



Inter-software and inter-threshold reliability of quantitative paraspinal muscle segmentation

Sihai Liu¹ · Hendrik Schmidt¹ · Katharina Ziegeler² · Tianwei Zhang¹ · Daishui Yang¹ · Nima Taheri^{1,3} · Matthias Pumberger³ · Luis Becker³

Received: 31 March 2023 / Revised: 17 October 2023 / Accepted: 12 November 2023 / Published online: 6 December 2023
© The Author(s) 2023

Abstract

Purpose Changes in the cross-sectional area (CSA) and functional cross-sectional area (FCSA) of the lumbar multifidus (MF) and erector spinae muscles (ES) are factors that can contribute to low back pain. For the assessment of muscle CSA and composition there are various software and threshold methods used for tissue segmentation in quantitative analysis. However, there is currently no gold standard for software as well as muscle segmentation. This study aims to analyze the measurement error between different image processing software and different threshold methods for muscle segmentation.

Methods Magnetic resonance images (MRI) of 60 patients were evaluated. Muscle CSA and FCSA measurements were acquired from axial T2-weighted MRI of the MF and ES at L4/L5 and L5/S1. CSA, FCSA, and FCSA/CSA ratio were measured independently by two observers. The MRI images were measured using two different software programs (ImageJ and Amira) and with two threshold methods (Circle/Overlap method) for each software to evaluate FCSA and FCSA/CSA ratio.

Results Inter-software comparisons revealed high inter-rater reliability. However, poor inter-rater reliability were obtained with different threshold methods. CSA, FCSA, and FCSA/CSA showed excellent inter-software agreement of 0.75–0.99 regardless of the threshold segmentation method. The inter-rater reliability between the two observers ranged between 0.75 and 0.99. Comparison of the two segmentation methods revealed agreement between 0.19 and 0.84. FCSA and FCSA/CSA measured via the Overlap method were significantly higher than those measured via the Circle method ($P < 0.01$).

Conclusion The present study showed a high degree of reliability with very good agreement between the two software programs. However, study results based on different threshold methods should not be directly compared.

Keywords Paraspinal muscle · Fatty infiltration · Magnetic resonance imaging · Software · Threshold

✉ Luis Becker
luis-alexander.becker@charite.de

Sihai Liu
Sihai.Liu@Charite.de

Hendrik Schmidt
Hendrik.Schmidt@Charite.de

Katharina Ziegeler
Katharina.Ziegeler@Charite.de

Tianwei Zhang
Tianwei.Zhang@Charite.de

Daishui Yang
Daishui.Yang@Charite.de

Nima Taheri
Nima.Taheri@Charite.de

Matthias Pumberger
Matthias.Pumberger@Charite.de

¹ Berlin Institute of Health, Julius Wolff Institute for Biomechanics and Musculoskeletal Regeneration, Charité – Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

² Department for Radiology, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

³ Center for Musculoskeletal Surgery, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

Introduction

Up to 80% of adults experienced chronic non-specific low back pain (LBP) at some point in their lives [1]. Global Burden of Disease's research shows, LBP is the main reason as the primary cause of disability [2]. Despite considerable research on the etiology of LBP, the pathomorphological relationships are not yet fully understood. One factor that might affect LBP is the composition and morphology of the paraspinal muscles [3, 4].

Currently, quantitative and qualitative methods are being used to assess the composition of paraspinal muscle tissue. Qualitative assessment refers to the utilization of visual grading methods to assess the extent of fatty infiltration. The reliability of Goutallier classification system (0–4 grading scale) measurements of the degree of fatty infiltration in the lumbar region has been questioned previously [5]. Therefore, the quantitative measurements such as the assessment of the cross-sectional area (CSA) and the functional cross-sectional area (FCSA), which results from the exclusion of the fat compounds of the muscles cross-sectional area have gained growing interest [6–10]. A quantitative assessment of the paraspinal muscle composition using MRI is performed by segregating the pixels within the region of interest that is thought to represent fat. These techniques could be performed using a manual segmentation method or different threshold methods [11, 12].

Although it is assumed that the measurement error is mainly related to the observer and the method used, another problem might be the use of different software, which can lead to incomparable results [13]. Therefore, it is important to verify whether direct comparisons can be made between the various freeware or commercial software packages used for this procedure. Whereas Fortin et al. [14] reported an excellent agreement between ImageJ and OsiriX in the assessment of paraspinal muscle CSA, composition, and side-to-side asymmetry, the amount of measurement error in between the U.S. Food and Drug Administration certified software package Amira (version 2019.4, Thermo Fisher Scientific Inc. Waltham, USA) and the freeware ImageJ (version 1.53, National Institutes of Health, Bethesda, Maryland, USA) is unknown. Therefore, the objective of this study was to elucidate the influence of inter-software differences between Amira and ImageJ as well as the influence of segmentation techniques.

Materials and method

Study design

In this retrospective study, we randomly selected 60 MRIs (39 women and 21 men) of the lumbar region from a sample

of a large cohort study, which was approved by the local ethics committee (EA1/058/21). MRI scans were conducted using a Siemens Avanto 1.5 T MRI system (Siemens AG, Erlangen, Germany) with T_2 -weighted turbo spin echo sequences for both axial and sagittal images. The axial T_2 parameters used were a repetition time of 4.000, an echo time of 113, and a slice thickness of 3 mm. As the vast majority of degenerative changes can be detected in the lower spine, the levels L4–L5 and L5–S1 were evaluated.

Muscle measurements and segmentation

All measurements were performed by two orthopedic residents, who were trained in the MRI muscle assessment. The MRI images were measured through the two different image processing programs (ImageJ and Amira). The two observers measured the MRIs in a random order for both investigators. The CSA of the multifidus muscle (MF) and erector spinae muscle (ES) was measured at mid-disk level L4/5 and L5/S1 (Fig. 1), the CSA was single measured before applying any thresholds. FCSA and FCSA/CSA were determined using two different segmentation thresholds for differentiating muscle fibers and fatty muscle infiltration.

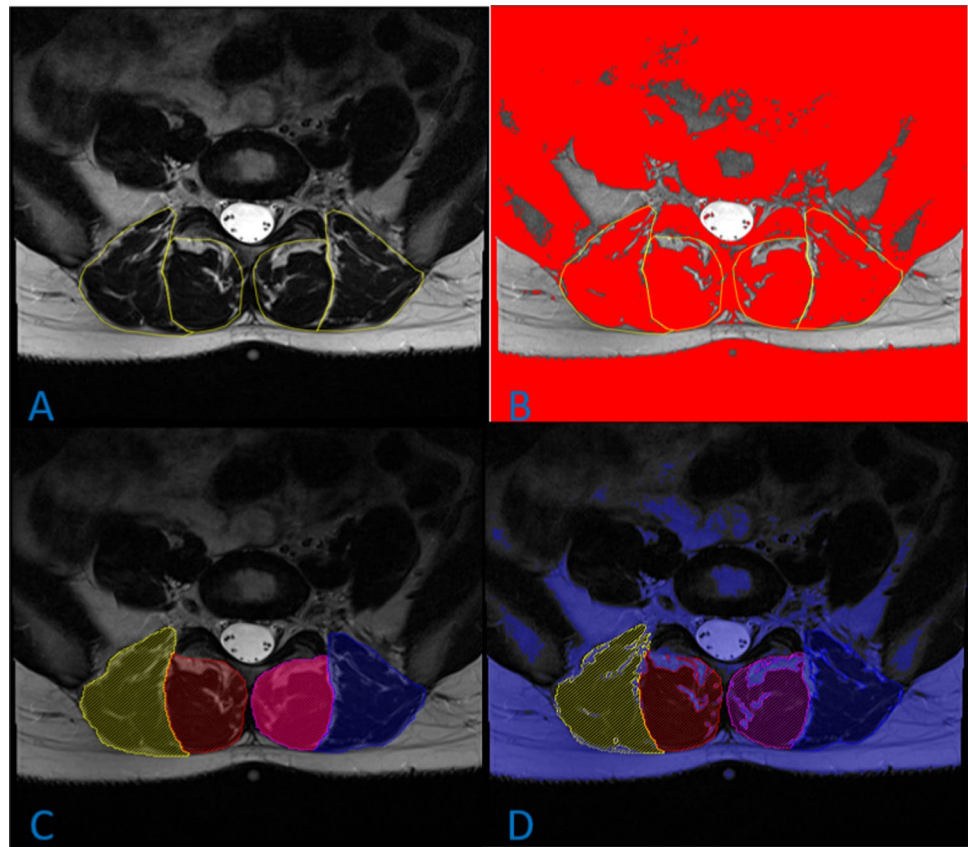
Circle method: Six regions of interest (ROIs) from the muscles of the MF and ES were taken from the visible areas of muscle tissue with least visual fatty infiltration. The maximum value that can be obtained from a sample ROI is regarded as the upper threshold to distinguish between muscle tissue and fat. Since the lower limit is usually 0 or close to 0, uniformly setting the lower limit at 0 is used to minimize errors (defined as Circle method) [14].

Overlap method: Outline CSA of paraspinal muscle (include ES and MF) and subcutaneous fat (SC) on both sides. By presenting the grayscale ranges for both CSA and SC as histograms and overlaying them, it was possible to identify signal intensities that were common to both images. The Overlapping area of the histograms represents the intensity of the fatty signal in the CSA (defined as Overlap method) [15].

Data analysis

For each measurement, descriptive statistics such as means (\bar{x}) and standard deviations (SD) were calculated. The inter-rater, inter-software, and inter-threshold reliability of the measurement were evaluated using intra-class correlation coefficient (ICC). Agreement was defined according to Portney and Watkins [16]: an ICC of 0.00–0.49 is considered poor, 0.50–0.74 is moderate, and 0.75–1.0 is excellent. As Bland and Altman suggested [17, 18], the 95% limits of agreement were used to evaluate the agreement between the measurements acquired by different raters using different

Fig. 1 L5/S1 MRI of the same subject, **A** and **B** are processed by ImageJ, **C** and **D** by Amira



software with different thresholds. The standard error of measurement (SEM) is a statistical metric used to estimate the expected error associated with a specific measurement ($SEM = S\sqrt{1 - r_{xx}}$), where S is the standard deviation of the test and r_{xx} represents the reliability of the test. In this study, the results were analyzed based on the muscles and spinal level that were investigated. The Wilcoxon Rank Sum Test is employed to analyze systematic differences between different thresholds. The statistical analysis was conducted using Statistical Package for the Social Sciences version 23.0 (SPSS Inc, Chicago, Illinois).

Based on Cohen's suggestions [19], By utilizing G^* Power version 3.1.3 (University of Düsseldorf, Düsseldorf, Germany), effect size conventions were provided in categories of "small," "medium," and "large" to determine the required sample size. In this study, with an effect size of 0.3, alpha error of 0.05, and a power (verification) of 0.8, the minimum sample size of 46 participants was determined. Therefore, the enrollment of 60 patients was considered adequate to achieve the desired statistical power.

Results

Inter-software reliability of muscle measurements using ImageJ and Amira

The outcomes of inter-software reliability (ICC), SEM values, and descriptive statistics (mean SD) are presented in Table 1. All ICC of CSA, FCSA, and FCSA/CSA of all the muscle composition measurements, regardless of the threshold methods, analyzed muscle or spinal level, showed excellent agreement, and varied between 0.75 and 0.99. SEM also showed good comparability for different software, muscle measurements, muscle analyzed, and spinal segments. In Figs. 2 and 3, Bland–Altman Plots illustrate the agreement between Amira and ImageJ and the solid line consistently crosses the y-axis above zero, indicating a systematic trend where the mean values of FCSA were consistently higher when utilizing the Amira.

Table 1 Inter-software reliability indexes between Amira and imageJ for the right MF and ES muscles

Parameter		Circle method				Overlap method			
		X_1 (SD)	X_2 (SD)	ICC (95% CI)	SEM	X_1 (SD)	X_2 (SD)	ICC (95% CI)	SEM
<i>MF muscle</i>									
CSA	L4/5	9.56 (1.75)	9.54 (1.73)	0.99 (0.98–0.99)	0.17	9.56 (1.75)	9.54 (1.73)	0.99 (0.98–0.99)	0.17
FCSA		7.58 (1.99)	7.36 (1.90)	0.98 (0.93–0.99)	0.27	8.75 (1.82)	8.44 (1.72)	0.96 (0.83–0.99)	0.35
FCSA/CSA		0.79 (0.13)	0.77 (0.12)	0.93 (0.84–0.97)	0.03	0.91 (0.07)	0.88 (0.06)	0.81 (0.30–0.93)	0.03
CSA	L5/S1	11.44 (1.86)	11.36 (1.86)	0.98 (0.96–0.99)	0.26	11.44 (1.86)	11.36 (1.86)	0.98 (0.96–0.99)	0.26
FCSA		8.74 (2.15)	8.49 (2.08)	0.98 (0.93–0.99)	0.30	10.35 (1.76)	10.04 (1.74)	0.97 (0.83–0.99)	0.30
FCSA/CSA		0.76 (0.12)	0.74 (0.12)	0.95 (0.90–0.98)	0.03	0.90 (0.05)	0.88 (0.05)	0.75 (0.40–0.88)	0.03
<i>ES muscle</i>									
CSA	L4/5	15.91 (3.32)	15.84 (3.38)	0.99 (0.99–1.00)	0.33	15.91 (3.32)	15.84 (3.38)	0.99 (0.99–1.00)	0.33
FCSA		12.92 (3.62)	12.65 (3.56)	0.99 (0.97–1.00)	0.36	14.7 (3.58)	14.28 (3.57)	0.98 (0.93–0.99)	0.50
FCSA/CSA		0.81 (0.11)	0.79 (0.10)	0.95 (0.91–0.97)	0.02	0.92 (0.08)	0.90 (0.08)	0.84 (0.67–0.91)	0.28
CSA	L5/S1	10.98 (3.10)	10.85 (3.03)	0.99 (0.99–1.00)	0.31	10.98 (3.10)	10.85 (3.03)	0.99 (0.99–1.00)	0.31
FCSA		7.75 (3.25)	7.63 (3.18)	0.99 (0.99–1.00)	0.32	9.87 (3.00)	9.50 (2.98)	0.98 (0.90–0.99)	0.42
FCSA/CSA		0.69 (0.16)	0.69 (0.15)	0.99 (0.97–0.99)	0.02	0.90 (0.07)	0.87 (0.06)	0.79 (0.47–0.90)	0.03

X_1 mean value in Amira, X_2 = mean value in ImageJ, ICC intra-class correlation coefficient, CI confidence interval, SEM standard error of measurement, CSA cross-sectional area, FCSA functional CSA, FCSA/CSA ratio, the unit of area is cm^2

Inter-rater reliability of muscle measurements using imageJ and Amira with Overlap method and Circle method

The results of inter-rater reliability (ICC), SEM values and descriptive statistics (mean \pm SD) data of L4–L5 and L5–S1 MF and ES between different software are given in Table 2. When measured using Amira, the ICC ranged from 0.75 to 0.99 for the inter-rater reliability of the Overlap method and 0.89 to 0.98 for the Circle method for both spinal levels. Inter-rater reliability of the Overlap method for both spinal levels ranged from 0.75 to 0.99 when measured using ImageJ, and ICC ranged from 0.88 to 0.98 for the Circle method. There were no significant differences observed in the ICC ranges for the inter-rater reliability analysis using two software and two threshold methods. However, compared with the Overlap method, the ICC of the Circle method is higher, and the SEM value is also slightly higher.

Inter-threshold reliability of muscle measurements using the Circle method and Overlap method

The ICC of MF and ES composition between the two different threshold methods showed poor or moderate agreement in both software programs (Table 3). Accordingly, the SEM value of the ES muscle and MF muscle in each software was high. All FCSA and FCSA/CSA measured using the Overlap method exhibits a notably greater extent compared to measurements taken with the Circle method, and this disparity demonstrates statistical significance ($P < 0.01$).

Discussion

During the quantitative assessment of lumbar paraspinal muscles composition, we compared the differences of segmentation of paraspinal muscles (MF and ES) by using different thresholding methods (Circle and Overlap) with

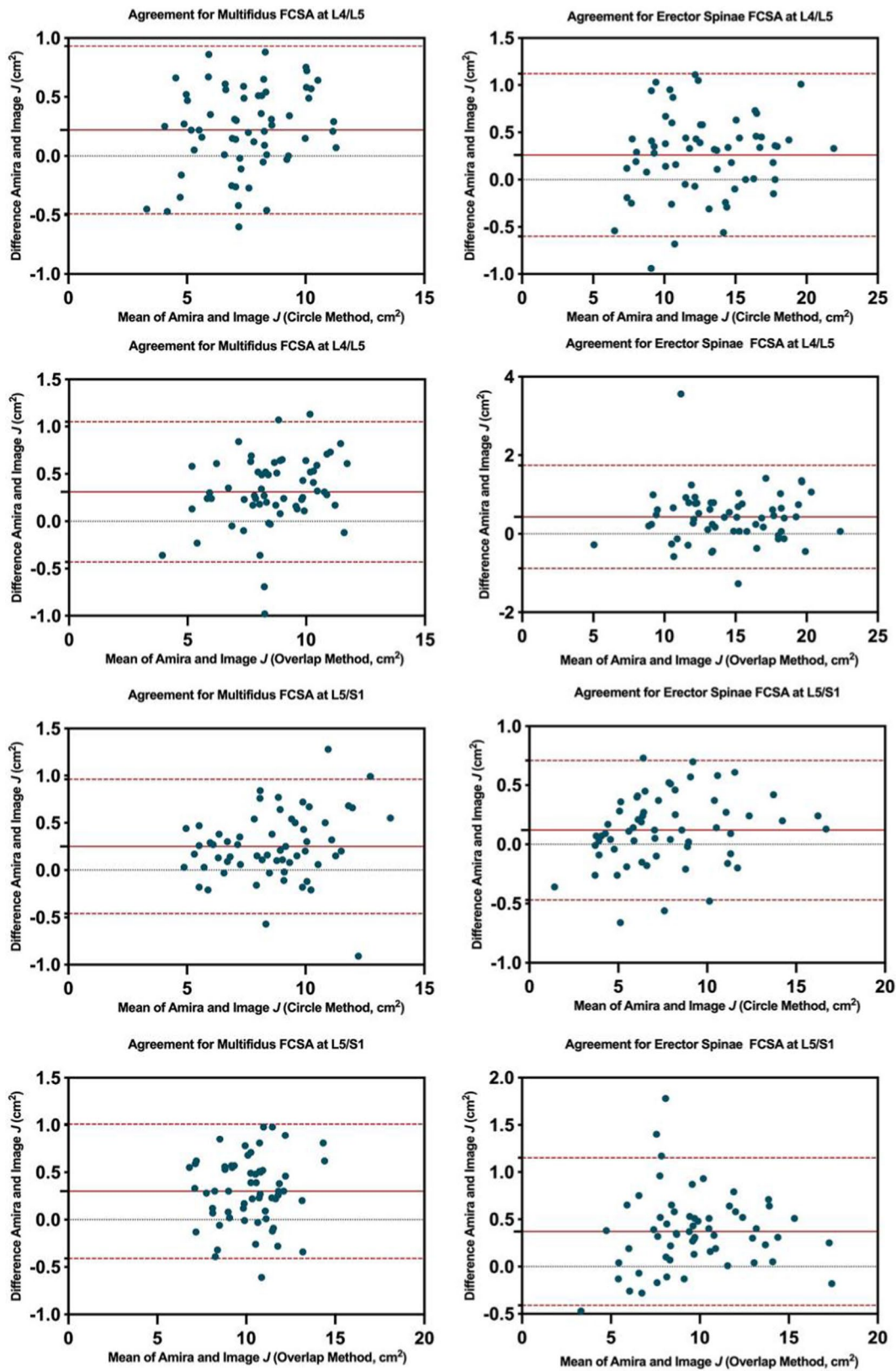


Fig. 2 Bland–Altman 95% limits of agreement plots for the FCSA measurements of the right MF and ES at L4–L5 and L5–S1. The solid line represents the mean difference between the two measure-

ment methods (i.e., Amira value-imageJ value), the dotted line represents the 95% limits of agreement for the difference (defined as the mean \pm 1.96 SD)

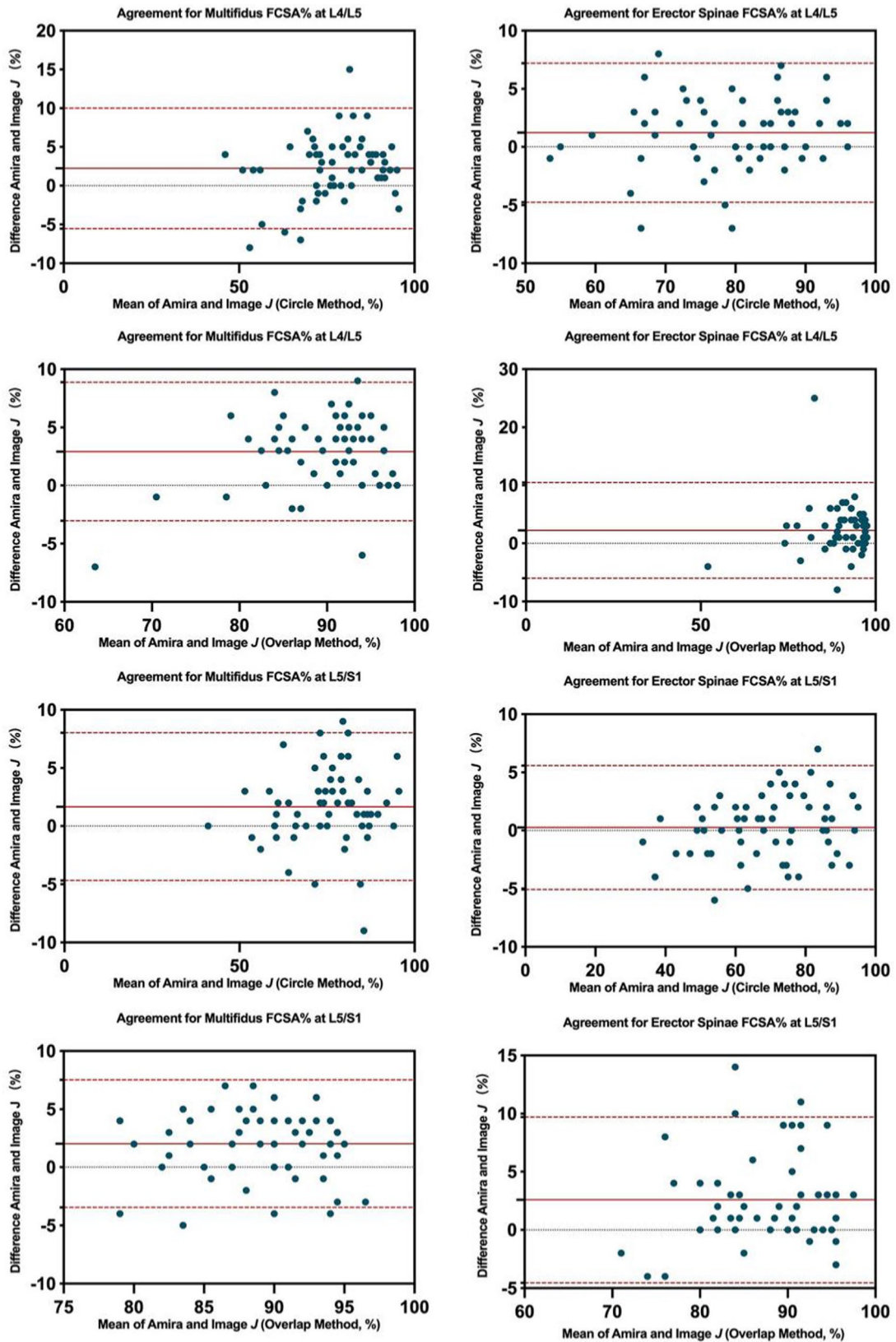


Fig. 3 Bland–Altman 95% limits of agreement plots for the FCSA% (100*FCSA/CSA) measurements of the right MF and ES at L4/L5 and L5/S1. The solid line represents the mean difference between

the two measurement methods (i.e., Amira value–ImageJ value), the dotted line represents the 95% limits of agreement for the difference (defined as the mean \pm 1.96 SD)

Table 2 Inter-rater reliability for different software and threshold segmentation

Parameter		Amira			ImageJ		
		X (SD)	ICC (95% CI)	SEM	X (SD)	ICC (95% CI)	SEM
<i>MF muscle</i>							
CSA	L4/5	9.56 (1.75)	0.98 (0.97–0.99)	0.25	9.54 (1.74)	0.96 (0.91–0.98)	0.35
FCSA circle		7.58 (2.00)	0.97 (0.95–0.98)	0.35	7.36 (1.92)	0.95 (0.90–0.97)	0.43
FCSA overlap		8.75 (1.83)	0.96 (0.92–0.98)	0.37	8.44 (1.73)	0.97 (0.94–0.98)	0.30
FCSA/CSA circle		0.79 (0.13)	0.93 (0.89–0.96)	0.03	0.77 (0.12)	0.88 (0.81–0.93)	0.04
FCSA/CSA overlap		0.91 (0.07)	0.83 (0.69–0.91)	0.03	0.88 (0.06)	0.83 (0.72–0.89)	0.02
CSA	L5/S1	11.44 (1.87)	0.97 (0.94–0.98)	0.32	11.36 (1.88)	0.96 (0.93–0.98)	0.38
FCSA circle		8.73 (2.16)	0.96 (0.94–0.98)	0.43	8.48 (2.09)	0.97 (0.93–0.98)	0.36
FCSA overlap		10.34 (1.77)	0.96 (0.94–0.98)	0.35	10.04 (1.75)	0.97 (0.96–0.98)	0.3
FCSA/CSA circle		0.76 (0.12)	0.89 (0.81–0.94)	0.04	0.74 (0.12)	0.91 (0.85–0.94)	0.04
FCSA/CSA overlap		0.90 (0.05)	0.75 (0.62–0.85)	0.03	0.88 (0.05)	0.77 (0.59–0.87)	0.02
<i>ES muscle</i>							
CSA	L4/5	15.91 (3.31)	0.99 (0.98–0.99)	0.33	15.83 (3.37)	0.99 (0.99–1.00)	0.34
FCSA circle		12.92 (3.62)	0.98 (0.97–0.99)	0.51	12.65 (3.57)	0.98 (0.96–0.99)	0.50
FCSA overlap		14.7 (3.58)	0.98 (0.97–0.99)	0.51	14.27 (3.57)	0.98 (0.97–0.99)	0.50
FCSA/CSA circle		0.81 (0.11)	0.91 (0.86–0.95)	0.03	0.79 (0.10)	0.89 (0.82–0.93)	0.03
FCSA/CSA overlap		0.92 (0.08)	0.92 (0.87–0.95)	0.02	0.90 (0.08)	0.89 (0.83–0.93)	0.03
CSA	L5/S1	10.98 (3.10)	0.99 (0.98–0.99)	0.31	10.85 (3.04)	0.97 (0.95–0.98)	0.53
FCSA circle		7.75 (3.25)	0.98 (0.97–0.99)	0.46	7.62 (3.18)	0.98 (0.97–0.99)	0.45
FCSA overlap		9.87 (3.00)	0.99 (0.98–0.99)	0.3	9.50 (2.97)	0.99 (0.98–0.99)	0.3
FCSA/CSA circle		0.69 (0.16)	0.91 (0.82–0.95)	0.05	0.69 (0.15)	0.91 (0.86–0.95)	0.05
FCSA/CSA overlap		0.90 (0.07)	0.83 (0.72–0.89)	0.03	0.87 (0.07)	0.75 (0.61–0.84)	0.04

X mean value, ICC intraclass correlation coefficient, CI confidence interval, SEM standard error of measurement, CSA cross-sectional area, FCSA functional CSA, FCSA/CSA ratio, the unit of area is cm²

different software (ImageJ and Amira). The agreement of the relevant paraspinal muscle measurements between these two distinct image processing programs demonstrated excellent reliability. These findings are supported by the Bland and Altman limit of agreement, which indicate that the agreement between the two software programs

is acceptable, and they can be used interchangeably. In addition, similar inter-rater and inter-software reliability coefficients and SEM indicated that the software used contributed little to measurement error. Supporting the results of prior research we found an excellent inter-rater reliability in CSA and FCSA measurements [20, 21]. However,

Table 3 Inter-threshold reliability indexes between Circle and Overlap method for the right MF and ES muscles

Parameter	Amira				ImageJ				
	Circle X (SD)	Overlap X (SD)	ICC (95% CI)	SEM	Circle X (SD)	Overlap X (SD)	ICC (95% CI)	SEM	
<i>MF muscle</i>									
FCSA	L4/5	7.58 (1.99)	8.75 (1.82)	0.70 (0.10–0.88)	1.09	7.36 (1.90)	8.44 (1.72)	0.63 (0.07–0.87)	1.25
FCSA/CSA		0.79 (0.13)	0.91 (0.07)	0.23 (0.08–0.50)	0.11	0.77 (0.12)	0.88 (0.06)	0.18 (0.09–0.46)	0.1
FCSA	L5/S1	8.74 (2.15)	10.35 (1.76)	0.63 (0.07–0.86)	1.29	9.49 (2.08)	10.04 (1.74)	0.71 (0.11–0.89)	1.02
FCSA/CSA		0.76 (0.12)	0.90 (0.05)	0.18 (0.09–0.45)	0.11	0.74 (0.12)	0.88 (0.05)	0.22 (0.08–0.49)	0.1
<i>ES muscle</i>									
FCSA	L4/5	12.92 (3.62)	14.70 (3.58)	0.63 (0.07–0.86)	1.29	12.65 (3.56)	14.28 (3.57)	0.85 (0.15–0.95)	1.41
FCSA/CSA		0.81 (0.11)	0.92 (0.08)	0.18 (0.09–0.45)	0.11	0.79 (0.10)	0.90 (0.08)	0.34 (-0.08–0.63)	0.08
FCSA	L5/S1	7.75 (3.25)	9.87 (3.00)	0.74 (0.06–0.92)	1.68	7.63 (3.18)	9.50 (2.98)	0.78 (0.04–0.94)	1.51
FCSA/CSA		0.69 (0.16)	0.90 (0.07)	0.18 (0.09–0.47)	0.14	0.69 (0.15)	0.87 (0.06)	0.24 (0.10–0.55)	0.13

X mean value, ICC intra-class correlation coefficient, CI confidence interval, SEM standard error of measurement, CSA cross-sectional area, FCSA functional CSA, FCSA/CSA ratio, the unit of area is cm²

relevant differences between the two threshold methods were observed. The agreement of related paraspinal muscles between these two methods is low or moderate.

Both threshold methods have been used in prior work and used in comparisons, although the agreement and reliability between the two threshold methods could not be confirmed by this study [14, 15]. However, the differences might result from the two different ways to determine the upper threshold and therefore lead to different results. However, research about reliability and agreement of inter-threshold comparisons is very rare. Fortin et al. [22] compared an automated thresholding algorithm with the Circle method, for which excellent agreement between 0.79 and 0.99 was reported. Besides that study, to the knowledge of the authors there is paucity in inter-threshold comparisons.

As the thresholding is crucial for muscle segmentation there are some studies proposing different manual, semi-automated, or automated approaches. Otsu et al. [23] presented a method to select a threshold automatically from a gray-level histogram. Cooley et al. [24] acquired an initial histogram for each image by first outlining both MF (connected via the subcutaneous fat but excluding any vertebral structures). Ranson et al. [25] proposed tissue differentiation based on manual segmentation of the three tissue types within the MRI vertebral bone, paraspinal muscles, and intermuscular fat. The resulting grayscale values for the three tissue types were then normalized to

the total number of pixels analyzed to determine the gray-scale range of MR signal intensities for the three tissue types across the scan set. Although these articles suggest methods that could be supposedly effective, they do not provide a gold standard for assessing the infiltration of fat in the paraspinal muscles.

The paper's main limitation is that only two image processing software for the quantitative assessment of paraspinal muscle composition was compared even if there exists a wide variety of different software approaches. Besides the two compared manual threshold methods there exists further manual, semi-automated or automated approaches which could not be compared within this paper.

Conclusion

In conclusion, the presented method to study paraspinal muscle CSA and composition has a high degree of reliability with very good agreement between the two software programs. However, the comparison between the two different thresholding approaches presented mostly moderate or poor reliability and therefore the results of these different thresholding methods should not be compared against each other.

Author contributions All authors contributed sufficiently to the work and take public responsibility for the following: SL, LB, and

HS contributed to the conception and design, SL, TZ, and DY were involved in the data acquisition, SL contributed to the analysis of data, SL and LB were involved in the drafting of the manuscript, SL, HS, KZ, TZ, DY, NT, MP, and LB contributed to the critical revision, MP and HS were involved in obtaining funding, MP and HS assisted in the administrative support, and MP, HS, and LB were involved in the supervision.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was financially supported by the German Research Foundation (DFG): PU 762/1-1, SCHM 2572/11-1, SCHM 2572/13-1. S.L., T.Z., D.Y. received China Scholarship Council (CSC, No. 202208080039, No. 202208080046, No. 202208080034).

Data availability All data generated or analyzed during this study are included in this published article.

Declarations

Conflict of interest All authors declare that there are no financial or personal relationships with other persons or organizations that could have inappropriately influenced this study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Traeger AC, Qaseem A, McAuley JH (2021) Low Back Pain. *JAMA* 326:286. <https://doi.org/10.1001/jama.2020.19715>
- Hoy D, March L, Brooks P, Blyth F, Woolf A, Bain C, Williams G, Smith E, Vos T, Barendregt J, Murray C, Burstein R, Buchbinder R (2014) The global burden of low back pain: estimates from the global burden of disease 2010 study. *Ann Rheum Dis* 73:968–974. <https://doi.org/10.1136/annrheumdis-2013-204428>
- Seyedhoseinpoor T, Taghipour M, Dadgoo M, Ebrahimi Takamjani I, Sanjari MA, Kazemnejad A, Hides J (2022) Relationship between the morphology and composition of the lumbar paraspinal and psoas muscles and lumbar intervertebral motion in chronic low-back pain: an exploratory study. *Clin Anat* 35:762–772. <https://doi.org/10.1002/ca.23893>
- Seyedhoseinpoor T, Taghipour M, Dadgoo M, Sanjari MA, Takamjani IE, Kazemnejad A, Khoshamooz Y, Hides J (2022) Alteration of lumbar muscle morphology and composition in relation to low back pain: a systematic review and meta-analysis. *Spine J* 22:660–676. <https://doi.org/10.1016/j.spinee.2021.10.018>
- Battaglia PJ, Maeda Y, Welk A, Hough B, Kettner N (2014) Reliability of the Goutallier classification in quantifying muscle fatty degeneration in the lumbar multifidus using magnetic resonance imaging. *J Manip Physiol Ther* 37:190–197. <https://doi.org/10.1016/j.jmpt.2013.12.010>
- Han G, Jiang Y, Zhang B, Gong C, Li W (2021) Imaging evaluation of fat infiltration in paraspinal muscles on MRI: a systematic review with a focus on methodology. *Orthop Surg* 13:1141–1148. <https://doi.org/10.1111/os.12962>
- Khil EK, Choi JA, Hwang E, Sidek S, Choi I (2020) Paraspinal back muscles in asymptomatic volunteers: quantitative and qualitative analysis using computed tomography (CT) and magnetic resonance imaging (MRI). *BMC Musculoskelet Disord* 21:403. <https://doi.org/10.1186/s12891-020-03432-w>
- Abbasi S, Hadian Rasanani MR, Olyaei GR, Ghotbi N, Rasouli O (2021) Reliability of ultrasound measurement of the lateral abdominal and lumbar multifidus muscles in individuals with chronic low back pain: a cross-sectional test-retest study. *J Bodyw Mov Ther* 26:394–400. <https://doi.org/10.1016/j.jbmt.2020.12.029>
- Erdagi K, Poyraz N (2021) Study on the cross-sectional area of the multifidus muscle at the L4–L5 level on ultrasound images of olympic-style weightlifting and judo athletes. *J Back Musculoskelet Rehabil* 34:689–696. <https://doi.org/10.3233/BMR-200101>
- Prasetyo M, Nindita N, Murdana IN, Prihartono J, Setiawan SI (2020) Computed tomography evaluation of fat infiltration ratio of the multifidus muscle in chronic low back pain patients. *Eur J Radiol Open* 7:100293. <https://doi.org/10.1016/j.ejro.2020.100293>
- Hu ZJ, He J, Zhao FD, Fang XQ, Zhou LN, Fan SW (2011) An assessment of the intra- and inter-reliability of the lumbar paraspinal muscle parameters using CT scan and magnetic resonance imaging. *Spine (Phila Pa 1976)* 36:E868–874. <https://doi.org/10.1097/BRS.0b013e3181ef6b51>
- Ropponen A, Videman T, Battie MC (2008) The reliability of paraspinal muscles composition measurements using routine spine MRI and their association with back function. *Man Ther* 13:349–356. <https://doi.org/10.1016/j.math.2007.03.004>
- Keller A, Gunderson R, Reikeras O, Brox JI (2003) Reliability of computed tomography measurements of paraspinal muscle cross-sectional area and density in patients with chronic low back pain. *Spine (Phila Pa 1976)* 28:1455–1460. <https://doi.org/10.1097/01.BRS.0000067094.55003.AD>
- Fortin M, Battie MC (2012) Quantitative paraspinal muscle measurements: inter-software reliability and agreement using OsiriX and ImageJ. *Phys Ther* 92:853–864. <https://doi.org/10.2522/ptj.20110380>
- Jun HS, Kim JH, Ahn JH, Chang IB, Song JH, Kim TH, Park MS, Chan Kim Y, Kim SW, Oh JK, Yoon DH (2016) The effect of lumbar spinal muscle on spinal sagittal alignment: evaluating muscle quantity and quality. *Neurosurgery* 79:847–855. <https://doi.org/10.1227/NEU.0000000000001269>
- Portney LG, Watkins MP (2009) *Foundations of clinical research: applications to practice*. Pearson/Prentice Hall Upper Saddle River, New Jersey
- Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1:307–310
- Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. *Stat Methods Med Res* 8:135–160. <https://doi.org/10.1177/096228029900800204>
- Abo T, Kawamura T (1997) Immune system and diseases that vary with circumstance and physical condition. *Rinsho Byori* 45:3–12
- Schomoller A, Risch L, Kaplick H, Wochatz M, Engel T, Schraplau A, Sonnenburg D, Huppertz A, Mayer F (2021) Inter-rater and inter-session reliability of lumbar paraspinal muscle composition in a mobile MRI device. *Br J Radiol* 94:20210141. <https://doi.org/10.1259/bjr.20210141>
- Suzuki K, Hasebe Y, Yamamoto M, Saita K, Ogihara S (2021) Inter-rater reliability between two examiners with different professional roles in the evaluation of fat infiltration in the lumbar

- paraspinal muscles using magnetic resonance imaging. *J Phys Ther Sci* 33:591–595. <https://doi.org/10.1589/jpts.33.591>
22. Fortin M, Omidyeganeh M, Battie MC, Ahmad O, Rivaz H (2017) Evaluation of an automated thresholding algorithm for the quantification of paraspinal muscle composition from MRI images. *Biomed Eng Online* 16:61. <https://doi.org/10.1186/s12938-017-0350-y>
 23. Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Trans Syst* 9:5. <https://doi.org/10.1109/tsmc.1979.4310076>
 24. Cooley JR, Hebert JJ, de Zoete A, Jensen TS, Algra PR, Kjaer P, Walker BF (2021) Assessing lumbar paraspinal muscle cross-sectional area and fat composition with $T1$ versus $T2$ -weighted magnetic resonance imaging: reliability and concurrent validity. *PLoS ONE* 16:e0244633. <https://doi.org/10.1371/journal.pone.0244633>
 25. Ranson CA, Burnett AF, Kerslake R, Batt ME, O'Sullivan PB (2006) An investigation into the use of MR imaging to determine the functional cross sectional area of lumbar paraspinal muscles. *Eur Spine J* 15:764–773. <https://doi.org/10.1007/s00586-005-0909-3>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.