5	2.8 ± 1.5
Grand total	6.0 ± 8.4	3.2 ± 2.0	2.8 ± 2.0	2.6 ± 1.5

Data presented as means ± standard deviations. Learning curves for DSC, RVD, ASSD, and HD95 for every N_{train} dataset are presented in Figure 3. Data for all N_{train} datasets are included in Figure S1.

Abbreviations: ASSD = average symmetric surface distance; DSC = dice similarity coefficient; HD95 = 95% Hausdorff distance; LH = long head; N_{train} = training set size; RVD = relative volume difference; SH = short head.

formula[16] and standard error of measurement (SEM) were computed to quantify consistency. Additionally, Bland-Altman analyses were performed to estimate bias and limits of agreement between Seg_A and Seg_M in the quantification of each biomarker.

Based on previous literature,[17,18] we developed cutoffs to determine the size of a training set needed to result in a reliable DL segmentation model. These cutoffs included: DSC $>$ 0.9, RVD $<$ 5%, ASSD $<$ 1 mm, HD95 $<$ 4 mm, and all ICCs $>$ 0.98. These cutoffs have all been described as excellent agreement between manual muscle segmentation and DL-based segmentations. To evaluate the potential for a near-perfect model performance, a stricter criterion was used, which included DSC $>$ 0.95, RVD $<$ 4%, ASSD $<$ 0.75, HD95 $<$ 3 mm, and all ICCs $>$ 0.99. Although both HD and HD95 were reported, only HD95 was used in the cutoff criteria as it provides a more stable and robust measure less affected by small outliers.[19]

Availability

The semi-automatic segmentation tool (SegPropa) and the final trained deep learning model (SegAI-Thigh) are openly available for research use under a CC BY-NC-SA 4.0 license. *SegPropa* is hosted at https://gitlab.com/augustin-c-ogier/segpropa (release v1.0), and *SegAI-Thigh* at https://github.com/EmorySPARC/SegAI-Thigh (release v1.0).

Results

The final sample included for the Athlete group was 94 male athletes (Age_{mean} = 25.0 ± 2.4 years, BMI_{mean} = 28.1 ± 4.1). The final ACL group used in the present analyses included 67 athletes with an ACL injury (males = 23 [34%], Age_{mean} = 17.0 ± 3.6 years, BMI_{mean} = 22.2 ± 3.1).

Geometric metrics

With just 5 training subjects (N_{train} = 5), the average DSC across all muscles exceeded 0.90 (0.91 ± 0.07). Performance improved rapidly with more data (Figure 4), reaching 0.94 ± 0.02 with N_{train} = 20 and consistently >0.95 with N_{train} = 30 or more.

Other metrics also improved (Table 1, Figure S1): HD95 dropped below 4 mm at N_{train} = 20 (3.2 ± 2.0) and $<$ 3 mm by N_{train} = 40; ASSD fell below 1 mm at N_{train} = 20 and $<$ 0.75 mm by N_{train} = 50. RVD was ∼11% at N_{train} = 5, dropping below 5% by N_{train} = 20 (4.9% ± 5.2%) and <4% by N_{train} = 50 (3.5% ± 3.5%), as shown in Figure 4.

Across all 4 geometric metrics, 2 training-set sizes emerged as key operating points. As shown in Figure 4, N_{train} = 20 was the smallest training size at which all reliability thresholds were met simultaneously (DSC $>$ 0.90, RVD $<$ 5%, ASSD $<$ 1 mm, HD95 $<$ 4 mm). The stricter criteria (DSC $>$ 0.95, RVD $<$ 4%, ASSD $<$ 0.75 mm, HD95 $<$ 3 mm) were first reached at N_{train} = 50.

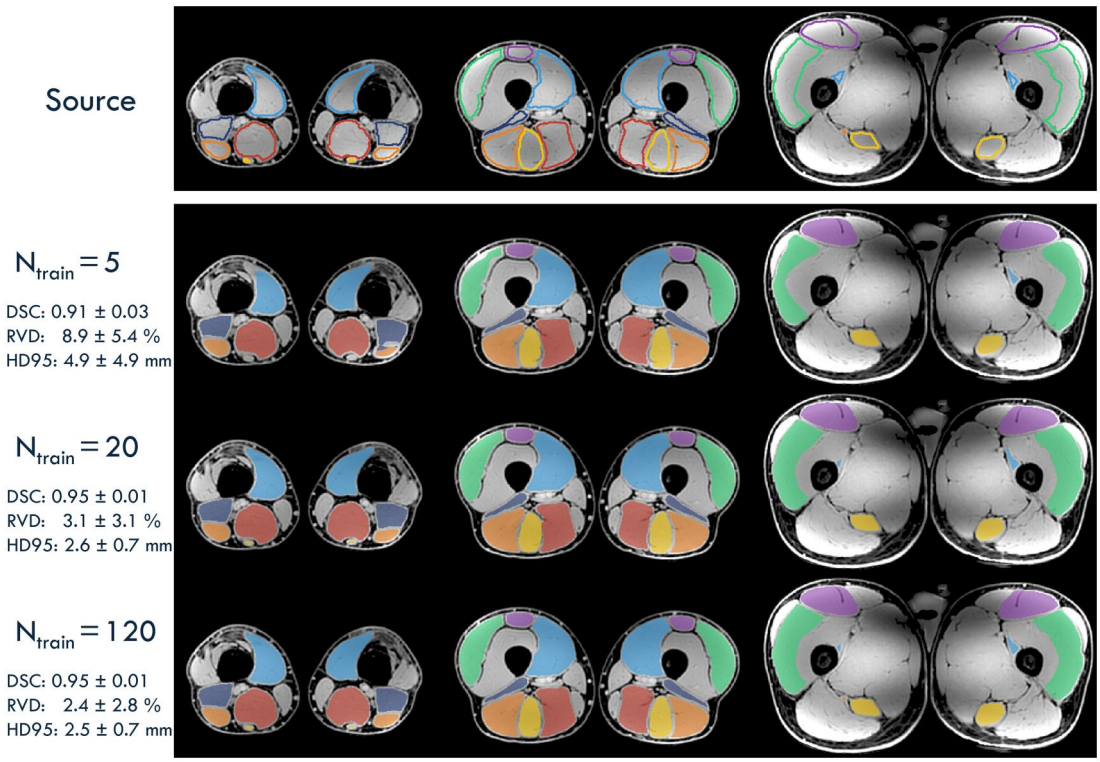

Figure 5 Representative distal, mid, and proximal thigh slices for 1 test subject from the athlete cohort, illustrating manual and automatic segmentations at different training-set sizes (N_{train}). The first row shows the source MR images with manual ground-truth boundaries overlaid as thin. Subsequent rows display the segmentation results for N_{train} = 5, 20, and 120. For each N_{train}, DSC, RVD, and HD95 are reported as mean ± SD across the 7 muscles. The segmented muscles include *biceps femoris* long head (orange), *biceps femoris* short head (dark blue), *rectus femoris* (purple), *semimembranosus* (red), *semitendinosus* (yellow), *vastus lateralis* (green), and *vastus medialis* (light blue). Abbreviations: DSC, dice similarity coefficient; HD95, 95% Hausdorff distance; RVD, relative volume difference.

Figure 5 illustrates Seg_M and Seg_A results for the smallest (N_{train} = 5), largest (N_{train} = 120), and high-performing models (N_{train} = 20, N_{train} = 50) as muscle cross-sections for a representative test subject.

For comparison, the recent open-source MuscleMap model[15] achieved an average DSC of 0.65 ± 0.21, RVD of 36.9 ± 28.2%, ASSD of 10.7 ± 15.3 mm, and HD of 67.3 ± 55.9 mm (Table S3).

Clinical metrics

Volume measurement ICCs were > 0.98 with N_{train} = 10 (ICC = 0.983) and > 0.99 with N_{train} = 20 (ICC = 0.995) (Figure S2). Bland-Altman analysis showed minimal FF differences between Seg_M and Seg_A without significant bias (Figures S3). FF ICCs were > 0.98 at N_{train} = 20 (ICC = 0.989) and > 0.99 at N_{train} = 50 (ICC = 0.991). DTI-derived metrics—FA, MD, and RD—had consistently high ICCs (> 0.995 at

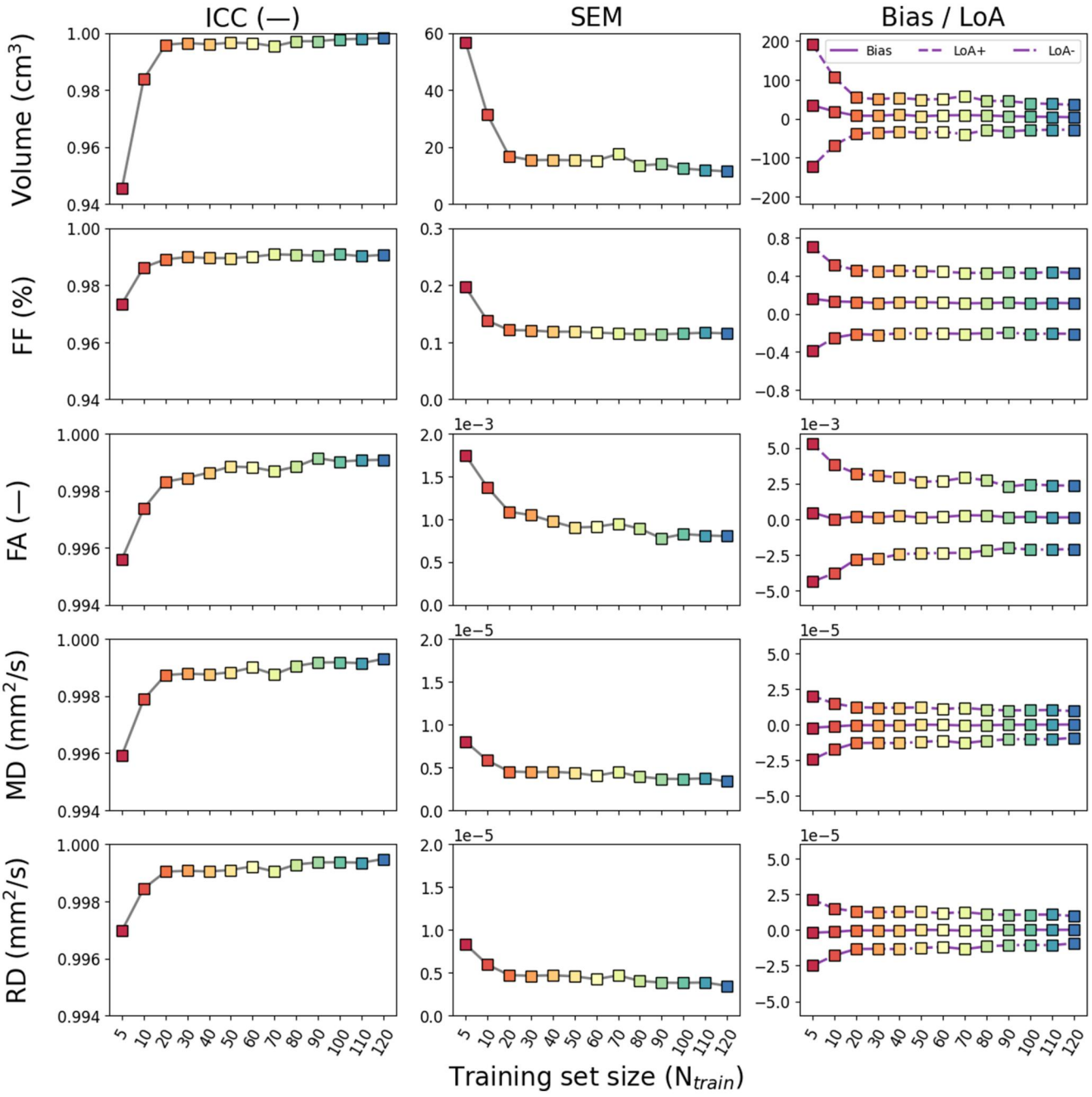

Figure 6 Intraclass correlation coefficient (ICC), standard error of measurement (SEM), and Bland-Altman bias and limits of agreement (LoA) for muscle segmentation-derived metrics across different training dataset sizes (N_{train} = 5 to 120). Each row represents comparisons between values derived from the assisted manual segmentation (SegM) and the automatic segmentation (SegA) for muscle volume, fat fraction (FF), fractional anisotropy (FA), mean diffusivity (MD), and radial diffusivity (RD), in that order. The columns display ICC values (unitless), SEM values, and Bland-Altman bias (solid line) with limits of agreement (dashed lines), the latter 2 expressed in the units of the respective metrics. Square markers represent individual training datasets.

N_{train} = 5), confirming DL segmentation reliably delineates muscle boundaries to assess muscle tissue characteristics (Figures S4-S6). ICCs, SEM, and Bland-Altman results for volume, FF, and DTI metrics are shown in Figure 6.

Discussion

This study presents a practical framework for transitioning from assisted-manual to fully automatic thigh muscle segmentation in professional athletes and individuals with ACL injury using deep learning. Utilizing the nnU-Net model,[10] we demonstrated that reliable performance can be achieved using as few as 20 training datasets and further improvements in accuracy and robustness are observed with 50. The ability to train on relatively small datasets is relevant for clinical and sports applications in which accurate muscle delineation supports recovery monitoring, muscle composition assessment, and treatment planning.

A key strength is the simulation of real-world conditions in which data accrue sequentially. A modest number of assisted-manual cases was sufficient to train models that deliver accurate results. Our results provide a significant step forward for developing applications requiring rapid and objective measurements of muscle health, such as rehabilitation following an ACL injury or athlete performance monitoring. Importantly, the validation of segmentation outputs through clinically relevant biomarkers, such as fat fraction (FF) and diffusion tensor imaging (DTI) metrics, confirms the model's ability to preserve muscle characteristics.

The segmentation performance obtained in this study is consistent with, and in many cases exceeds, previously published results on automatic segmentation of individual leg athletes have reported DSCs at 0.85-0.91 ± 0.03-0.09,[8,20,21] whereas studies focused on inclusion of participants with varying levels of fat infiltration have reported DSCs at 0.88-0.90 ± 0.03-0.09,[22–24] which is known to increase segmentation difficulty.[25] In comparison, our model inclusive of healthy and injured athletes trained with datasets of N_{train} = 120 subjects achieved a mean DSC of 0.95 ± 0.02, having attained 0.94 ± 0.02 at N_{train} = 20, which compares favorably with previously reported methods in terms of geometric accuracy. For context, the open-source MuscleMap model achieved a mean DSC of 0.65 ± 0.21 on the same test set, suggesting that generic models may not yet achieve the same level of accuracy as dataset-specific training.

The accuracy of the current models is particularly notable given that our approach relied entirely on the default nnU-Net configuration,[10] without any architectural modification or extensive hyperparameter tuning. In contrast, previously published methods often required significant implementation effort, involving custom pipelines such as multi-stage networks with separate localization and segmentation models,[8] contrastive pretraining strategies to reduce annotation needs,[26] or the design and comparison of multiple architectures including U-Net, DeepLab v3+, or PSPNet variants.[21] Others introduced architectural complexities such as dense or residual blocks,[8] additional boundary-aware modules,[22] or fine-tuning of pretrained classification networks like AlexNet,[20] often coupled with manual optimization of learning schedules, loss functions, and data augmentations. While prior optimization approaches may provide incremental improvements, our findings indicate that off-the-shelf solutions like nnU-Net not only simplify deployment but also deliver state-of-the-art accuracy.

Our chosen methodological approaches are not without limitations. First, we used a sequential data acquisition strategy for selecting training subjects, reflecting real-world clinical studies but lacking the efficiency of active learning, which can improve dataset quality by prioritizing informative samples and reducing annotation burden.[27] Second, the model's generalizability may be limited by the single-site design and the relatively homogeneous cohort of generally healthy young individuals, with a subset experiencing an acute injury. However, the objective of this study was not to develop a universal thigh segmentation model but to quantify the annotation requirements for achieving reliable deep learning–based automation. Future studies should extend this approach to more heterogeneous populations, such as individuals with neuromuscular disorders, for which the same semi-automatic framework has previously enabled accurate DL segmentation,[25] to assess annotation needs under greater anatomical variability.

To facilitate the adoption of DL-based muscle segmentation, we provide the semi-automatic 3D propagation tool (SegPropa) used in this study to create the annotated training database. The availability of pre-trained models for similar MRI protocols can further simplify implementation, minimizing the need for retraining. Open-source platforms like Dafne,[28] which provide lower limb MRI models and support model refinement and updates, offer promising paths for shared and continuously improving DL tools. As such, we have openly provided the final nnU-Net model trained on N_{train} = 120 subjects with 1 example dataset (*SegAI-Thigh*), to facilitate use and independent external validation by other groups. This model is currently utilized for ongoing research data collection and has been successfully applied to $>$ 600 datasets.

In conclusion, this work provides a practical, data-driven approach to developing DL models for fully automatic thigh muscle segmentation, demonstrating its feasibility and utility in developing clinical applications in the care of healthy and injured athlete populations. By balancing annotation effort with segmentation performance, the study highlights the potential of DL to streamline muscle health assessment, reduce inter-rater variability, and enable scalable monitoring of rehabilitation and training outcomes. Future advancements, including the incorporation of pre-trained models and active learning strategies, will further reduce the need for large, annotated datasets, paving the way for the widespread adoption of automated segmentation tools in sports medicine clinical practice and research.

Author contributions

Alexis B. Slutsky-Ganesh (Conceptualization, Data curation, Methodology, Project administration, Writing—original draft, Writing—review & editing), Salomé Baup (Formal analysis, Methodology, Validation, Visualization, Writing—review & editing), Upasana Bharadwaj (Methodology, Writing—original draft, Writing—review & editing), Jake A. Slaton (Data curation, Visualization, Writing—review & editing), Melanie Valencia (Data curation, Writing—review & editing), Jed A. Diekfuss (Data curation, Investigation, Writing—review & editing), Taylor Zuleger

(Data curation, Investigation, Writing—review & editing), Shayla Warren (Project administration, Writing—review & editing), Kim D. Barber Foss (Project administration, Writing—review & editing), Kyle Hammond (Project administration, Resources, Writing—review & editing), John W. Xerogeanes (Project administration, Resources, Writing—review & editing), Ruud B. van Heeswijk (Project administration, Resources, Software, Writing—review & editing), Gregory D. Myer (Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Software, Supervision, Writing—review & editing), and Augustin C. Ogier (Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *Radiology Advances* online.

Funding

There was no external funding received to support this work.

Conflicts of interest

A. B. S.-G. has no conflicts of interest to disclose. Additional conflicts of interest are reported separately by each co-author on the ICMJE Disclosure Forms provided as Supplementary Material.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1. Lenchik L, Heacock L, Weaver AA, et al. Automated segmentation of tissues using CT and MRI: a systematic review. *Acad Radiol*. 2019;26(12):1695-1706. https://doi.org/10.1016/j.acra.2019.07.006.
2. Garcia SA, Moffit TJ, Vakula MN, Holmes SC, Montgomery MM, Pamukoff DN. Quadriceps muscle size, quality, and strength and self-reported function in individuals with anterior cruciate ligament reconstruction. *J Athl Train*. 2020;55(3):246-254. https://doi.org/10.4085/1062-6050-38-19
3. Chavarro-Nieto C, Beaven M, Gill N, Hebert-Losier K. Hamstrings injury incidence, risk factors, and prevention in Rugby Union players: a systematic review. *Phys Sportsmed*. 2023;51(1):1-19. https://doi.org/10.1080/00913847.2021.1992601
4. Barnouin Y, Butler-Browne G, Voit T, et al. Manual segmentation of individual muscles of the quadriceps femoris using MRI: a reappraisal. *J Magn Reson Imaging*. 2014;40(1):239-247. https://doi.org/10.1002/jmri.24370
5. Muzic SI, Paoletti M, Solazzo F, et al. Reproducibility of manual segmentation in muscle imaging. *Acta Myol*. 2021;40(3):116-123. https://doi.org/10.36185/2532-1900-052
6. Liu X, Xing F, Marin T, Fakhri GE, Woo J. Variational inference for quantifying inter-observer variability in segmentation of anatomical structures. *Proc SPIE Int Soc Opt Eng*. 2022;12032. https://doi.org/10.1117/12.2604547
7. Ogier AC, Hostin MA, Bellemare ME, Bendahan D. Overview of MR image segmentation strategies in neuromuscular disorders. *Front Neurol*. 2021;12:625308. https://doi.org/10.3389/fneur.2021.625308
8. Ni R, Meyer CH, Blemker SS, Hart JM, Feng X. Automatic segmentation of all lower limb muscles from high-resolution magnetic resonance imaging using a cascaded three-dimensional deep convolutional neural network. *J Med Imaging (Bellingham)*. 2019;6(4):044009. https://doi.org/10.1117/1.JMI.6.4.044009
9. Philbrick KA, Weston AD, Akkus Z, et al. RIL-Contour: a medical imaging dataset annotation tool for and with deep learning. *J Digit Imaging*. 2019;32(4):571-581. https://doi.org/10.1007/s10278-019-00232-0
10. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. https://doi.org/10.1038/s41592-020-01008-z
11. Liu CY, McKenzie CA, Yu H, Brittain JH, Reeder SB. Fat quantification with IDEAL gradient echo imaging: correction of bias from T(1) and noise. *Magn Reson Med*. 2007;58(2):354-364. https://doi.org/10.1002/mrm.21301
12. Ogier A, Sdika M, Foure A, Le Troter A, Bendahan D. Individual muscle segmentation in MR images: a 3D propagation through 2D non-linear registration approaches. *Annu Int Conf IEEE Eng Med Biol Soc*. 2017;2017:317-320. https://doi.org/10.1109/EMBC.2017.8036826
13. Ogier AC, Heskamp L, Michel CP, et al. A novel segmentation framework dedicated to the follow-up of fat infiltration in individual muscles of patients with neuromuscular disorders. *Magn Reson Med*. 2020;83(5):1825-1836. https://doi.org/10.1002/mrm.28030
14. Ma J, Chen J, Ng M, et al. Loss odyssey in medical image segmentation. *Med Image Anal*. 2021;71:102035. https://doi.org/10.1016/j.media.2021.102035
15. McKay MJ, Weber KA, Wesselink EO, et al. MuscleMap: an open-source, community-supported consortium for whole-body quantitative MRI of muscle. *J Imaging*. 2024;10(11):262. https://doi.org/10.3390/jimaging10110262
16. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation—a discussion and demonstration of basic features. *PLoS One*. 2019;14(7):e0219854. https://doi.org/10.1371/journal.pone.0219854
17. Henson WH, Li X, Lin Z, Guo L, Mazza C, Dall'Ara E. Automatic segmentation of lower limb muscles from MR images of post-menopausal women based on deep learning and data augmentation. *PLoS One*. 2024;19(4):e0299099. https://doi.org/10.1371/journal.pone.0299099
18. Riem L, Feng X, Cousins M, et al. A deep learning algorithm for automatic 3D segmentation of rotator cuff muscle and fat from clinical MRI scans. *Radiol Artif Intell*. 2023;5(2):e220132. https://doi.org/10.1148/ryai.220132

19. Ramachandran P, Eswarlal T, Lehman M, Colbert Z. Assessment of optimizers and their performance in autosegmenting lung tumors. *J Med Phys*. 2023;48(2):129-135. https://doi.org/10.4103/jmp.jmp_54_23
20. Ghosh S, Ray N, Boulanger P. A structured deep-learning based approach for the automated segmentation of human leg muscle from 3D MRI. 2017;2017:117-123. https://doi.org/10.1109/CRV.2017.32
21. Kasahara J, Ozaki H, Matsubayashi T, Takahashi H, Nakayama R. Semantic segmentation for individual thigh skeletal muscles of athletes on magnetic resonance images. *Radiol Phys Technol*. 2025;18(2):496-504. https://doi.org/10.1007/s12194-025-00901-6
22. Guo Z, Zhang H, Chen Z, et al. Fully automated 3D segmentation of MR-imaged calf muscle compartments: neighborhood relationship enhanced fully convolutional network. *Comput Med Imaging Graph*. 2021;87:101835. https://doi.org/10.1016/j.compmedimag.2020.101835
23. Ding J, Cao P, Chang HC, Gao Y, Chan SHS, Vardhanabhuti V. Deep learning-based thigh muscle segmentation for reproducible fat fraction quantification using fat-water decomposition MRI. *Insights Imaging*. 2020;11(1):128. https://doi.org/10.1186/s13244-020-00946-8
24. Agosti A, Shaqiri E, Paoletti M, et al. Deep learning for automatic segmentation of thigh and leg muscles. *MAGMA*. 2022;35(3):467-483. https://doi.org/10.1007/s10334-021-00967-4
25. Hostin MA, Ogier AC, Michel CP, et al. The impact of fatty infiltration on MRI segmentation of lower limb muscles in neuromuscular diseases: a comparative study of deep learning approaches. *J Magn Reson Imaging*. 2023;58(6):1826-1835. https://doi.org/10.1002/jmri.28708
26. Casali N, Scalco E, Taccogna MG, et al. Positional contrastive learning for improved thigh muscle segmentation in MR images. *NMR Biomed*. 2024;37(10):e5197. https://doi.org/10.1002/nbm.5197
27. Boehringer AS, Sanaat A, Arabi H, Zaidi H. An active learning approach to train a deep learning algorithm for tumor segmentation from brain MR images. *Insights Imaging*. 2023;14(1):141. https://doi.org/10.1186/s13244-023-01487-6
28. Santini F, Wasserthal J, Agosti A, et al. Deep anatomical federated network (Dafne): an open client/server framework for the continuous collaborative improvement of deep-learning-based medical image segmentation. 2023. https://doi.org/10.48550/arXiv.2302.06352.

Radiology Advances, 2026, 3, 1–11
https://doi.org/10.1093/radadv/umag005
Original Research