Cite as: Sobolewski EJ, Wein LD, Crow JM, Carpenter KM: Intra-rater and inter-rater reliability of the process of obtaining cross-sectional area and echo intensity measurements of muscles from ultrasound images. J Ultrason 2021; 21: e7–e11. doi: 10.15557/JoU.2021.0002.

Submitted: 22.09.2020
Accepted: 16.01.2021
Published: 08.03.2021

Intra-rater and inter-rater reliability of the process of obtaining cross-sectional area and echo intensity measurements of muscles from ultrasound images

Eric J. Sobolewski, Leah D. Wein, Jacquelyn M. Crow, Kaitlyn M. Carpenter

Molnar Human Performance Lab, Furman University, Greenville SC, USA

Correspondence: Assistant Prof. Dr. Eric Sobolewski, Furman University 3300 Poinsett Hwy, Greenville, SC, USA 29613; tel.: +18642943602, fax: +18642942942, e-mail: eric.sobolewski@furman.edu

DOI: 10.15557/JoU.2021.0002

Keywords
muscle quality, muscle size, ultrasonography, echogenicity

Abstract

Introduction: The use of ultrasound images for analyzing muscle quality and size is continuing to grow in the literature. However, many of these manuscripts fail to properly describe their measurement techniques and steps involved in analyzing ultrasound images. **Aim of this study:** To evaluate the intra- and inter-rater reliability of the steps involved when analyzing ultrasound images to measure cross-sectional area and echo intensity. **Material and methods:** Twenty ultrasound images of the rectus femoris and vastus lateralis images were blinded and replicated, and then analyzed by experienced raters. The raters then were asked to analyze the images using open-source software for scaling measurements, subcutaneous fat thickness, cross-sectional area, and echo intensity. Matched image values for each measurement where compared for intra- and inter-rater reliability. **Results:** Intra-rater reliability ranged from fair ($ICC_{3,1} = 0.32$) to high (0.98), with echo intensity values being the least reliable (>0.55), and scaling and depth measurements being the most reliable (<0.85). Inter-rater reliability ranged from good (0.77) to high (0.97). Conclusion: Ultrasound- derived measures of cross-sectional area and echo intensity can be measured reliably, with echo intensity being the most difficult to replicate. However, reliability measures are unique to the rater and study and, therefore, should be clearly reported in every paper.

Introduction

More and more studies rely on ultrasonography (US) to measure muscle quality and size due to the imaging being a noninvasive and inexpensive procedure easily available to clinicians(1). Specifically, cross-sectional area (CSA) and echo intensity (EI) of skeletal muscles obtained with the use of ultrasound imaging provide crucial information regarding muscle composition and quality(2). CSA is a measurement of how large the muscle is, while EI is an index of muscle quality obtained through a gray-scale analysis of individual pixels within the US image. US images have been shown to be a valid measure compared to MRIs and CT scans for both CSA and EI(3–7).

Very often, the reliability of measurements is reported in the literature(8), yet studies sometimes simply specify intraclass correlation coefficients (ICCs) and other reliability statistics(9). When working with US imaging, it is important to understand that it is a two-step process. First, the technician completes the scan, and then someone analyzes the image using dedicated software. Image analysis also contains multiple steps, and it is imperative that all steps are done in a reliable and repeatable manner. Other key information related to reliability is if the technician is a highly skilled scanner, if the scanner's reliability statistics are high, if the reliability of the analyzer is high, or all of the above. Some researchers report ICCs from a single measurement with multiple raters ($ICC_{1,1}$)(10), several report a single rater with a single measurement ($ICC_{2,1}$)[11,12], and others report a single rater with the average of the means ($ICC_{2,k}$)[6]. ICCs tend to range from moderate (0.50–0.75) to excellent (>0.90). CSA tends to be the highest recorded ICCs with a range of (0.81–0.99)[3,7,8,12–14], while EI values tend to have a lower range of (0.71–0.98)[6,11,12,15,16]. The vast majority of these studies evaluate intra- and inter-rater reliability of the same muscle but measured on different days, and no study to date has evaluated the rater's reliability in measuring the same image.

In most method sections, the process of analyzing images is not well described due to the process having multiple steps and referring to "images were analyzed using". The first step is to set a scale by measuring the known difference between two points. If raters are analyzing the same image with varying scales, it can ultimately result in different values and hence a lower reliability value. Once the scale is set, raters use the polygon function to trace the muscle using the fascia as a guide[12]. If raters do not follow the fascia border, they may include the fascia and thus arbitrarily increase CSA and EI; conversely, if they are too conservative, they may miss parts of the muscle. Very often EI is normalized to subcutaneous fat, but where and how each researcher records the depth plays a major role in determining the correct EI values[16]. All of these integral steps in analyzing US images are a potential place of error in data analysis. Therefore, researchers performing US imaging need to be more informative of their reliability and the process to ensure there is a high reliability rate of US imaging across the literature. Consequently, the aim of this study is to understand the intra- and inter-rater reliability of each step in the process of analyzing US images to obtain EI and CSA values.

Material and methods

Ultrasound assessment

Twenty subjects (mean ± SD: age: 20.5 ± 2, height: 173 ± 7.3 cm, weight: 60 ± 6 kg) volunteered for this study. Prior to any testing, the participants read and signed an informed consent form and a health history questionnaire. All the subjects were free of any neurological disease or musculoskeletal injuries. The study was approved by the Institutional Review Board for protection of human participants. (FUIRB #: 10818).

US images were taken with a portable B-mode imaging device (GE Logiq e BT12, GE Healthcare, Milwaukee, WI, USA) and a multi-frequency linear-array probe (12 L-RS, 5–13 MHz, 38.4-mm field of view, GE Healthcare, Milwaukee, WI, USA). The panoramic function was used to obtain images of the right Rectus Femoris (RF) and Vastus Lateralis (VL) in the transverse plane. The images were taken at 1/2 of the distance between the anterior superior iliac spine and the superior border of the patella. A high-density foam pad was secured around the right thigh with an adjustable Velcro strap to ensure probe movement in the transverse plane. US settings (frequency: 10 MHz, gain: 45 dB, dynamic range: 72) were kept consistent across the participants. To scanning depth was standardized to 3.5, as all subjects' muscles fit in this window. A generous amount of water-soluble transmission gel (Aquasonic 100 ultrasound transmission gel, Parker Laboratories, Inc., Fairfield, NJ, USA) was applied to the skin, so that it immersed the probe surface during testing in order to enhance acoustic coupling.

The US images were digitized and examined with ImageJ Software (version 1.46, National Institutes of Health, Bethesda, MD, USA). First, the images were scaled to 1 cm using the line function and distance marks on the image. Next, subcutaneous fat thickness measurements were taken at three locations and averaged using the line function[16]. The polygon function was used to outline the border of the RF and VL, and then both the EI and CSA were measured and assessed by computer-aided grey-scale analysis using the histogram function. The EI values were determined as the corresponding index of muscle quality ranging between 0 and 255 A.U. (black = zero, white = 255) (Fig. 1).

Raters

Two raters underwent two weeks of US analysis training during which they analyzed over 120 images containing panoramic US images of the VL and RF muscles. Once their training was deemed satisfactory (consistent analysis that resulted in similar values across multiple images), they were assigned to analyze 40 images. Twenty original images were blinded and doubled, so that there was a total of 40 images to analyze, with each image being analyzed twice. The images were then unblinded and used to compare values for the purpose of evaluating reliability.

Statistical analysis

Intra-rater and inter-rater reliability were analyzed using model "3,1" for both raters, 1 & 2[9]. Reliability analysis was conducted on scaled units, depth (subcutaneous fat thickness), and EI & CSA, for both the RF and VL. The statistics of interests were intraclass correlation coefficients (ICCs), the coefficient of variation, the standard error of measurement (SEM), and the minimum difference (MD) values. In addition, SEM values were also expressed as percentages of the mean. Test-retest reliability data were analyzed using dedicated software (Microsoft Excel, Microsoft Corporation, Redmond, WA, USA). Systematic variability for each variable across the testing days was examined using separate one-way repeated-measure ANOVAs[17]. Alpha levels were set a priori at $p \leq 0.05$ to determine statistical significance.

Results

All results for intra-rater reliability are displayed in Tab. 1. The results for inter-rater reliability are shown in Tab. 2.

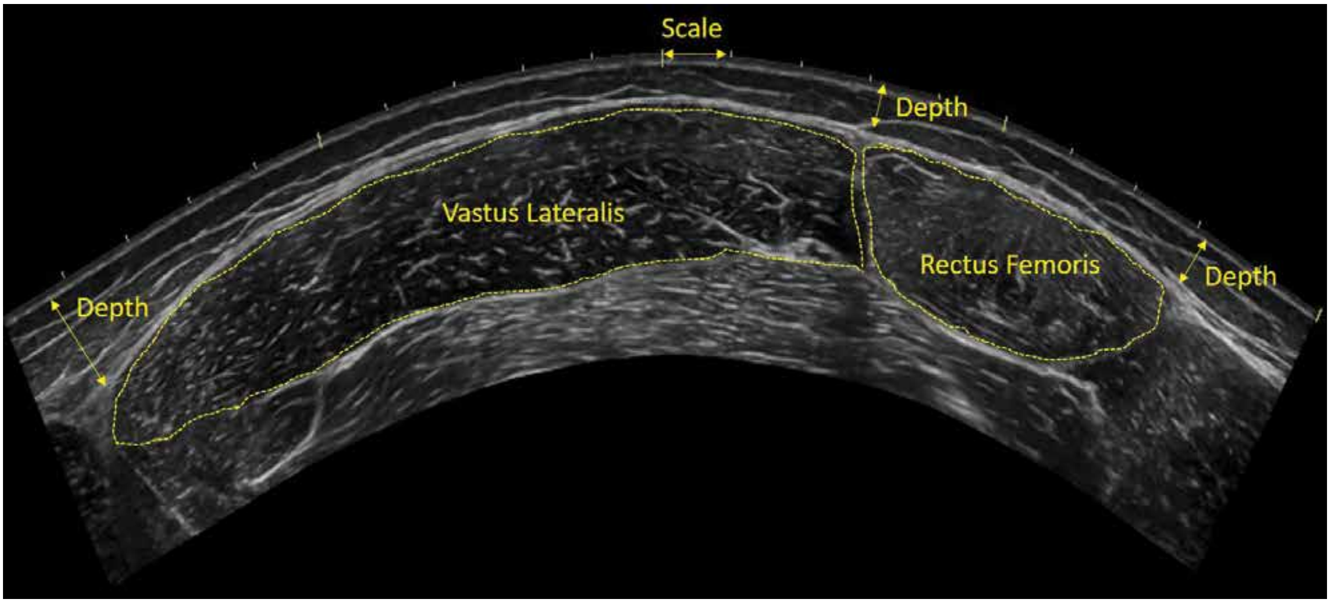

Fig. 1. *Ultrasound image measurements analyzed for reliability*

Discussion

As this is one of the first studies to address the reliability of the process of analyzing US images, we demonstrated relatively high levels of reliability when it came to setting the scale and depth, with intra-rater reliability of (0.85–0.96), and inter-rater reliability of (0.87–0.97). This demonstrates that when the scales are set on the same image, the process can be considered reliable across raters. This simple step, which is often overlooked, could be a point of error, but as the results showed, there is a high reliability level in trained raters.

Regarding reliability measurements of CSA, the ICC values were (0.67–0.88) for intra-rater, and (0.87–0.92) for inter-rater analysis. Our intra-rater reliability showed just how varied a rater could be when analyzing the same image. In comparison to other studies[11,12,16,18–20], we had similar ICC values ranging from an inter-rater reliability of 0.75–0.99 and an intra-rate reliability of 0.551–0.92. For EI values, our results were lower, showing poor intra-rater reliability (0.32–0.55), but moderate inter-rater reliability (0.77–0.91). Research indicates that the reliability for EI values ranges from 0.72 to 0.92[11,12,15,21,22], with Jenkins *et al*.[11] reporting that the confidence of an ICC for EI is 0.44–0.92, in comparison with the confidence interval of CSA amounting to 0.96–0.99.

The coefficient of variation (CV) values were small in setting the pixel scale (1.89–2.51%), while the depth had a larger variation, ranging from 7.59 to 10.97%. For CSA, the CV values were in the range of 3.36–11.66%, and for EI they ranged from 2.95 to 9.93%. Our studies CV compared to others[6,18,23,24] within a range of 2–5% appears to be larger in variation, so even though EI has a lower reliability rate, it tends to have lower CV values, which indicates reproducibility between measurements. Both CSA and EI have low levels of variation, which shows good relative reliability[11].

Another measurement of reliability is standard error of measurement. For the scaling and depth measurements, the relative SEM (%) was good, but we are unable to compare this finding to other data, for this is the only study to look at the scaling process. However, CSA having SEM (%) reported between 5–12% isin agreement with the literature[3,11,12]. EI also has low relative SEM values, in the range of 2–10%, with absolute SEM values of (0.56–1.05), which are similar to published data[8,11,12,25–27].

Based on the results of this study and the available literature, the reliability values of EI are lower than CSA. This could be due to the sensitivity of the measurement process. CSA is based on the scale and tracing area of the muscle, so if you slightly overtrace, the impact on total CSA might be one hundredth of a centimeter off. However, if you overtrace with EI, you may include white pixels that are part of the fascia, thus inflating the values. We demonstrated that EI is a more sensitive measurement, as the MD values (0.78–2.92) in the current study were lower than what has been reported in the literature[11,12,15,27]. This difference could be due to the fact that in previous studies the reliability values referred to comparing different images over multiple days, while this study compared the same images evaluated twice. Our study resulted in similar MD values for CSA (1–4 cm^2) as in the previous literature[11,12,16]. MD values often fail to be reported in reliability statistics, but they are a useful tool when analyzing changes over time, offering a reference to what is clinically significant compared to what is statistically significant based on the reliability of data[17].

The majority of reliability studies address US reliability by analyzing different images of the same muscle, and with that comes the variability of image, as not every image is the same. Factors that are known to affect US scans include the curvature of the limb, angling of the probe[28], pressure of the probe, and location of the probe[11]. All of these

Tab. 1. *Intra-rater reliability statistics ultrasound derived measurements of the leg extensor muscles*

	Scale		Depth		Rectus femoris				Vastus lateralis			
Variable	**(pixels/cm)**		**(cm)**		**CSA (cm^2)**		**EI (AU)**		**CSA (cm^2)**		**EI (AU)**	
RATER	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
Average	41.5	41.4	0.82	0.77	6.29	6.21	10.04	10.02	21.78	21.66	12.04	11.91
CV %	2.51	1.89	10.97	10.12	11.66	6.72	6.11	9.93	6.01	5.54	6.95	5.61
***P* value**	0.45	0.42	0.86	0.52	0.47	0.25	0.40	0.36	0.13	0.07	0.32	0.22
ICC$_{3,1}$	0.85	0.88	0.96	0.98	0.74	0.67	0.36	0.32	0.85	0.88	0.55	0.47
SEM	1.69	1.60	0.05	0.04	0.68	0.77	1.00	1.05	1.44	1.38	0.71	0.72
SEM (%)	4.07	3.86	6.10	5.19	10.81	12.40	9.96	10.48	6.61	6.37	5.90	6.05
MD	4.68	4.43	0.14	0.11	1.88	2.15	2.79	2.92	4.00	3.83	1.97	2.00

CV – coefficient of variation expressed as %; *p* value – type 1 error rate for the repeated measures ANOVA across images; ICC$_{3,1}$ – intraclass correlation coefficient model$_{3,1}$; SEM – standard error of measurement; SEM (%) – standard error of measurement expressed as percentage of the mean; MD – minimum difference to be consider real

Tab. 2. *Inter-rater reliability statistics ultrasound derived measurements of the leg extensor muscles*

	Scale	Depth	Rectus Femoris		Vastus Lateralis	
Variable	**(pixels/cm)**	**(cm)**	**CSA (cm^2)**	**EI (AU)**	**CSA (cm^2)**	**EI (AU)**
Average	41.54	0.79	6.30	10.00	21.72	11.98
CV %	2.01	7.59	7.82	3.91	3.36	2.95
***P* value**	0.68	0.4	0.12	0.21	0.98	0.45
ICC$_{3,1}$	0.87	0.97	0.87	0.77	0.92	0.91
SEM	1.62	0.04	0.47	0.56	1.11	0.28
SEM (%)	3.90	5.06	7.46	5.60	5.11	2.34
MD	4.49	0.13	1.30	1.56	3.06	0.78

CV – coefficient of variation expressed as %; *p* value – type 1 error rate for the repeated measures ANOVA across images; ICC$_{3,1}$ – intraclass correlation coefficient model$_{3,1}$; SEM – standard error of measurement; SEM (%) – standard error of measurement expressed as percentage of the mean; MD – minimum difference to be consider real

aspects affect reliability over time when analyzing different images of the same muscle. Our study, which is the first of its kind, evaluated the ability to analyze the same image with different raters. The goal was to understand areas of variability and the potential for errors in the analysis of US images.

Conclusions

Overall, measuring reliability for each step during the research process is key to determining whether US imaging is a suitable method for CSA and EI measurements. Based on the results of this study and current literature[8,16,27], CSA and EI have moderate to high reliability values. This is the first study to address the reliability of scale and depth measurements, both of which displayed a high level of reliability. However, as all reliability values are specific to the measurement and the rater, these factors need to be taken into consideration when interpreting US-derived measurements. Every paper that utilizes US measurements should include their own reliability statistics, which should be reported in the results section. In studies relying on US to monitor changes over time or between groups, it is highly suggested that minimum difference to be considered real (MD) be used as part of the analysis, considering that clinical significance and statistical significance in these measurements can vary.

Conflict of interest

The authors do not report any financial or personal connections with other persons or organizations which might negatively affect the contents of this publication and/or claim authorship rights to this publication.

References

1. Mechelli F, Arendt-Nielsen L, Stokes M, Agyapong-Badu S: Validity of ultrasound imaging versus magnetic resonance imaging for measuring anterior thigh muscle, subcutaneous fat, and fascia thickness. Methods Protoc 2019; 2: 58.
2. Mayans D, Cartwright MS, Walker FO: Neuromuscular ultrasonography: quantifying muscle and nerve measurements. Phys Med Rehabil Clin N Am 2012; 23: 133–148.
3. Ahtiainen JP, Hoffren M, Hulmi JJ, Pietikäinen M, Mero AA *et al.*: Panoramic ultrasonography is a valid method to measure changes in skeletal muscle cross-sectional area. Eur J Appl Physiol 2010; 108: 273–279.
4. Reeves ND, Maganaris CN, Narici MV: Ultrasonographic assessment of human skeletal muscle size. Eur J Appl Physiol 2004; 91: 116–118.
5. Pillen S, Tao RO, Zwarts MJ, Lammens MM, Verrijp KN, Arts IM *et al.*: Skeletal muscle ultrasound: correlation between fibrous tissue and echo intensity. Ultrasound Med Biol 2009; 35: 443–446.
6. Young HJ, Jenkins NT, Zhao Q, Mcully KK: Measurement of intramuscular fat by muscle echo intensity. Muscle Nerve 2015; 52: 963–971.

7. Scott JM, Martin DS, Ploutz-Snyder R, Caine T, Matz T, Arzeno NM *et al.*: Reliability and validity of panoramic ultrasound for muscle quantification. Ultrasound Med Biol 2012; 38: 1656–1661.
8. English C, Fisher L, Thoirs K: Reliability of real-time ultrasound for measuring skeletal muscle size in human limbs in vivo: a systematic review. Clin Rehabil 2012; 26: 934–944.
9. Caresio C, Molinaro F, Emanuel G, Minetto MA: Muscle echo intensity: reliability and conditioning factors. Clin Physiol Funct Imaging 2015; 35: 393–403.
10. Fukumoto Y, Ikezoe T, Yamada Y, Tsukagoshi R, Nakamura M, Mori N *et al.*: Skeletal muscle quality assessed from echo intensity is associated with muscle strength of middle-aged and elderly persons. Eur J Appl Physiol 2012; 112: 1519–1525.
11. Jenkins ND, Miller JM, Bucker SL, Sochrane KC, Bergstrom HC, Hill EC *et al.*: Test–retest reliability of single transverse versus panoramic ultrasound imaging for muscle size and echo intensity of the biceps brachii. Ultrasound Med Biol 2015; 41: 1584–1591.
12. Rosenberg JG, Ryan ED, Sobolewski EJ, Scharville MJ, Thompson BJ, King GE: Reliability of panoramic ultrasound imaging to simultaneously examine muscle size and quality of the medial gastrocnemius. Muscle Nerve 2014; 49: 736–740.
13. Radaelli R, Bottaro M, Wilhelm EN, Wagner DR, Pinto RS: Time course of strength and echo intensity recovery after resistance exercise in women. J Strength Cond Res 2012; 26: 2577–2584.
14. Korhonen MT, Mero AA, Alén M, Sipilä S, Häkkinen KLiikavainio T *et al.*: Biomechanical and skeletal muscle determinants of maximum running speed with aging. Med Sci Sports Exerc 2009; 41: 844–856.
15. Vieira A, Siqueira A, Ferreira-Junior JB, Pereira P, Wagner D, Bottaro M: Ultrasound imaging in women's arm flexor muscles: intra-rater reliability of muscle thickness and echo intensity. Braz J Phys Ther 2016; 20: 535–542.
16. Burton AM, Stock MS: Consistency of novel ultrasound equations for estimating percent intramuscular fat. Clin Physiol Funct Imaging 2018; 38: 1062–1066.
17. Weir JP: Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005; 19: 231–240.
18. e Lima KM, da Matta TT, de Oliveira LF: Reliability of the rectus femoris muscle cross-sectional area measurements by ultrasonography. Clin Physiol Funct Imaging 2012; 32: 221–226.
19. Gellhorn AC, Carlson MJ: Inter-rater, intra-rater, and inter-machine reliability of quantitative ultrasound measurements of the patellar tendon. Ultrasound Med Biol 2013; 39: 791–796.
20. Kreienbrink K, Ahrens J, Everson E, Barmore D, [Kernozek T, Ragan R]: Intra-and inter-rater reliability of ultrasound measurements of Achilles cross-sectional area. UWL J Undergrad Res 2019; 22.
21. Sobolewski EJ, Hall AB, Rodriguez GC, Richard MO: Ultrasound derived muscle cross-sectional area and echo intensity is unable to detect differences among acute aerobic exercise bouts with varying duration and intensity. Int J Sports Sci 2020; 10: 31–37.
22. Strasser EM, Draskovits T, Praschak M, Quittan M, Graf A: Association between ultrasound measurements of muscle thickness, pennation angle, echogenicity and skeletal muscle strength in the elderly. Age (Dordr) 2013; 35: 2377–2388.
23. Watanabe Y, Yamada Y, Fukumoto Y, Ishihara T, Yokoyama K, Yoshida T *et al.*: Echo intensity obtained from ultrasonography images reflecting muscle strength in elderly men. Clin Interv Aging 2013; 8: 993–998.
24. Varanoske AN, Fukuda DH, Boone CH, Beyer KS, Stout JR, Hoffman JR: Scanning plane comparison of ultrasound-derived morphological characteristics of the vastus lateralis. Clin Anat 2017; 30: 533–542.
25. Scanlon TC, Fragala MS, Stout JR, Emerson NS, Beyer KS, Oliveira LP *et al.*: Muscle architecture and strength: adaptations to short-term resistance training in older adults. Muscle Nerve 2014; 49: 584–592.
26. Jajtner AR, Hoffman JR, Scanlon TC, Wells AJ, Townsend JR, Beyer KS *et al.*: Performance and muscle architecture comparisons between starters and nonstarters in National Collegiate Athletic Association Division I women's soccer. J Strength Cond Res 2013; 27: 2355–2365.
27. Stock MS, Whitson M, Burton AM, Dawson NT, Sobolewski EJ, Thompson BJ: Echo intensity versus muscle function correlations in older adults are influenced by subcutaneous fat thickness. Ultrasound Med Biol 2018; 44: 1597–1605.
28. Ishida H, Suehiro T, Suzuki K, Watanabe S: Muscle thickness and echo intensity measurements of the rectus femoris muscle of healthy subjects: Intra and interrater reliability of transducer tilt during ultrasound. J Bodyw Mov Ther 2018; 22: 657–660.