**ACTA MYOLOGICA 2021**; XL: p. 116-123
doi:10.36185/2532-1900-052

ORIGINAL ARTICLES

# Reproducibility of manual segmentation in muscle imaging

Shaun Ivan Muzic[1]*, Matteo Paoletti[2]*, Francesca Solazzo[1], Eugenio Belatti[1], Raimondo Vitale[1], Niels Bergsland[3,4], Stefano Bastianello[1,2], Anna Pichiecchio[1,2]

*[1] University of Pavia, Pavia, Italy; [2] Department of Neuroradiology, IRCCS Mondino Foundation, Pavia, Italy; [3] Buffalo Neuroimaging Analysis Center, Department of Neurology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, Buffalo, NY, USA; [4] Don Carlo Gnocchi ONLUS Foundation IRCCS, Milan, Italy*

**These authors have contributed equally to the study*

Received: March 29, 2021
Accepted: September 12, 2021

**Correspondence**
Anna Pichiecchio
Department of Neuroradiology, IRCCS Mondino Foundation, via Mondino 2, 27100 Pavia, Italy
E-mail: anna.pichiecchio@mondino.it

**How to cite this article:** Muzic SI, Paoletti M, Solazzo F, et al. Reproducibility of manual segmentation in muscle imaging. Acta Myol 2021;40:116-123. https://doi.org/10.36185/2532-1900-052



**Purpose. To assess the reproducibility of a manual muscle MRI segmentation method that follows a specific set of recommendations developed in our center.**

**Materials and methods. Nine healthy volunteers underwent a muscle MRI examination that included a TSE T2 sequence of the thighs. Muscle segmentation was performed by three operators: an expert operator (OP1) with 3 years of experience and two radiology residents (OP2 and 3) who were both given basic segmentation instructions, whereas only OP2 underwent additional supervised training from OP1. Intra- and inter-operator Dice similarity coefficient (DSC) was calculated.**

**Results. OP1 showed the highest average intra-operator DSC values (0.885), whereas OP2 had higher average DSC (0.856) compared to OP3 (0.818). The highest inter-operator agreement was observed between Operators 1 and 2 (0.814) and the lowest between OP2 and OP3 (0.702). Confidence interval (CI) analysis showed that the most experienced operator also had the least variability in drawing the ROIs, whereas OP2 showed both higher intra-operator reproducibility compared to OP3 and higher inter-operator agreement with OP1. The muscles that showed the least reproducibility were the *semimembranosus* and the short head of the *biceps femoris*.**

**Discussion. Following specific recommendations such as these ones derived from our single-center experience leads to an overall high reproducibility of manual muscle segmentation and is helpful in improving both intra-operator and inter-operator reproducibility in less experienced operators.**

**Key words:** muscle MRI, segmentation, reproducibility, neuromuscular diseases, qMRI

## Introduction

Quantitative magnetic resonance imaging (qMRI) methods are being increasingly used in the field of neuromuscular diseases to assess quantitative features of the muscle such as fat fraction, T1 and T2 relaxation. Additional parameters can also be derived from advanced MRI techniques [1-3], the majority of which are progressively recognized and used as biomarkers for clinical trials [4]. An accurate and reproducible segmentation of muscles, achieved through the drawing of regions of interest (ROI), is a necessary condition for the technological advancement of qMRI methods, and would allow a wider diffusivity and a broader application of its techniques.

Muscle segmentation has thus far been mostly an operator-dependent, manual operation; however, the process of drawing the ROIs is time consuming, usually needs a dedicated operator, and training times for operators can be quite long before a good performance is achieved [5].

Even though automatic and semi-automatic processes are being developed with promising results through the application of neural networks and machine learning techniques [6,7], the process still needs to be supervised. Furthermore, the availability of an extensive dataset of muscle ROIs realized by expert hand drawing represents a fundamental step in training neural networks, hence the importance of an adequate ROI drawing process.

Few studies described the details of the technique of ROI drawing, and those same studies often apply differing approaches. The differences between the techniques concerned not only the portion of the muscle being segmented – ranging from partial muscle delimitation in the form of circular [8,9] or square-shaped ROIs [10,11] to the complete delimitation of the segmented muscles – but also the number of segmented slices [1,2,12], the level at which the segmentation was performed [4,13] and the MRI sequence used for segmentation [14], although some studies underline the importance of maintaining adequate distance from the muscle fascia [4]. A consensus on the optimal criteria for drawing ROIs over the examined muscles, especially in the case of complete muscle delimitation, is still lacking. In particular, no specific indication exists regarding how the ROI border should be drawn within the muscle area (i.e., the exact distance from the muscle fascia, the inclusion of fatty infiltration or vascular structures), a detail that is particularly important, especially when considering the aforementioned aim of training neural networks.

In this study, we aimed to assess the reproducibility of a manual muscle segmentation method in a small cohort of healthy volunteers using a specific set of recommendations suggested by our single-center experience. The muscle segmentation was centered at the thigh level, as this region is the most commonly examined area in the field of neuromuscular diseases [5,15-17].

## Materials and methods

We enrolled 9 healthy volunteers that gave their formal informed consent for the participation in the study, which was approved by the local ethics committee. None of these subjects had significant muscle pathologies, and neurological examination of the muscle function was normal. All subjects underwent a muscle MRI protocol with a 3 Tesla scanner (Skyra Siemens, Erlangen Germany) including a TSE T2 multi-echo sequence with EPG fitting (17 echo times), developed for T2 mapping. Sequences were centered on the thigh muscles. The 32-channels spine coil and a 18-channels phased-array coil positioned upon the thighs were used. The first echo image (TE = 10.9 ms) was arbitrarily used for segmentation.

The drawing of ROIs was performed on the axial slice that was approximately equidistant from the uppermost part of the femoral heads and the lowermost part of

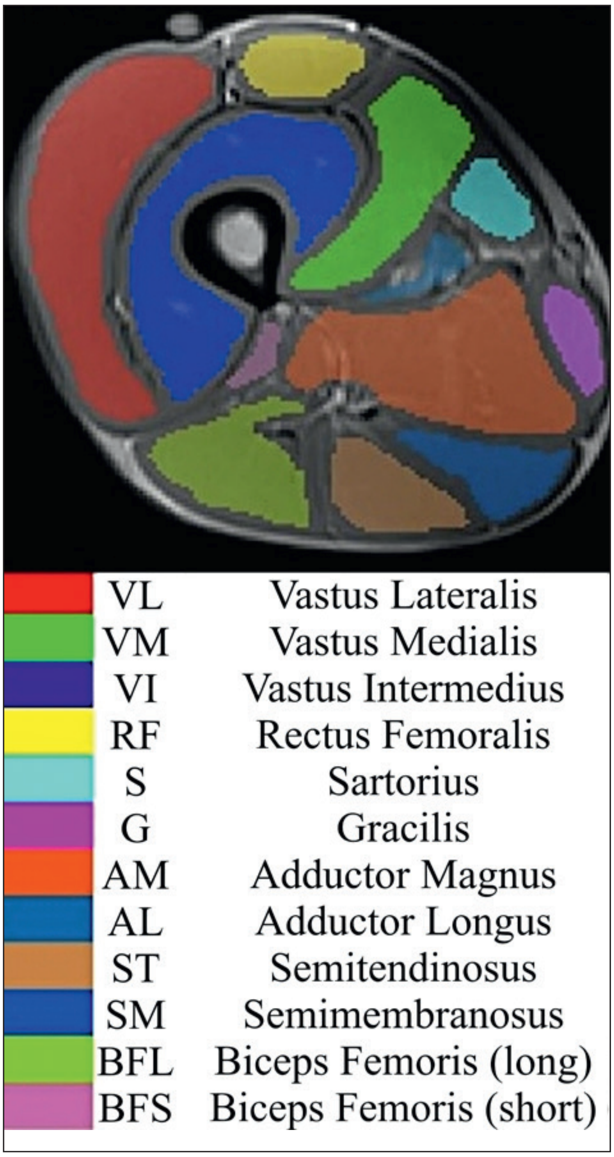


**Figure 1.** Example of ROIs obtained during manual thigh muscle segmentation superimposed on a TSE T2 weighted image. The color code indicates all 12 ROIs that were considered, which were drawn on both the left and right thigh, with the corresponding muscle indicated in the figure.

the femoral condyles, in order to capture a level where the muscle bellies have a wider spatial representation. Twelve ROIs, one for each muscle of both left and right thigh (Fig. 1), were manually drawn using the ITK-SNAP software (v 3.8.0) [18] by three operators: one operator with 3 years of experience in muscle segmentation (OP1), and two radiology residents. Both radiology residents (OP 2 and OP3), underwent two hours of basic training on muscle segmentation by OP1.

In this study, operators were trained to leave a thin 1-2 mm border of unsegmented muscle tissue between the margins of each ROI and the T2 hypointense fascia and cortical bone of the femur. The same thin border was to be applied between two ROIs of contiguous muscles if no evident anatomical margin was observed (e.g., in the case of the delimitation between the *adductor magnus* and *semimembranosus* in some cases in the young adult subgroup). Secondary instructions for muscle delimitation included avoiding extra-muscular connective tissue, as well as internal fascia invaginations, which at times resulted in a less regular morphology of the ROI but is presumably more representative of the actual muscle tissue. Fatty muscle degeneration within the borders of the muscle fascia in the form of thin, non-confluent internal T2 hyperintensities (though less hyperintense compared to the extra-muscular fat tissue) was to be included as well. In case of doubt, the less experienced operators were instructed to refer to previous segmentations of other patients present in the internal database, for reference. As a last resort, if serious doubts persisted about the location and size of a specific muscle, less experienced operators were trained not to segment that particular muscle.

The instructions also included a review of the anatomy of the thigh muscles, an MRI-based atlas of the lower limbs and hands-on training in the use of the segmentation software. The same radiology residents were also given a reference dataset of muscle MRI studies that were already segmented by the expert operator (OP1). Only OP2 additionally received advanced practical training in the form of twenty practice cases to be segmented independently, with subsequent corrections of the ROIs and further advice by OP1. Such segmentation instructions included leaving a thin border of one to two voxels of muscle tissue while avoiding extra-muscular connective tissue, as well as internal fascia invaginations. All three operators then performed the segmentation process twice on the same slice for each patient, at the beginning of the study and 72 hours after the first segmentation.

Both intra- and inter-operator agreement of the spatial overlap of the ROIs were evaluated with Dice Similarity Coefficient (DSC) (average value between the left and right thigh). The agreement was scored in the range between 0 and 1 as follows: 0 “no agreement”, below 0.4 “poor agreement”, 0.4-0.6 “moderate agreement”, 0.6-0.8 “substantial agreement”, and 0.8-1.0 “almost perfect” [19].

Confidence intervals (CI) of intra-operator DSC values were calculated for each operator as a means to define their global reproducibility performance. A non-parametric test for independent samples (Kruskal-Wallis test) was also performed to compare medians of average intra-operator DSC values. An additional more specific analysis was performed comparing pairwise operators in terms of paired differences between average intra-operator DSC of each muscle. The alpha value was conventionally set as 0.05. Considering the low population size, the p values were adjusted with the Benjamini-Hochberg procedure to decrease false discovery rate.

## Results

The nine enrolled subjects had a median age of 30 years (20-65 y range, 7 F, 2 M). To assess whether the age of the examined subjects had an impact on the reproducibility of muscle segmentation, we also identified two subgroups based on age: a first group of 6 volunteers between 20 and 30 (4 F, 2 M), hereinafter referred to as the young adult subgrou, and a second group of 3 volunteers aged 45 to 65 years (3F) (the adult subgrou). Inter-operator and intra-operator agreement is expressed as the fraction of the overlapping voxels between two comparisons, ranging from 0 to 1.

### *Intra-operator agreement*

OP1 had the highest intra-operator agreement in drawing the ROIs (0.885, CI 0.8845 ± 0.0093) with high DSC in nearly all the muscles of the thigh of the entire cohort, the only exception being the short head of the *biceps femoris* (BFS)(0.749). Intra-operator DSC values respectively for OP2 and OP3 were also quite high, respectively of 0.856 (CI 0.8558 ± 0.0154) and 0.818 (CI 0.8178 ± 0.0194). When examining muscles separately, OP2 showed the lowest reproducibility again for BFS (0.629), whereas OP3 showed the lowest DSC values for *vastus medialis* (VM) (0.765), *semimembranosus (*SM) (0.719) and for BFS (0.550).

When examining the young adult subgroup, OP1 outperformed OP2 and even more OP3 in the reproducibility (DSC respectively of 0.901, 0.851 and 0.801). By contrast, when examining the adult subgroup OP3 showed the highest DSC values (0.897), with OP1 and OP2 having only slightly lower values (0.851 and 0.875, respectively). The pairwise analysis showed a significant difference between the intra-operator DSC of OP1 and OP2 ($p = 0.0483$), OP1 and OP3 ($p = 0.0073$), but no difference between OP2 and OP3 ($p = 0.0806$).

*Inter-operator agreement*

When considering the entire cohort, we found a high (almost perfect) agreement between OP1 and OP2 (0.814) and only substantial agreement between OP1 and OP3 (0.762) and OP2 and 3 (0.702). With regard to single muscles, OP1 and OP2 showed almost perfect agreement in the majority of muscles, with the exception of *semimembranosus* (0.773), *gracilis* (0.747), BFS (0.734) and *sartorius* muscles (0.719). OP1 and OP3 showed a high agreement in most of the muscles, with a substantial agreement of the *gracilis* (0.779), *sartorius* (0.722), *adductor longus* (0.634), *vastus medialis* (0.632), and only fair agreement for BFS segmentation (0.495). OP2 and OP3 had substantial agreement for long head of the *biceps femoris* (BFL) (0.786), *gracilis* (0.700), *adductor longus* (0.626) and *sartorius* (0.626), moderate agreement of the *semimembranosus* (0.599), *vastus medialis* (0.571), and BFS (0.417), as shown in Table I.

When considering the two age subgroups, OP1 and OP2 had slightly higher inter-operator concordance for the young adult subgroup than the adult subgroup (0.827 *vs* 0.775), while higher DSC values were observed in the adult subgroup (compared to young adults) when concordance between OP1 and OP3 (0.775 *vs* 0.751) as well as between OP2 and OP3 (0.785 *vs* 0.680) was evaluated.

## Discussion

We aimed to assess the reproducibility of manual muscle MRI segmentation while applying a coherent, well-defined set of drawing recommendations based on our single-center clinical experience. In this experience focused on a small cohort of healthy volunteers we showed that the application of these recommendations leads to not only a high degree of reproducibility in experienced operators, but also to a significant improvement in reproducibility of relatively inexperienced operators, when trained appropriately.

*Intra-operator reproducibility*

In this study, the highest overall intra-operator reproducibility in ROI drawing was shown for the most expert operator (OP1), which also had the lowest width of the confidence interval. Such a finding reflects the importance of experience in muscle segmentation, which is known to be a strongly operator-dependent process and with a relatively low reproducibility in time.

Among the less experienced operators, OP2 showed higher overall intra-operator agreement compared to OP3, but the pairwise analysis failed to show a significant difference between the distribution of intra-operator DSC values ($p > 0.05$). When examining the confidence intervals for OP2 and OP3, however, we found that they were not overlapping, and that the width of the CI was lower for OP2 compared to that of OP3. If we assume a roughly equal baseline reproducibility between the two less experienced operators, even considering the small dataset, we still can hypothesize that the greater amount of supervised training received actually helped in improving OP2's segmentation reproducibility.

When examining the reproducibility in drawing each muscle separately we found that *semimembranosus* (SM) and the BFS created the higher difficulties to all three operators (also to the expert OP1). We think this might be due to the difficulty in delimitating SM from *adductor magnus* (AM) and, for what concerns BFS, to the small sectional area of the muscle at the considered level for segmentation, compared to other muscles. Arguments could be raised regarding the advantages of omitting certain muscles from the segmentation as BFS in this case, but such decision should follow the purpose of each specific study.

*Inter-operator reproducibility*

The higher inter-operator DSC was shown between OP1 and OP2 (0.814) compared to OP1 and OP3 (0.762) and OP2 and OP3 (0.702). Even considering the small population, the higher degree of training that Operators 2 received seemed effective in increasing the reproducibility of ROI drawing (though, as aforementioned, the difference between the reproducibility between OP2 and OP3 was not statistically significant).

The two muscles with the worst inter-operator reproducibility (average of all operators) were the *sartorius* (0.689), and the BFS (0.642). Whereas the small sectional area of the BFS undoubtedly makes it difficult to obtain a good reproducibility, we interpreted the low results obtained for the *sartorius* as due to difficult belly delimitation from the adjacent semimembranosus (SM), which also presented low inter-operator DSC values (0.699).

*Young versus adult population*

The reproducibility of ROI drawing for the young adult group was lower compared to the adult group when examining intra-operator agreement of OP2 and OP3 and the inter-operator agreement between OP2 and OP3 and also OP1 and OP3. By contrast, OP1 had a higher intra-operator reproducibility for the young adult compared to the adult subgroup and also a higher inter-operator agreement with OP2 compared to the adult subgroup. Such a result was not unexpected and seems to confirm our clinical experience suggesting a higher difficulty in drawing replicable muscle ROIs in younger subjects who have more trophic muscles and where muscle borders may

**Table I.** Average intra-operator and inter-operator DSC test results between ROI segmentation of each thigh muscle at 0 and 72 hours, and average inter-operator DSC test between each operator.

| | | | Single Muscles | | | | | | | | | | | | All Muscles |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **VL** | **VM** | **VI** | **RF** | **Sa** | **G** | **AM** | **SM** | **ST** | **BFL** | **BFS** | **AL** | |
| Average intra-operator DSC value | Operator 1 | Age 20-30 | 0.953 | 0.919 | 0.923 | 0.918 | 0.906 | 0.887 | 0.947 | 0.890 | 0.927 | 0.895 | 0.743 | 0.909 | 0.901 |
| | | Age 45-65 | 0.886 | 0.909 | 0.890 | 0.896 | 0.845 | 0.822 | 0.912 | 0.748 | 0.884 | 0.844 | 0.760 | 0.814 | 0.851 |
| | | Whole cohort | 0.931 | 0.916 | 0.912 | 0.911 | 0.886 | 0.865 | 0.935 | 0.843 | 0.913 | 0.878 | 0.749 | 0.877 | **0.885** |
| | Operator 2 | Age 20-30 | 0.934 | 0.911 | 0.920 | 0.926 | 0.812 | 0.836 | 0.916 | 0.821 | 0.794 | 0.853 | 0.660 | 0.833 | 0.851 |
| | | Age 45-65 | 0.941 | 0.898 | 0.920 | 0.924 | 0.870 | 0.917 | 0.949 | 0.813 | 0.947 | 0.906 | 0.617 | 0.793 | 0.875 |
| | | Whole cohort | 0.937 | 0.908 | 0.917 | 0.925 | 0.822 | 0.853 | 0.924 | 0.809 | 0.832 | 0.863 | 0.629 | 0.851 | **0.856** |
| | Operator 3 | Age 20-30 | 0.904 | 0.714 | 0.866 | 0.920 | 0.790 | 0.791 | 0.860 | 0.740 | 0.859 | 0.868 | 0.550 | 0.749 | 0.801 |
| | | Age 45-65 | 0.945 | 0.906 | 0.945 | 0.946 | 0.872 | 0.911 | 0.958 | 0.733 | 0.948 | 0.930 | 0.807 | 0.867 | 0.897 |
| | | Whole cohort | 0.913 | 0.765 | 0.887 | 0.924 | 0.807 | 0.820 | 0.884 | 0.719 | 0.881 | 0.881 | 0.550 | 0.783 | **0.818** |
| Average intra-operator values (all operators) | | | 0.927 | 0.863 | 0.905 | 0.920 | 0.838 | 0.846 | 0.915 | 0.790 | 0.875 | 0.874 | 0.642 | 0.837 | 0.853 |
| Average inter-operator DSC value | Operator 1 and 2 | Age 20-30 | 0.902 | 0.842 | 0.911 | 0.855 | 0.728 | 0.780 | 0.901 | 0.785 | 0.833 | 0.825 | 0.722 | 0.837 | 0.827 |
| | | Age 45-65 | 0.811 | 0.783 | 0.829 | 0.783 | 0.698 | 0.686 | 0.867 | 0.740 | 0.833 | 0.753 | 0.751 | 0.766 | 0.775 |
| | | Whole cohort | 0.882 | 0.825 | 0.892 | 0.837 | 0.719 | 0.747 | 0.894 | 0.773 | 0.834 | 0.806 | 0.734 | 0.828 | **0.814** |
| | Operator 1 and 3 | Age 20-30 | 0.869 | 0.578 | 0.820 | 0.883 | 0.699 | 0.750 | 0.848 | 0.785 | 0.855 | 0.831 | 0.495 | 0.602 | 0.751 |
| | | Age 45-65 | 0.811 | 0.783 | 0.829 | 0.783 | 0.698 | 0.686 | 0.867 | 0.740 | 0.833 | 0.753 | 0.751 | 0.766 | 0.775 |
| | | Whole cohort | 0.883 | 0.632 | 0.838 | 0.887 | 0.722 | 0.779 | 0.860 | 0.726 | 0.866 | 0.824 | 0.495 | 0.634 | **0.762** |
| | Operator 2 and 3 | Age 20-30 | 0.808 | 0.498 | 0.780 | 0.842 | 0.581 | 0.672 | 0.801 | 0.657 | 0.766 | 0.770 | 0.417 | 0.575 | 0.680 |
| | | Age 45-65 | 0.885 | 0.799 | 0.883 | 0.858 | 0.712 | 0.787 | 0.914 | 0.539 | 0.903 | 0.856 | 0.690 | 0.591 | 0.785 |
| | | Whole cohort | 0.822 | 0.571 | 0.805 | 0.844 | 0.626 | 0.700 | 0.829 | 0.599 | 0.800 | 0.786 | 0.417 | 0.626 | **0.702** |
| Average inter-operator values (all operators) | | | 0.862 | 0.676 | 0.845 | 0.856 | 0.689 | 0.742 | 0.861 | 0.699 | 0.833 | 0.805 | 0.549 | 0.696 | 0.760 |

The latter values represent an average of DSC test results between all combinations of both ROIs of each operator for each muscle and are expressed as an average value between both the left and the right thigh. An average DSC value was provided in the last column as a single value that summarizes spatial overlap of all ROIs. A separate row underneath each sub-table reports the average single-muscle DSC values both for intra-operator and inter-operator tests. The colors represent the agreement class for each value, with dark green representing near perfect agreement (DSC values ranging between 0.8 and 1.0), light green representing substantial agreement (DSC value between 0.6 and 0.8) and yellow representing moderate agreement (DSC value between 0.4 and 0.6). *VL: Vastus lateralis; VM: Vastus medialis; VI: Vastus Intermedius; RF: Rectus femoris; Sa: Sartorius; G: gracilis; AM: Adductor Magnus; SM: Semimembranosus; ST: Semitendinosus; BFL: Long head of Biceps femoris; BFS: Short head of Biceps femoris; AL: Adductor Longus.*

be uncertain between different muscle bellies, leading to a decrease of intra- and inter-operator agreement, especially for less experienced operators. We surmise that the score difference between the two subgroups may be due to inter-muscular adipose tissue being more prominent in the adult group, with fat acting as a natural contrast facilitating the segmentation process. An example of this can be seen in OP3's lower repeatability in segmenting the VM muscle, which is usually clearly distinguishable. Upon further inspection, this error was more pronounced in the younger subgroup, due to the fact that the demarcation between the VM and the vastus intermedius (VI) muscles is less obvious in the younger population.

Initial stages of certain myopathies may have a similar appearance, rendering segmentation slightly easier, whereas advanced stages might have the opposite effect, since the extensive fatty infiltration might render the localization of muscles even more difficult. The most expert operator (OP1) by contrast showed a higher reproducibility in the youngest subcohort, presumably reflecting a higher ability in correctly separating muscle bellies. This finding was also mirrored by the higher reproducibility between OP1 and OP2 in the younger cohort compared to the older one.

These data are in line with our data-driven hypothesis of a sort of "U-shaped" curve of difficulty in drawing ROIs in the field of neuromuscular diseases, with higher difficulty presenting when adipose replacement is very low or very high; this observation is in line with our clinical experience.

*The ROI-drawing process: our experience*

Manual muscle segmentation can be classified into three approaches: segmentation of all muscles as a single ROI, segmentation of each muscle compartment and segmentation of each single muscle separately. The latter approach was the one evaluated in this study. The main disadvantage of single-muscle segmentation is the time required to draw all the ROIs, making it difficult for a single dedicated operator to apply such an approach in routine care (especially when segmentation is applied to several slices to cover a wide range of selected muscles). Multiple operators that are part of the same research (or clinical) center are consequently frequently used to hasten such a procedure. This underlines the importance of establishing clear segmentation criteria to maximize spatial overlap between operators. As software based on machine learning are progressively applied in the segmentation process, segmentation speed might soon, however, not be a relevant issue. In order to develop such algorithms, however, the software must learn from previous adequate and highly reproducible ROI datasets. Algorithms training times are also reduced when datasets received as inputs are more coherent and more representative of the "ground truth" that can be obtained preferably by consensus between human operators with an expert knowledge of muscle segmentation. In this sense, one of the main variables influencing reproducibility of muscle ROIs is the distance of ROI margin from the muscle fascia. Few studies in recent literature specify in sufficient detail the inclusion of anatomical muscle borders or lack thereof [20,21]. One of the exceptions is a recent study [4] that not only explicitly specifies to hold adequate distance to the muscle fascia to avoid chemical shift artifacts but also compared inter-operator and intra-operator agreement [4]. The majority of other studies (including those with segmentation performed by more than one operator) only measured inter-operator agreement, often by measuring the intraclass correlation coefficient (ICC) [22,23]. Even if a high ICC agreement is generally considered to be reliable for follow-up evaluations of fatty infiltration [12,24], we have a stronger interest in the reproducibility of the ROI drawing. The position and the morphology of ROIs, in fact, can be highly relevant when it comes to evaluating the degree of pathology, especially in longitudinal follow-up.

We therefore consider the specification of the distance between the ROI borders and the muscle fascia as the most important factor that influences ROI reproducibility.

*Study limitations*

Several limitations apply to our study. Firstly, in this study we analyzed the ROI segmentation process of the thigh muscles, as this region is the most frequently studied in the field of neuromuscular disease. Though the focus on a single muscle group represents a limitation of this study, we believe that the presented recommendations are applicable to most muscle groups in several contexts, though we hope to test this hypothesis in future studies.

Secondly, the small number of operators dedicated to ROI drawing and the small numbers of subjects included are certainly to be underlined as limiting factors for generalization of results. Thirdly, the main parameter by which we evaluated the validity of this set of recommendations is reproducibility. Due to the aforementioned lack of a standard set of recommendations concerning manual segmentation, we chose not to perform an accuracy analysis and to concentrate our work on reproducibility by assessing DSC values. A clearly identifiable consensus or guideline for segmentation with muscle volume as its primary objective, is still lacking. As such, a gold standard for comparison with the presented results is missing. A high level of precision and reproducibility is nonetheless highly desirable in the clinical context of ROI drawing, especially when follow-up MRI studies are performed. Such an element becomes particularly relevant in subjects suffering from dystrophic diseases, in which the possibility to detect even small changes over time have an impact on improving the prediction of the patient's prognosis [25].

## Conclusions

In this study we showed how, following specific recommendations, the process of muscular ROI drawing can be highly reproducible and reliable; less experienced operators dedicated to ROI drawing can be trained with specific instructions with promising results in order to gather vast ROI datasets for analysis and as reference. Though only thigh muscles were selected as the target region to be segmented in this study, these same results can be generalized to other anatomical districts and applied to muscles that are more difficult to segment. The measure of distance from the muscle fascia in the drawing process appears, in our opinion, as the main criterion to follow in order to guarantee the reproducibility of ROIs, as well as a major cause of inter-operator low agreement in the cases in which it is not adequately defined. Secondary recommendations include avoiding internal fascia invaginations, while including areas of fatty muscle degeneration within the borders of the muscle fascia. We believe that following these recommendations allows for increased reproducibility of manual segmentation in muscle imaging, especially in the case of young adults, in which the intermuscular connective tissue is less prominent and delimitation between certain muscles is not straightforward.

*Ethical consideration*

All subjects gave their formal informed consent for the participation in the study, which was approved by the local ethics committee.

*Funding*

This work was supported by grants of the Italian Ministry of Health (Ricerca Corrente 2017-2019 and Ricerca Corrente 2020).

*Financial interest*

The Authors have no relevant financial or non-financial interests to disclose.

*Conflict of interest*

The Authors have no conflicts of interest to declare that are relevant to the content of this article.

*Acknowledgments*

This work was supported by grants of the Italian Ministry of Health (Ricerca Corrente 2017-2019 and Ricerca Corrente 2020).

*Author contributions*

A.P. and S.B. contributed to the conceptualization and supervision of this study. F..S, E.B. and R.V. contributed to the investigation, and resource collection. M.P. and N.B. contributed to the methodology, formal analysis, and data curation. S.I.M., M.P. and A.P. contributed to the writing of the original draft, while A.P. and S.B. contributed to the review and editing of the final draft.

## References

1 Bonati U, Hafner P, Schädelin S, et al. Quantitative muscle MRI: A powerful surrogate outcome measure in Duchenne muscular dystrophy. Neuromuscul Disord 2015;25:679-685. https://doi.org/10.1016/j.nmd.2015.05.006

2 Duijnisveld BJ, Henseler JF, Reijnierse M, et al. Quantitative Dixon MRI sequences to relate muscle atrophy and fatty degeneration with range of motion and muscle force in brachial plexus injury. Magn Reson Imaging 2017;36:98-104. https://doi.org/10.1016/j.mri.2016.10.020

3 Carlier PG. Global T2 versus water T2 in NMR imaging of fatty infiltrated muscles: different methodology, different information and different implications. Neuromuscul Disord 2014;24:390-392. https://doi.org/10.1016/j.nmd.2014.02.009

4 Schroeder J, Tobler P, Stalder A-L, et al. Intra-rater and inter-rater reliability of quantitative thigh muscle magnetic resonance imaging. Imaging Med 2019;11.

5 Rodrigues R, Pinheiro AMG. Segmentation of skeletal muscle in thigh dixon MRI Based on texture analysis (Published online 2019. http://arxiv.org/abs/1904.04747).

6 Ghosh S, Boulanger P, Acton ST, et al. Automated 3D muscle segmentation from MRI data using convolutional neural network (Published online 2017). https://doi.org/10.1109/ICIP.2017.8297121.

7 Mesbah S, Shalaby AM, Stills S, et al. Novel stochastic framework for automatic segmentation of human thigh MRI volumes and its applications in spinal cord injured individuals. PLoS One 2019;14. https://doi.org/10.1371/journal.pone.0216487

8 Kim HK, Lindquist DM, Serai SD, et al. Magnetic resonance imaging of pediatric muscular disorders. Recent advances and clinical applications. Radiol Clin North Am 2013;51:721-742. https://doi.org/10.1016/j.rcl.2013.03.002

9 Larmour S, Chow K, Kellman P, et al. Characterization of T1 bias in skeletal muscle from fat in MOLLI and SASHA pulse sequences: quantitative fat-fraction imaging with T1 mapping. Magn Reson Med 2017;77:237-249. https://doi.org/10.1002/mrm.26113

10 Kim HK, Serai S, Lindquist D, et al. Quantitative skeletal muscle MRI: part 2, MR spectroscopy and T2 relaxation time mapping-comparison between boys with duchenne muscular dystrophy and healthy boys. Am J Roentgenol 2015;205:W216-W223. https://doi.org/10.2214/AJR.14.13755

11 Forbes SC, Willcocks RJ, Triplett WT, et al. Magnetic resonance imaging and spectroscopy assessment of lower extremity skeletal muscles in boys with duchenne muscular dystrophy: a multicenter cross sectional study. PLoS One 2014;9. https://doi.org/10.1371/journal.pone.0106435

12 Fischmann A, Hafner P, Fasler S, et al. Quantitative MRI can detect subclinical disease progression in muscular dystrophy. J Neurol 2012;259:1648-1654. https://doi.org/10.1007/s00415-011-6393-2

13 Hooijmans MT, Niks EH, Burakiewicz J, et al. Non-uniform muscle fat replacement along the proximodistal axis in Duchenne muscular dystrophy. Neuromuscul Disord 2017;27:458-464. https://doi.org/10.1016/j.nmd.2017.02.009

14 Garrood P, Hollingsworth KG, Eagle M, et al. MR imaging in Duchenne muscular dystrophy: quantification of T 1-weighted signal, contrast uptake, and the effects of exercise. J Magn Reson Imaging 2009;30:1130-1138. https://doi.org/10.1002/jmri.21941

15 Orgiu S, Lafortuna CL, Rastelli F, et al. Automatic muscle and fat segmentation in the thigh from T1-Weighted MRI. J Magn Reson Imaging 2016;43:601-610. https://doi.org/10.1002/jmri.25031

16 Keller S, Wang ZJ, Aigner A, et al. Diffusion tensor imaging of dystrophic skeletal muscle: comparison of two segmentation methods adapted to chemical-shift-encoded water-fat MRI. Clin Neuroradiol 2019;29:231-242. https://doi.org/10.1007/s00062-018-0667-3

17 Schlaeger S, Freitag F, Klupp E, et al. Thigh muscle segmentation of chemical shift encoding-based water-fat magnetic resonance images: the reference database MyoSegmenTUM. PLoS One 2018;13:1-19. https://doi.org/10.1371/journal.pone.0198200

18 Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006;31:1116-1128. https://doi.org/10.1016/j.neuroimage.2006.01.015

19 Pajula J, Kauppi J, Tohka J. Inter-subject correlation in fMRI : method validation against stimulus-model based analysis. PLoS One 2012;8:e41196. https://doi.org/10.1371/journal.pone.0041196

20 Wary C, Azzabou N, Giraudeau C, et al. Quantitative NMRI and NMRS identify augmented disease progression after loss of ambulation in forearms of boys with Duchenne muscular dystrophy. NMR Biomed 2015;28:1150-1162. https://doi.org/10.1002/nbm.3352

21 Ponrartana S, Ramos-Platt L, Wren TAL, et al. Effectiveness of diffusion tensor imaging in assessing disease severity in Duchenne muscular dystrophy: preliminary study. Pediatr Radiol 2015;45:582-589. https://doi.org/10.1007/s00247-014-3187-6

22 Janssen B, Voet N, Geurts A, et al. Quantitative MRI reveals decelerated fatty infiltration in muscles of active FSHD patients. Neurology 2016;86:1700-1707. https://doi.org/10.1212/WNL.0000000000002640

23 Mul K, Vincenten SCC, Voermans NC, et al. Adding quantitative muscle MRI to the FSHD clinical trial toolbox. Neurology 2017;89:2057-2065. https://doi.org/10.1212/WNL.0000000000004647

24 Fischer MA, Pfirrmann CWA, Espinosa N, et al. Dixon-based MRI for assessment of muscle-fat content in phantoms, healthy volunteers and patients with achillodynia: comparison to visual assessment of calf muscle quality. Eur Radiol 2014;24:1366-1375. https://doi.org/10.1007/s00330-014-3121-1

25 Carlier PG, Marty B, Scheidegger O, et al. Skeletal muscle quantitative nuclear magnetic resonance imaging and spectroscopy as an outcome measure for clinical trials. J Neuromuscul Dis 2016;3:1-28. https://doi.org/10.3233/JND-160145