Evaluation of interrater reliability of different muscle segmentation techniques in diffusion tensor imaging

Johannes Forsting1 | Robert Rehmann1 | Marlena Rohm1 | Martijn Froeling2 | Lara Schlaffke1

1Department of Neurology, BG-University Hospital Bergmannsheil, Ruhr-University Bochum, Bochum, Germany
2Department of Radiology, University Medical Centre Utrecht, Utrecht, The Netherlands

Introduction: Muscle diffusion tensor imaging (mDTI) is a quantitative MRI technique that can provide information about muscular microstructure and integrity. Ultrasound and DTI studies have shown intramuscular differences, and therefore separation of different muscles for analysis is essential. The commonly used methods to assess DTI metrics in muscles are manual segmentation and tract-based analysis. Recently methods such as volume-based tractography have been applied to optimize muscle architecture estimation, but can also be used to assess DTI metrics.

Purpose: To evaluate diffusion metrics obtained using three different methods—volume-based tractography, manual segmentation-based analysis and tract-based analysis—with respect to their interrater reliability and their ability to detect intramuscular variance.

Materials and methods: 30 volunteers underwent an MRI examination in a 3 T scanner using a 16-channel Torso XL coil. Diffusion-weighted images were acquired to obtain DTI metrics. These metrics were evaluated in six thigh muscles using volume-based tractography, manual segmentation and standard tractography. All three methods were performed by two independent raters to assess interrater reliability by ICC analysis and Bland-Altman plots. Ability to assess intramuscular variance was compared using an ANOVA with muscle as a between-subjects factor.

Results: Interrater reliability for all methods was found to be excellent. The highest interrater reliability was found for volume-based tractography (ICC ≥ 0.967). Significant differences for the factor muscle in all examined diffusion parameters were shown in muscles using all methods (main effect p < 0.001).

Conclusions: Diffusion data can be assessed by volume tractography, standard tractography and manual segmentation with high interrater reliability. Each method...
produces different results for the investigated DTI parameters. Volume-based tractography was superior to conventional manual segmentation and tractography regarding interrater reliability and detection of intramuscular variance, while tract-based analysis showed the lowest coefficients of variation.

**KEYWORDS**

interrater reliability, manual segmentation, MRI, muscle diffusion tensor imaging (mDTI), tractography

1 | INTRODUCTION

Quantitative muscle diffusion tensor imaging (mDTI) is a possible biomarker in the diagnosis and monitoring of neuromuscular diseases. Certain neuromuscular diseases have specific patterns of muscle involvement that can only be distinguished by analyzing individual muscles. To be able to extract whole-muscle DTI data and capture muscle specific diffusion changes, muscles must be delineated and reconstructed prior to muscle specific DTI analysis. Since the incidence of neuromuscular diseases is low, standardization of analysis protocols—especially muscle segmentation—is needed to sufficiently pool data from different centers.

The muscle segmentation must be done manually as there are currently no reliable automated protocols for fully automated muscle segmentation. Currently there are three methods to perform muscle segmentation. The most popular technique is the manual segmentation-based analysis (MSB). In MSB analysis muscle contours of individual muscles are delineated on every slice of T1-weighted (T1w) images, which results in a three-dimensional muscle volume. By superimposing these muscle volumes on mDTI maps the diffusion metrics of the voxels within these masks can be extracted and analyzed.

Muscle segmentation can also be done using mDTI-based fiber tractography or “tract-based analysis” (TBA). In TBA, T1w images are registered on diffusion weighted imaging (DWI) images. Using the T1w images multiple selection regions of interest (ROIs) are defined to select tracts obtained by whole volume tractography or the entire muscle. TBA has the advantage to additionally deliver information about muscle macrostructure due to the visualization of muscle fiber tracts. Additionally, DTI metrics can be obtained for whole muscle structure as with MSB. Still, in terms of anatomical accuracy, there have been some concerns, because reconstructed tracts often continue along the aponeurosis where physiological muscle fibers usually stop.

The application of “anatomical constraints” for fiber tractography was examined by Bolsterlee et al to optimize anatomical accuracy. Limiting tractography to a previously defined muscle volume showed reliable measurements of architecture parameters in human calf muscles. This is the basis for the third concept of muscle delineation, called “volume-based tractography” (VBT). This method combines the manual and tract-based segmentation. For VBT the delineation of individual muscle volumes must be done manually on a T1w dataset equal to MSB segmentation. The muscle volumes are then registered to a corresponding DTI dataset on which whole muscle tractography is performed with the tracts constrained by the segmented muscle volume. VBT has the advantage that the tractography is constrained to the defined anatomically correct muscle volume. The muscle delineation on a T1w image is more precise than with the tractography analysis, and muscle tendons and connective tissue can be excluded more easily. The resulting muscle volumes that are registered to the DTI dataset thus contain only the contractile muscle volume without connective tissue.

All three segmentation methods have been previously used in studies in the assessment of mDTI data. However, little is known about how they compare. Criteria for a good assessment in healthy subjects typically include a high interrater reliability, which allows comparison between different studies, pooling of data and data analysis by different raters. To date there are only a few studies assessing interrater reliability in muscle segmentation. Ponrartana et al showed high interrater agreement in the evaluation of healthy children’s leg muscles with tractography (intraclass-correlation coefficient, ICC > 0.93). Källin et al analyzed DTI data of rotator cuff muscles by manual segmentation using ROIs and with tractography. For manual segmentation excellent interrater reliability for fractional anisotropy (FA) and mean diffusivity (MD) was shown (ICC > 0.92), while tractography showed moderate results (ICC > 0.71). Manual segmentation and tractography results correlated positively (r > 0.628).

Furthermore, muscles of human thigh have different architectures shown by different anatomical, ultrasound and DTI studies. This results in differences of DTI parameters of different muscles, which are commonly observed in different studies. Therefore, we assume that an analysis method should be able to reflect these differences between muscles in the assessment of healthy participants.

The aim of this study was to compare the three described methods of muscle segmentation, VBT, MSB and TBA, in a group of 30 healthy volunteers. For each method we investigated (i) their interrater dependence and (ii) their ability to detect intramuscular variance.
2 | MATERIALS AND METHODS

2.1 | Study population

Thirty healthy subjects (15 females) participated in this study. Mean age was 31 years (SD 7 years, range 21-53). Inclusion criteria were no medical history of muscle or nerve disease, no strength exercise five days prior to examination, no prior injuries of human calf or thigh 12 months prior to study enrollment and no regular medication except for oral contraceptives. The study protocol was approved by the local ethics committee (Ruhr University Bochum No 15-5,281).

2.2 | Data acquisition

MRI was performed using a 3 T MRI system (Achieva 3 T X, Philips) and a 16-channel torso XL coil. Participants were instructed to lie still in a feet-first supine position. The MRI protocol included proton density-weighted (PD), T2-weighted (T2w) and DWI in an axial slice order from proximal to distal (total acquisition time 27 min). To avoid shimming artifacts due to large field of view (FOV), the upper leg region was divided into three FOVs of 480 × 264 × 150 mm3 along the z-axis (stacks). For accurate merging the stacks had an overlap of 10 mm.

High resolution anatomical images were acquired using a PD sequence using a turbo spin echo (TSE) readout with the following parameters: voxel size 1.5 × 1.5 × 3.0 mm³; repetition time/echo time (TR/TE) 1646/15 ms. A T2w TSE sequence with fat suppression (SPAIR) followed: voxel size 1.5 × 1.5 × 3.0 mm³; TR/TE 11 422/53 ms. Finally, a diffusion-weighted spin echo-echo planar imaging acquisition was performed using the following parameters: voxel size 3.0 × 3.0 × 6.0 mm³; TR/TE 3819/46 ms; SPAIR fat suppression; SENSE: 2; 18 gradient directions with b = 400 s/mm² and three non-diffusion-weighted images (b = 0 s/mm²). A noise map was obtained, using the same imaging parameters as the DWI, but without RF power and gradients (only acquisition channels open).

2.3 | Data processing

Data was preprocessed similarly to Schlaffke et al using QMRITools software (Mathematica 11). In short, DWI images were merged, denoised and motion corrected by registration to the T2 image. The diffusion data was corrected for subject motion and eddy current distortions using affine registration and aligned with T2w data using non-rigid registration (1000 iterations, b-spline spacing 120, 80, 80) including the rotation of the b-matrix. Tensors were estimated using the MATLAB-based toolbox ExploreDTI using iterative weighted least squares with outlier rejection (REKINDLE).

2.4 | Muscle segmentation

Segmentation of six thigh muscles (biceps femoris, rectus femoris, semimembranosus, semitendinosus, vastus lateralis and vastus medialis) was performed by two independent raters using three methods (for an overview see Figure 1).

2.4.1 | Method 1 (MSB)

The muscles were manually segmented slice by slice on the PD image, avoiding subcutaneous fat and fascia (3D-slicer 4.4.0, https://www.slicer.org). The manual segmentations were smoothed and eroded by one voxel to avoid partial volume effects of non-muscular tissue and registered to the diffusion metrics of FA, MD, λ1 and radial diffusivity (RD) for each muscle. The segmentations were then registered to the DWI data using sequential rigid and b-spline transformations (elastix, http://elastix.isi.uu.nl).

2.4.2 | Method 2 (TBA)

Deterministic whole muscle volume tractography of the entire upper leg was performed using ExploreDTI with a seed grid of 3.0 × 3.0 × 3.0 mm³. The fiber tracking was performed using an angle to step ratio of 10°/mm (with a maximum angle 15° and step size 1.5 mm) and an
FA range of 0.1–0.6. The PD data was first registered to the DWI data using sequential rigid and b-spline transformations. Using these co-registered DWI and PD data, ROIs in the form of “seed” gates were manually drawn and used to select the tracts belonging to each of the six upper leg muscles of interest. The muscle DTI parameters were extracted for each individual muscle using tract-based sampling. Mean diffusion metrics were obtained by averaging the tracts at each point along all muscle fiber tracts.

2.4.3 Method 3 (VBT)

VBT is a combination of MSB and TBA. The segmentations from the MSB were registered to the diffusion space using sequential rigid and b-spline transformations (elastix). The preprocessed diffusion data were segmented based on the segmentations and whole muscle tractography was performed only within the resulting segments of diffusion data using the same parameters as the TBA. Diffusion metrics were then obtained using tract-based sampling, similar to TBA.

2.5 Statistical analysis

All statistical analyses were performed using IBM SPSS V24. Correlation analysis (Pearson correlation coefficient, ICC) and Bland-Altman plots were used to assess interrater reliability. To determine differences in DTI-derived parameters (FA, MD, $\lambda_2$, RD) between different methods (VBT, MSB, TBA) we used analysis of variance (ANOVA). For significant main effects, post hoc t-tests with Bonferroni correction for multiple comparisons were performed. To determine intramuscular differences an ANOVA was done with muscle as a between-subjects factor. Effect size was calculated using partial eta squared ($\eta^2$). The square of the correlation ratio ($\eta^2$) measures the proportion of the variation of DTI values in muscles between subjects. The significance level for all tests was set at $p < 0.05$. We also calculated the coefficient of variation for all three methods as standard deviation/mean value.

3 RESULTS

All scans were successfully performed in all participants. After visual inspection one data set had to be excluded due to motion artefacts. Manual segmentation and DTI-based tractography were successfully used to separate six different thigh muscles (biceps femoris, rectus femoris, semimembranosus, semitendinosus, vastus lateralis, vastus medialis) in all remaining datasets.
Using all three methods we observed FA values between 0.12 and 0.44 (mean 0.20-0.22), and MD values between 1.03 and 1.98 (mean 1.64-1.66) for all six thigh muscles (see Table 1). Coefficients of variance were lowest for TBA and highest for MSB (see Table 2). To illustrate the interrater variance, we also displayed our interrater data on FA, MD, $\lambda_1$ and RD as boxplots in Figure S1. Boxplot diagrams comparing FA, MD, $\lambda_1$ and RD (N = 29) for each examined muscle of the human upper leg and each method are shown in Figure 2.

When evaluating differences of diffusion metrics for VBT, MSB and TBA, we found significant differences for methods in FA, MD and RD (main effect $p < 0.001$), but not for $\lambda_1$ (main effect $p = 0.358$; see Table 3). Post hoc tests showed significant differences between all methods for FA AND between TBA and MSB for MD ($p < 0.05$). For RD no significant difference between MSB and VBT was found, while TBA differs from both MSB and VBT ($p < 0.05$).

In the assessment of rater dependence, interrater reliability for all methods was found to be good (ICC ≥ 0.864). The highest interrater reliability for all diffusion parameters was found for VBT, which showed excellent agreement (ICC ≥ 0.967). The scatter plots in Figure 3 illustrate the correlation of interrater measurements including Pearson correlation coefficient ($r$) and ICC in each graph. The Pearson correlation coefficient showed strong correlation between raters ($r ≥ 0.775$). The highest correlation between raters was found for VBT ($r ≥ 0.937$).

Figure 4 depicts the Bland-Altman plots showing differences between interrater measurements of FA, MD, $\lambda_1$ and RD for VBT, MSB and TBA. Bland-Altman plots did not reveal any bias between raters. The smallest limits of agreement (LoA) were observed for VBT [LoA [FA] = −0.012; 0.018; LoA [MD] = −0.062; 0.064; LoA [$\lambda_1$] = −0.078; 0.093; LoA [RD] = −0.058; 0.055; see Figure 4]. Typical muscle shapes and volumes of the three methods and differences between raters are visualized in Figure 5.

In the evaluation of intramuscular differences, we found significant differences for muscle in all examined diffusion parameters, FA, MD, $\lambda_1$ and RD, using all three methods, ie VBT, MSB and TBA (main effect $p < 0.001$). Highest effect sizes were consistently found for VBT ($\eta^2$[FA] = 0.656; $\eta^2$[MD] = 0.277; $\eta^2$[$\lambda_1$] = 0.452; $\eta^2$[RD] = 0.350; see Figure 2).

## DISCUSSION

Muscle segmentation is important for the analysis of DTI data. Thigh muscles differ in fiber length, fiber angle and muscle volume, and architectural intermuscular differences also result in different muscle DTI parameters. To analyze muscle specific diffusion changes and to be able to compare them with other studies reliable muscle segmentation protocols are necessary.

In this study we evaluated the stability and interrater reliability of VBT, MSB and TBA for muscle segmentation. All methods were able to segment the thigh muscles with good to excellent interrater reliability. Furthermore, all methods also showed significant variance for diffusion metrics in different muscles. The highest effect sizes were shown for VBT.

However, significant differences were shown for the main diffusion parameter FA between all methods, possibly due to the different weighting of diffusion information within the muscle volume. For MD significant differences were only found between TBA and MSB. This may be explained since VBT is a combination of TBA and MSB and therefore has similarities with both techniques. No significant main effect was found for $\lambda_1$ (also referred to as AD). RD values showed significant differences between VBT, MSB and TBA and therefore seem to be more susceptible to different data analysis methods compared with $\lambda_1$. Both RD and FA are susceptible to low-signal-to-noise-ratio (SNR) regions.

### TABLE 1
Overview of DTI parameters obtained using three different segmentation techniques

| Method | FA | MD | $\lambda_1$ | RD |
|--------|----|----|-------------|----|
| VBT    | 0.21 ± 0.04 | 1.65 ± 0.09 | 2.04 ± 0.14 | 1.46 ± 0.09 |
| TBA    | 0.20 ± 0.03 | 1.66 ± 0.09 | 2.03 ± 0.12 | 1.47 ± 0.09 |
| MSB    | 0.22 ± 0.05 | 1.64 ± 0.11 | 2.04 ± 0.16 | 1.44 ± 0.11 |

### TABLE 2
Coefficients of variation for three different segmentation techniques

| Method | FA | MD | $\lambda_1$ | RD |
|--------|----|----|-------------|----|
| VBT    | 0.1905 | 0.0545 | 0.0686 | 0.0616 |
| TBA    | 0.1500 | 0.0542 | 0.0591 | 0.0612 |
| MSB    | 0.2272 | 0.0670 | 0.0784 | 0.0763 |
According to Fieremans et al, RD would change due to myofiber diameters while AD would remain unaffected. Therefore, a potential cause of variance may also be the result of sensitivity to fiber diameters in RD and FA, which cannot be detected in AD.

The main advantage of MSB is its simplicity: PD masks are superimposed on DTI images and no additional tractography algorithms are required. However, MSB, similarly to TBA and VBT, is susceptible to poor data quality when SNR is low. In MSB each voxel is weighted equally, meaning that diffusion information is measured once for each voxel and voxels with low quality are weighted equally to reliable voxels. Thus, in comparison to TBA—where voxel diffusion information is counted multiple times depending on how many fiber tracts travel through it—fewer
TABLE 3  $p$-values of differences in diffusion metrics between segmentation techniques

|                  | ANOVA Main effect | Post hoc t-tests VBT-MSB | VBT-TBA | MSB-TBA |
|------------------|-------------------|--------------------------|---------|---------|
| FA               | <0.001            | 0.035                    | 0.01    | <0.001  |
| MD               | 0.033             | 0.579                    | 0.283   | 0.038   |
| $\lambda_1$     | 0.358             | 1                        | 1       | 0.572   |
| RD               | <0.001            | 0.055                    | 0.004   | <0.001  |

FIGURE 3  Scatter plots of correlation between interrater measurements for VBT, MSB and TBA with $r$ and ICC included in each graph
data points are taken into account for analysis and all are weighted equally. With fewer data points MSB is more susceptible to a higher interrater variance and to partial volume effects (Figure 4).

Similarly to MSB, TBA achieves robust results in the assessment of intramuscular variability and rater dependence; however, in contrast to MSB, TBA weights the diffusion information. In regions with a continuous reliable diffusion signal, e.g., in the tightly packed muscle belly of a healthy muscle, more and longer fiber tracts are generated as in regions where the diffusion data is of low quality or at the muscle border. Therefore, fiber density is low in unreliable areas; however, it can increase in muscle aponeurosis and fascia. Regions within a muscle with high fiber tract density thus are overrepresented and influence the result of total muscle diffusion data. This could result in a lower influence of partial volume effects, which is supported by the lower coefficients of variation in TBA. These low coefficients of variation suggest that TBA has the highest accuracy of the examined methods, although it is important to keep in mind that, since we did not compare our data with muscle biopsies, we cannot be certain that TBA more closely approximates the ground than the other segmentation methods. However, it has been shown that in high quality data tract density is very constant in muscle. Nonetheless, TBA is susceptible to SNR, since regions with low SNR produce fewer and shorter fiber tracts and an SNR > 25 in DWI is necessary to achieve reliable tractography results. The fiber tractography outcome in TBA furthermore depends on the selection of predefined stop parameter settings (e.g., step size, maximum curvature angle, FA limits). Another limitation of TBA is that the segmentation is performed based on DWI images that are superimposed on a fitted T1w image. Thus, in TBA the segmentation is influenced by changes of acquisition parameters, postprocessing procedures and tractography algorithms. In contrast to MSB, TBA provides additional visual and quantitative information—such as fiber tract length, tract density, tract count, pennation angle—about muscle

FIGURE 4  Bland-Altman plots of interrater measurements for VBT, MSB and TBA. The x-value shows the mean of two raters and the y-value the difference between the raters. The colored line shows the mean of the paired difference; the black lines show LoAs from $-1.96 \, s$ to $1.96 \, s$
macrostructure by reconstructing fiber tracts that reflect the individual architecture of skeletal muscle fibers in healthy and diseased muscles. A major limitation of TBA is that reconstructed tracts often continue—physiologically implausible—beyond the aponeurosis and can be reconstructed to other muscles. Therefore diffusion information is not solely derived from muscle tissue or the muscle of interest.

VBT combines the anatomically accuracy of MSB with the advantages of fiber tractography of the TBA method. By delineating individual muscles on a PD slice, non-muscle tissue (aponeurosis, blood vessels, nerves) can be excluded and tractography is done only in muscle tissue. In this study we showed that VBT was able to separate different human thigh muscles with higher effect size than MSB and TBA. Although TBA had a slightly lower coefficient of variation in all diffusion parameters, interrater reliability for VBT was found to be excellent (ICC ≥ 0.92) and LoAs in Bland-Altman plots were smaller in comparison with TBA and MSB. These findings support a high precision of VBT, which is necessary to ensure comparability between studies. A recently published study by Bolsterlee et al used a technique called "anatomically constrained DTI tractography," which is comparable to VBT. Limiting the tractography algorithm to a previously defined volume, this method was not as susceptible as TBA regarding stop criterion settings of tractography and their influences on mDTI parameters. Sahrmann et al also successfully applied this technique in soleus muscles in patients with cerebral palsy and healthy controls. Therefore, VBT allows us to additionally obtain more accurate information about muscle macrostructure with less dependence on fiber tracking stop criteria. However, if the anatomical segmentations on high resolution anatomical images are not properly aligned with the diffusion data on DWI this can result in erroneous tractography and quantification, which is a limitation of VBT.

The most important limitation in muscle segmentation is the time- and cost-intensive manual segmentation process. All three presented methods require a manual muscle segmentation, which is a limitation to clinical applicability. In healthy muscles automatic and semi-automatic muscle segmentation methods are under investigation, but there is still no extensive use of these methods. As segmentation methods were only studied in healthy participants, further validation of optimal segmentation approach in diseased muscles is needed. If such methods become available, they may reduce the time of analysis in MSB and VBA analysis considerably.

In diseased skeletal muscle, eg in several muscular dystrophies, fatty infiltration influences fiber tractography and DTI metrics. In muscle diseases where advanced fatty degeneration is a key aspect of disease manifestation VBT could be a method to consider. With VBT the viable muscle segmentation can be done more accurately as with TBA, and the tractography is limited to a predefined volume. This higher anatomically accuracy could allow tractography-based analysis despite fatty infiltration, since the predefined muscle borders restrict the tractography algorithm and prevent detection of implausible fiber tracts. However, it should be noted that fiber tractography in fat infiltrated muscle can be challenging due to the typically lower SNR resulting from fat partial volume effects.

Our results show that using different segmentation methods provides different results of mDTI metrics in muscle specific analysis. Thus, current publications regarding muscle specific mDTI values are hard to compare. In future studies it is important to state what muscle segmentation method was used to gain mDTI values.
The highest interrater reliability could be reached with VBT compared with the other segmentation methods. When examining muscular diseases in future studies, the higher effect size of VBT in mDTI values between muscle groups could be beneficial. Our data supports the assumption that whole muscle DTI analysis is best performed with the VBT approach regarding precision and with TBA regarding accuracy. These approaches should also be evaluated in multi-center and clinical studies.

5 | CONCLUSION

In conclusion, we have shown that DTI metrics can be assessed by all three methods with good to excellent interrater variability. VBT was superior to conventional manual segmentation and tractography regarding interrater reliability and detection of intramuscular variance, while TBA showed the lowest coefficients of variation.

Neuromuscular disorders are rare diseases. This makes it difficult to collect sufficient data from different individuals in a single center. Further standardization of analysis protocols is needed to ensure comparable data quality between individual centers. The results of this study show that VBT is a promising technique for standardizing mDTI analysis in future studies.

ACKNOWLEDGEMENTS

We thank Philips Germany for continuous scientific support and specifically Dr. Burkhard Mädler for valuable discussion. JF and LS received funding from the Deutsche Forschungsgemeinschaft SFB874 (TP-A1 to JF, TP-A5 to LS). Open access funding enabled and organized by Projekt DEAL.

ORCID

Johannes Forsting https://orcid.org/0000-0001-6647-3167
Martijn Froeling https://orcid.org/0000-0003-3841-0497
Lara Schlaffke https://orcid.org/0000-0002-0716-3780

REFERENCES

1. Hooijmans MT, Damon BM, Froeling M, et al. Evaluation of skeletal muscle DTI in patients with Duchenne Muscular Dystrophy. NMR Biomed. 2015; 28(11):1589-1597.
2. Maggi L, Moscatelli M, Frangiamore R, et al. Quantitative muscle MRI protocol as possible biomarker in Becker Muscular Dystrophy. Clin Neuroradiol. 2020. https://doi.org/10.1007/s00062-019-00875-0
3. Wattjes MP, Kley RA, Fischer D. Neuromuscular imaging in inherited muscle diseases. Eur Radiol. 2010;20(10):2447-2460.
4. Díaz-Manera J, Llauger J, Gallardo E, Illa I. Muscle MRI in muscular dystrophies. Acta Myol. 2015;34(2-3):95-108.
5. Okamoto Y, Kunimatsu A, Miki S, Shindo M, Niitsu M, Minami M. Fractional anisotropy values of calf muscles in normative state after exercise: preliminary results. Magn Reson Med Sci. 2008;7(3):157-162.
6. Schlaffke L, Rehmann R, Froeling M, et al. Diffusion tensor imaging of the human calf: variation of inter- and intramuscle-specific diffusion parameters. J Magn Reson Imaging. 2017;46(4):1137-1148.
7. Bhatt JM. The epidemiology of neuromuscular diseases. Neurol Clin. 2016;34(4):999-1021.
8. Pons C, Borotikar B, Garetier M, et al. Quantifying skeletal muscle volume and shape in humans using MRI: a systematic review of validity and reliability. PloS ONE. 2018;13(11):e0207847.
9. Südhoff I, de Guise JA, Nordez A, et al. 3D-patient-specific geometry of the muscles involved in knee motion from selected MRI images. Med Biol Eng Comput. 2009;47(6):579-587.
10. Moal B, Raya JG, Jolivet E, et al. Validation of 3D spino-pelvic muscle reconstructions based on dedicated MRI sequences for fat-water quantification. IRBM. 2014;35(3):119-127.
11. Monte JR, Hooijmans MT, Froeling M, et al. The repeatability of bilateral diffusion tensor imaging (DTI) in the upper leg muscles of healthy adults. Eur Radiol. 2020;30(3):1709-1718.
12. Rehmann R, Schlaffke L, Froeling M, et al. Muscle diffusion tensor imaging in glycogen storage disease V (McArdle disease). Eur Radiol. 2019;29(6):3224-3232.
13. Forsting J, Rehmann R, Froeling M, Vorgerd M, Tegenthoff M, Schlaffke L. Diffusion tensor imaging of the human thigh: consideration of DTI-based fiber tracking stop criteria. Magn Reson Mater Phys Biol Med. 2019;33(3):343-355.
14. Damon BM, Froeling M, Buck AKW, et al. Skeletal muscle diffusion tensor-MRI fiber tracking: rationale, data acquisition and analysis methods. NMR Biomed. 2017;30(3):e3563.
15. Oudemans J, Mazzoli V, Marra MA, et al. A novel diffusion-tensor MRI approach for skeletal muscle fascicle length measurements. Physiol Rep. 2016;4(24):e13012.
16. Bolsterlee B, D’Souza A, Herbert RD. Reliability and robustness of muscle architecture measurements obtained using diffusion tensor imaging with anatomically constrained tractography. J Biomech. 2019;86:71-78.
17. Malis V, Sinha U, Caspo R, Narici M, Smitaman E, Sinha S. Diffusion tensor imaging and diffusion modeling: application to monitoring changes in the medial gastrocnemius in disuse atrophy induced by unilateral limb suspension. J Magn Reson Imaging. 2019;49(6):1655-1664.
18. Froeling M, Oudemans J, Strijkers GJ, et al. Muscle changes detected with diffusion-tensor imaging after long-distance running. Radiology. 2015;274(2):548-562.
19. Ponrartana S, Andrade KE, Wren TAL, et al. Repeatability of chemical-shift-encoded water-fat MRI and diffusion-tensor imaging in lower extremity muscles in children. Am J Roentgenol. 2014;202(6):W567-W573.

20. Kalin PS, Huber FA, Hamie QM, et al. Quantitative MRI of visually intact rotator cuff muscles by multi-echo Dixon-based fat quantification and diffusion tensor imaging. J Magn Reson Imaging. 2019;49(1):109-117.

21. Wickiewitz Roy RR, Powell PL, Edgerton VR. Architecture of the human lower limb. Clin Orthop Relat Res. 1983(179):275-283.

22. Lieber RLJ. Clinical significance of skeletal muscle architecture. Clin Orthop Relat Res. 2001;383:140-151.

23. Lee D, Li Z, Sohail QZ, Jackson K, Fiume E, Agur A. A three-dimensional approach to pennation angle estimation for human skeletal muscle. Comput Methods Biomech Biomed Eng. 2015;18(13):1474-1484.

24. Charles JP, Moon C-H, Anderst W. Determining subject-specific lower-limb muscle architecture data for musculoskeletal models using diffusion tensor MRI. J Biomech Eng. 2019;141(6):060905.

25. Huoijmans MT, Monte JRC, Froeling M, et al. Quantitative MRI reveals microstructural changes in the upper leg muscles after running a marathon. J Magn Reson Imaging. 2020;52(2):407-417.

26. Froeling M, Tax CMW, Vos SB, Luijten PR, Leemans A. “MASSIVE” brain dataset: multiple acquisitions for standardization of structural imaging validation and evaluation. Magn Reson Med. 2017;77(5):1797-1809.

27. Schlafkne L, Rehmann R, Rohm M, et al. Multi-center evaluation of stability and reproducibility of quantitative MRI measures in healthy calf muscles. NMR Biomed. 2019;32(9):e4119.

28. Froeling M. QMRTools: a Mathematica toolbox for quantitative MRI analysis. J Open Source Softw. 2019;4(38):1-8, 1204.

29. Tax CMW, Otte WM, Viergever MA, Dijkhuizen RM, Leemans A. REKINDLE: Robust Extraction of Kurtosis INDices with Linear Estimation. Magn Reson Med. 2015;73(2):794-808.

30. Leemans A, Jeuruissen B, Sijbers J, Jones DK. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. Proc Int Soc Magn Reson Med. 2009;17:3537.

31. Klein S, Starling M, Pluim JWP. Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. IEEE Trans Image Process. 2007;16(12):2879-2890.

32. Basser PJ, Pajevic S, Pierpaoli C, Duda J, Aldroubi A. In vivo fiber tractography using DT-MRI data. Magn Reson Med. 2000;44(4):625-632.

33. Richardson JTE. Eta squared and partial eta squared as measures of effect size in educational research. Educ Res Rev. 2011;6(2):135-147.

34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2010;9(2):157-164.

35. Fieremans E, Lemberskiy G, Veraart J, Sigmund EE, Gyftopoulos S, Novikov DS. In vivo measurement of membrane permeability and myofiber size in human muscle using time-dependent diffusion tensor imaging and the random permeable barrier model. NMR Biomed. 2017;30(3):e3612.

36. Froeling M, Nederveen AJ, Nicolay K. DTI of human skeletal muscle: the effects of diffusion encoding parameters, signal-to-noise ratio and T2 on tensor indices and fiber tracts. NMR Biomed. 2013;26(11):1339-1352.

37. Heemskerk AM, Sinha TK, Wilson KJ, Ding Z, Damon BM. Quantitative assessment of DTI-based muscle fiber tracking and optimal tracking parameters. Magn Reson Med. 2009;61(2):467-472.

38. Oudeman J, Nederveen AJ, Strijkers GJ, Maas M, Luijten PR, Froeling M. Techniques and applications of skeletal muscle diffusion tensor imaging: a review. J Magn Reson Imaging. 2016;43(4):773-788.

39. Bolsterlee B, Finni T, D’Souza A, Eguchi J, Clarke EC, Herbert RD. Three-dimensional architecture of the whole human soleus muscle in vivo. PeerJ. 2018;6:e4610.

40. Sahrmann AS, Stott NS, Besier TF, Fernandez JW. Handsfield GG. Soleus muscle weakness in cerebral palsy: muscle architecture revealed with diffusion tensor imaging. PLoS ONE. 2019;14(2):e0205944.

41. Orglu S, Lafortuna CL, Rastelli F, Cadioli M, Falini A, Rizzo G. Automatic muscle and fat segmentation in the thigh from T1-weighted MRI. J Magn Reson Imaging. 2016;43(3):601-610.

42. Froeling M, Schlafkne L, Rohm M, Isgum I, Kan H, Wolterink J. Evaluation of input data and UNet based convolutional network architectures for automated muscle annotation in 2D and 3D. Paper presented at: ISMRM 27th Annual Meeting and Exhibition; May 11-16, 2019; Montreal, Quebec, Canada.

43. 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(2):231-242.

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Forsting J, Rehmann R, Rohm M, Froeling M, Schlaffke L. Evaluation of interrater reliability of different muscle segmentation techniques in diffusion tensor imaging. NMR in Biomedicine. 2021;34:e4430. https://doi.org/10.1002/nbm.4430