Journal of Bodywork & Movement Therapies 48 (2026) 1045–1050

Contents lists available at ScienceDirect

Journal of Bodywork & Movement Therapies

journal homepage: www.elsevier.com/jbmt

Intra- and inter-rater reliability of ultrasound assessment of rectus femoris muscle thickness by physiotherapists: influence of operator experience

Daniele Conte [a,*], Giulia Costantini [b], Chiara Pinzini [b,c], Giuliana De Maio [b,c], Maria Parpinel [a], Daniele Vidi [b,c]

[a] Department of Medicine, University of Udine, Udine, 33100, Italy
[b] School of Physiotherapy, Department of Medicine, University of Udine, Udine, 33100, Italy
[c] Institute of Physical Medicine and Rehabilitation "Gervasutta", Friuli Centrale University Health Authority, Udine, 33100, Italy

A R T I C L E I N F O

Keywords:
Reliability
Ultrasonography
Rehabilitative ultrasound imaging
Physical therapists
Quadriceps muscle
Muscle thickness

A B S T R A C T

Background: Rehabilitative ultrasound imaging is increasingly used in physiotherapy as an objective method for muscle assessment. However, its reliability across different levels of operator experience requires further clarification.
Objective: To evaluate the intra- and inter-rater reliability of rectus femoris thickness measurements among expert and novice physiotherapists.
Methods: Four physiotherapists (two experts, two novices) assessed rectus femoris thickness in healthy individuals. Reliability was evaluated using intraclass correlation coefficients (ICC) and limits of agreement (LoA). The standard error of measurement (SEM) and minimal detectable change (MDC) were calculated to quantify measurement precision.
Results: Both expert raters demonstrated good-to-excellent intra-rater reliability (Rater A: $ICC_{2,1}$ 0.84–0.98; Rater B: $ICC_{2,1}$ 0.81–0.98). Novice raters showed good-to-excellent (Rater C: $ICC_{2,1}$ 0.80–0.98) and moderate-to-excellent reliability (Rater D: $ICC_{2,1}$ 0.73–0.97). The pooled SEM was 0.48 mm, with no clinically relevant differences between raters. The MDC was 1.34 mm. Inter-rater reliability was moderate-to-good for single measurements and good-to-excellent for average measurements, with similar patterns across expertise groups. LoA ranged from −3.13 to 3.05 mm for experts and from −3.25 to 2.09 mm for novices.
Conclusions: Ultrasound assessment of rectus femoris thickness demonstrates high intra-rater reliability among physiotherapists, supporting its use in clinical practice. Our preliminary findings suggest that novice raters may approach expert-level reliability following a short training protocol. Inter-rater reliability was acceptable and improved when averaging multiple measurements. For longitudinal monitoring, assessments are likely to be more consistent when performed by the same operator; when this is not feasible, averaging repeated measurements may help reduce inter-operator variability.

1. Introduction

Rehabilitative ultrasound imaging (RUSI) has gained increasing attention within musculoskeletal (MSK) care as a non-invasive, real-time method for the assessment of muscle morphology and function. Its clinical applications include, but are not limited to, the evaluation of muscle thickness (MT), cross-sectional area (CSA), and activation patterns, supporting both diagnostic processes and the monitoring of rehabilitation outcomes. Compared to other imaging modalities, ultrasound (US) offers advantages in terms of cost and feasibility at the point of care, making it particularly suitable for repeated assessments in clinical settings (Whittaker et al., 2019).

Within MSK rehabilitation, quadriceps muscle function plays a central role in determining patient outcomes. Adequate quadriceps strength is essential for activities of daily living, functional mobility, and return to sport. For example, following anterior cruciate ligament (ACL) reconstruction, evidence suggests that quadriceps strength is a strong determinant of patient-reported function during the later stages of

* Corresponding author. Department of Medicine, University of Udine, Via Colugna, 50, 33100, Udine, Italy.
E-mail addresses: conte.daniele@spes.uniud.it (D. Conte), costantini.giulia@spes.uniud.it (G. Costantini), chiara.pinzini@uniud.it (C. Pinzini), giuliana.demaio@uniud.it (G. De Maio), maria.parpinel@uniud.it (M. Parpinel), daniele.vidi@uniud.it (D. Vidi).

https://doi.org/10.1016/j.jbmt.2026.08.020
Received 12 May 2026; Received in revised form 3 July 2026; Accepted 12 August 2026
Available online 13 August 2026

rehabilitation (Chaput et al., 2021). Similarly, in individuals with knee osteoarthritis, strength training programs have been shown to reduce pain and improve function (Hegde et al., 2025). Among the quadriceps muscles, the rectus femoris (RF) is frequently used as a representative site for US assessment due to its superficial location and accessibility, making it a practical marker of overall quadriceps morphology. Additionally, RF measurements have been shown to correlate with knee extensor strength (Şekir et al., 2022).

Beyond orthopaedics and sports rehabilitation, the assessment of quadriceps muscle mass, often including RF thickness, has become increasingly relevant in critical care settings. In intensive care units (ICUs), US-based evaluation of MT is used to monitor muscle wasting, sarcopenia, and functional decline in critically ill patients. Reductions in quadriceps muscle size have been associated with decreased strength, prolonged recovery, and poorer functional outcomes, highlighting the importance of accurate and reliable measurement techniques in these populations (Parry et al., 2015; Puthucheary et al., 2013).

Despite these clinically relevant applications, the use of US by physiotherapists (PTs) remains variable across healthcare systems. In some contexts, its implementation is limited by regulatory restrictions, variability in training pathways, and uncertainty regarding scope of practice (Whittaker et al., 2019). In addition, although a growing body of literature supports the reliability of US measurements, relatively few studies have specifically examined reliability in PTs with differing levels of experience. Existing evidence often focuses on experienced operators or mixed professional groups, which may limit the generalizability of findings to routine physiotherapy practice.

Establishing the reliability of US measurements when performed by PTs is a key step toward broader clinical integration. Reliable measurement is essential to distinguish true physiological changes from measurement error, particularly when monitoring longitudinal adaptations such as muscle atrophy or hypertrophy (Franchi et al., 2018; Mokkink et al., 2023). Furthermore, understanding how operator experience influences measurement reliability may inform training strategies and clinical implementation.

Therefore, the primary aim of this study was to evaluate the intra- and inter-rater reliability of US assessment of RF muscle thickness among PTs. The secondary aim was to determine the influence of operator experience by comparing the reliability achieved by expert and novice raters. By providing initial data specific to this professional group in healthy individuals, this study seeks to strengthen the evidence base supporting the use of RUSI in physiotherapy practice and to contribute to the development of a more consistent and evidence-informed approach to its clinical application.

2. Methods

The protocol of the present study was approved by the Institutional Review Board of the Department of Medicine at the University of Udine (Reference: Prot IRB 259/2025, Tit III, cl 13, fasc.24/2025). All participants provided written informed consent prior to enrolment. The study was reported in accordance with the Guidelines for Reporting Reliability and Agreement Studies (GRAAS) (Kottner et al., 2011).

2.1. Participants

A convenience sample of healthy volunteers was recruited from the bachelor's degree program in Physiotherapy at the [Blinded for Review]. Participant recruitment was conducted via a formal announcement posted on the official notice board of the physiotherapy course, inviting students to volunteer for the study.

Potential participants were screened for eligibility based on predetermined exclusion criteria. Individuals were excluded if they reported a history of major lower limb injury within the preceding six months that required prolonged rest, a period of physical inactivity, or the use of assistive devices (e.g., crutches).

The required sample size was calculated according to Walter et al. (1998) using the following parameters: three replicates, $\rho0 = 0.6$, $\rho1 = 0.95$, $\alpha = 0.05$, power $(1\text{-}\beta) = 80\%$. $\rho0$ was determined based on the threshold for clinical relevance in reliability measurements (Munro, 2005) whereas $\rho1$ was selected based on data from previous literature. Based on these parameters, the minimum required sample size was determined to be 6 subjects.

2.2. Raters

Measurement procedures were performed by four PTs (designated as A, B, C and D), each holding both a bachelor's degree in Physiotherapy and a master's degree in Rehabilitation Sciences. The raters were further categorized by their previous expertise in US assessment. Raters A and B were considered experts, as both possessed over 10 years of clinical experience in musculoskeletal ultrasound and had completed formal postgraduate training in the field. Conversely, Raters C and D were novices, with no prior practical expertise to US imaging.

Before the beginning of the study, the novice raters underwent a 6-h training program divided into two 3-h sessions, delivered by an experienced physiotherapist with over 10 years of clinical ultrasound practice. This training focused on the basic principles of MSK-US and dedicated 5 h to hands-on practice of the rectus femoris acquisition protocol. While a rigid competency criterion was not established, the trainer qualitatively assessed that both novices were fully proficient with the fundamental procedures prior to data collection.

2.3. Study procedures

All data collection was performed during a single experimental session. US images were acquired using a portable US system (LOGIQ e, GE Medical Systems, software version BT12) equipped with a 12 L-RS wideband multifrequency linear array probe (8–13 MHz). Imaging was performed in B-mode with the frequency set at 10 MHz, image depth at 6 cm, gain at 74%, and a single focus positioned at 3.5 cm. Probe pressure was minimized during acquisition to avoid muscle tissue deformation.

Upon entering the laboratory, participants underwent an initial screening where demographic information and physical activity related data were recorded. Body composition was assessed via bioelectrical impedance analysis (BIA).

Participants were positioned supine on a standard examination table, with the head section of the bed elevated to an angle between 20° and 30°. The lower limbs were maintained in a neutral, fully extended position.

The target site for ultrasound imaging was RF of the right leg. The probe positioning site was defined as the midpoint between the anterior superior iliac spine (ASIS) and the superior margin of the patella (Pardo et al., 2018). This distance was measured using a flexible tape, and the exact location was marked on the skin with a dermatographic pencil (eye pencil) which allowed for easy removal between assessments. Each rater independently identified, measured, and marked the anatomical site for each assessment; the mark was completely removed between sessions to ensure raters remained entirely blinded to previous landmarking.

Among each participant, three separate imaging capturing sessions were carried out. Each session involved a different, computer-generated random sequence of the four raters. The randomization was constrained to ensure that no rater performed two consecutive acquisitions (i.e., a rater could not be the final operator of one session and the initial operator of the subsequent session). All raters were blinded to the image acquisition procedure of their peers.

Each rater captured a single transverse image of the RF per session, resulting in 3 images per rater for each subject. Given an approximate acquisition time of 2 min per rater, a minimum interval of 6 min was maintained between measurements by the same operator for any given subject. Participants were repositioned between every individual measurement.

2.4. Image measurement

Following the experimental session, all US images were exported and processed for thickness measurements. An investigator who was not involved in the image acquisition process performed the measurements and, to ensure complete blinding during the process, all files were renamed using a unique alphanumeric code, and the measurement sequence was randomized.

MT of the RF was quantified using ImageJ software (Schneider et al., 2012). For each image, three separate measurements were taken within a 1 cm horizontal range centered on the middle of the ultrasound frame; specifically, one measurement was taken at the center and two at the lateral margins of this 1 cm reference line. Thickness was defined as the distance between the internal surfaces of the superior and inferior connective tissue borders surrounding the muscle. The mean of these three measurements was calculated to provide a single representative value for each image, which was used for all subsequent statistical analyses. An example of the measurement procedure is presented in Fig. 1.

2.5. Statistical analysis

All statistical analyses were conducted using R version 4.5.2 (R Core Team, 2025) within the RStudio integrated development environment version 2026.01.2 + 418 (Posit team, 2026).

Reliability and agreement analyses were performed using the SimplyAgree package (Caldwell, 2022), applying linear mixed-effects models (LMM) for variance component estimation. Compared to the traditional sums of squares (i.e., analysis of variance) based approaches, LMM provide more robust estimates in the presence of repeated measurements by accounting for within-subject correlation and allows for the handling of missing data without the need for a complete-case approach, as required in ANOVA (Chen et al., 2018).

Normality of the data distribution was assessed using the Shapiro–Wilk test. Relative reliability was evaluated using the intraclass correlation coefficient (ICC). Following the framework proposed by Koo and Li (2016), ICC model selection was based on the "Model," "Type," and "Definition." A two-way model was used for all analyses. Inter-rater reliability was framed within a two-way random-effects model to allow for broader clinical generalization. Conversely, intra-rater reliability was defined by a two-way mixed-effects model, reflecting a research design where the specific raters involved are treated as fixed factors for the purpose of assessing the reliability of their individual measurements.

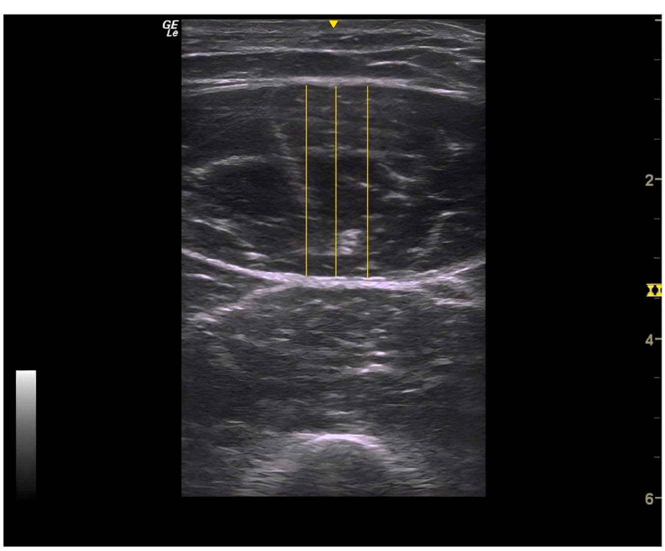

Fig. 1. Example of ultrasound image acquisition of the rectus femoris muscle. The three yellow vertical lines represent the positions used for thickness measurement. These measurements are placed along a 1 cm horizontal reference line (one in the center and two at the extremities). The three measurements were then averaged to obtain a single value for the image. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

An absolute agreement definition was adopted for all analyses to evaluate the extent to which raters provide identical measurement values, rather than merely maintaining consistent ranking across subjects.

For intra-rater reliability, the "single measurement" type was selected to reflect the reality of clinical practice, where a clinician typically performs and relies upon a single image acquisition. Conversely, inter-rater reliability was assessed using both single-measure and average-measure ICCs to allow for a comparison of distinct clinical workflows. The single-measure ICC was utilized to model a scenario where different clinicians each obtain a single scan and must agree on the resulting value. The average-measure ICC ($k = 3$) was calculated from the three repeated acquisitions per rater. This measure was included as a comparative scenario to evaluate if inter-rater agreement is significantly enhanced when clinicians utilize a more rigorous protocol, specifically averaging three image replicates before reporting a final value.

According to the Shrout and Fleiss (1979) convention, these selections correspond to ICC(2,1) for both single-measure intra- and inter-rater reliability, as the mathematical formulations for absolute agreement in single-measure models are identical across both random and mixed-effects frameworks, and ICC(2,k) for average-measure inter-rater reliability.

To allow comparison of measurement precision independent of absolute muscle thickness, the coefficient of variation (CV%) was calculated as an index of relative variability. Absolute measurement precision was quantified using the standard error of measurement (SEM), derived from the mean square error (MSE) of the model. The minimal detectable change at the 95% confidence level (MDC_{95}) was calculated to define the threshold for true change beyond measurement error.

Agreement between raters was evaluated through systematic bias and 95% limits of agreement (LoA). The Method of Variance Estimates Recovery (MOVER) was employed to calculate the 95% confidence intervals (CI) of the LoA, as this approach provides a more conservative estimate compared to the traditional Bland-Altman method (Donner and Zou, 2012). LoA were specifically computed to compare raters within the same experience group (experts and novices) in order to assess agreement associated with operator experience. Given the presence of replicated measurements within subjects (i.e., multiple repeated measurements per rater on the same subject, where the underlying true value was not expected to change), the model accounted for within-subject correlation.

ICC values were interpreted according to the following benchmarks: <0.50 (poor), 0.50–0.75 (moderate), 0.75–0.90 (good), and >0.90 (excellent) (Koo and Li, 2016).

3. Results

3.1. Sample characteristics

A total of eight healthy subjects (four males and four females) were recruited. Based on our a priori power analysis, the minimum necessary sample size estimated was six subjects but, to account for potential scheduled absences on the day of assessment, this sample size was selected to ensure the study remained adequately powered in the event of unexpected dropouts.

The study population consisted of young adults with a mean age of 20.9 ± 0.4 years and a mean body mass index (BMI) of 22.7 ± 1.2 kg/m^2. Participants were physically active, reporting an average of 6.1 ± 2.0 h of training per week. Body composition analysis revealed a mean fat-free mass (FFM) of 82.4 ± 5.1% and a skeletal muscle mass (SMM) of 43.8 ± 4.7%. Detailed demographic and anthropometric characteristics, stratified by sex, are presented in Table 1.

3.2. Intra-rater reliability

The intra-rater reliability results for each of the four raters are presented in Table 2.

For the expert raters, Rater A demonstrated an ICC(2,1) of 0.94 (95% CI: 0.84, 0.98), while Rater B showed an ICC(2,1) of 0.92 (95% CI: 0.81, 0.98). Based on the 95% CI, expert reliability ranged from good to excellent. The expert group showed a CV ranging from 1.74% to 2.14%, with an SEM between 0.42 mm and 0.52 mm, and an MDC_{95} between 1.17 mm and 1.45 mm.

Regarding the novice raters, Rater C demonstrated an ICC(2,1) of 0.92 (95% CI: 0.80, 0.98), and Rater D showed an ICC(2,1) of 0.89 (95% CI: 0.73, 0.97). Based on the 95% CI, novice reliability ranged from moderate to excellent. The novice group exhibited a CV between 1.68% and 2.33%, with an SEM ranging from 0.41 mm to 0.58 mm, and an MDC_{95} between 1.13 mm and 1.61 mm.

3.3. Inter-rater reliability

The inter-rater reliability results for all four rater and the experts and novice groups are presented in Table 3.

Total inter-rater reliability resulted in an ICC(2,1) of 0.69 (95% CI: 0.55–0.81) and an ICC(2,3) of 0.90 (95% CI: 0.83–0.95), indicating moderate-to-good and good-to-excellent agreement, respectively.

For experts, ICC(2,1) was 0.71 (95% CI: 0.49–0.84) and ICC(2,3) was 0.83 (95% CI: 0.66–0.92), with reliability ranging from poor to excellent and from good to excellent respectively. Mean bias was −0.04 mm (95% CI: −1.28; 1.19) and LoA from −3.13 to 3.05 mm.

For novices, ICC(2,1) was 0.60 (95% CI: 0.33-0.78) and ICC(2,3) was 0.75 (95% CI: 0.50–0.88), representing poor to good and moderate to good reliability respectively. Mean bias was −0.58 mm (95% CI: −1.61; 0.45) and LoA from −3.25 to 2.09 mm.

4. Discussion

The findings of this study indicate that US-based measurement of RF thickness is a reliable tool when performed by PTs. Although the overall results suggest that level of expertise does not represent a clinically meaningful barrier to achieving high reliability, a more detailed examination of individual rater performance provides important insights into measurement precision among novice users and contributes to a more comprehensive interpretation of the findings.

4.1. Main reliability findings and influence of operator experience

From an overall perspective, a key finding is the absence of clinically meaningful differences in absolute reliability between expertise levels.

Table 1
General characteristics of the study sample, total and stratified by sex.

	Males (n = 4)	Females (n = 4)	Total (n = 8)
Age (y)	20.8 ± 0.5	21.0 ± 0.0	20.9 ± 0.4
Height (cm)	177.5 ± 3.3	165.0 ± 3.6	171.3 ± 7.4
Weight (kg)	72.9 ± 3.9	60.3 ± 1.7	66.6 ± 7.3
BMI (kg/m^2)	23.2 ± 1.5	22.2 ± 0.7	22.7 ± 1.2
FM (%)	15.7 ± 6.3	19.5 ± 3.5	17.6 ± 5.1
FFM (%)	84.4 ± 6.3	80.5 ± 3.5	82.4 ± 5.1
SMM (%)	46.2 ± 5.5	41.5 ± 3.1	43.8 ± 4.7
Wk training (h)	6.9 ± 2.4	5.2 ± 1.3	6.1 ± 2.0

Y = years; kg = kilograms; cm = centimeters; m = meters; FM = fat mass; FFM = fat-free mass; SMM = skeletal muscle mass; Wk = week; h = hours.

The pooled SEM for experts (0.47 mm) was nearly identical to that of novices (0.49 mm), with a difference of only 0.02 mm, which is negligible from a clinical standpoint. This suggests that, when a standardized protocol is followed, the measurement error of PTs might not be heavily influenced by years of experience alone. However, while a high level of reliability appears achievable across raters, it was not uniformly observed. For example, Rater C (novice) achieved a level of precision (SEM = 0.41 mm) comparable to, or slightly better than, that of the expert raters, whereas Rater D (novice) demonstrated the lowest precision (SEM = 0.58 mm). These findings indicate that brief standardized training may allow novices to approach expert-level reliability, although this is not consistently achieved. It should also be noted that these differences are small in relative terms: Rater C and Rater D demonstrated coefficients of variation of 1.68% and 2.33%, respectively, suggesting that even the least precise rater exhibited limited variability that is unlikely to meaningfully affect clinical interpretation.

The inter-rater analysis provides additional insight into agreement between raters. When considering single-measure reliability, both groups demonstrated moderate agreement, with slightly higher values in the expert group. When average-measure reliability was considered, both groups reached higher levels of agreement, indicating that reliability improves when measurements from multiple raters are combined.

Agreement analysis based on the 95% LoA showed comparable ranges between groups. The expert group (A + B) demonstrated LoA from −3.13 to 3.05 mm, while the novice group (C + D) showed LoA from −3.25 to 2.09 mm. The magnitude of systematic bias was small in both groups (−0.04 mm for experts and −0.58 mm for novices), with confidence intervals including zero, indicating no statistically significant bias. These findings suggest that, at a group level, agreement between raters is similar regardless of experience.

4.2. Contextualisation within the existing literature

Our findings are consistent with and extend the existing literature supporting the use of RUSI. With respect to the RF, our results align with those of Takahashi et al. (2021), who reported high reliability between trained PTs. While that study demonstrated the proficiency of experienced operators, our findings suggest that similar levels of reliability may also be achieved by less experienced clinicians after a short training protocol.

The accessibility of US for less experienced raters is further supported by Pinto-Ramos et al. (2022), who reported high reliability in point-of-care RF assessments following minimal training. Although the background of their raters was not clearly specified, the present study adds to this evidence by indicating that less experienced clinicians may acquire these skills relatively rapidly when guided by a structured protocol. Comparable findings have also been reported in other professional groups, as shown by Stausholm et al. (2024), where non PTs raters achieved similar reliability.

The reliability patterns observed in this study are consistent with those reported in other MSK-US applications. For example, Valera-Calero et al. (2020) demonstrated comparable intra- and inter-rater reliability between expert and novice raters when assessing the cervical multifidus. Together, these findings suggest that the difference in performance between experienced and novice operators may be relatively small across different muscle groups. In addition, studies of upper limb structures, such as the supraspinatus tendon, have shown that experienced raters can achieve high measurement precision (Ahmad et al., 2017), supporting the overall technical consistency of MSK-US.

4.3. Clinical implications

From a clinical perspective, these findings have practical implications. A novice PT may be able to reliably monitor changes in RF

Table 2
Intra-rater reliability of muscle thickness measurements.

Rater		$ICC_{2,1}$ (95% CI)	CV (%)	SEM (mm)			MDC_{95} (mm)		
				Individual	Group	Total	Individual	Group	Total
Expert	A	0.94 (0.84; 0.98)	2.14	0.52	0.47	0.48	1.45	1.31	1.34
	B	0.92 (0.81; 0.98)	1.74	0.42			1.17		
Novice	C	0.92 (0.80; 0.98)	1.68	0.41	0.49		1.13	1.37	
	D	0.89 (0.73: 0.97)	2.33	0.58			1.61		

ICC=Interclass Correlation Coefficient; CI=Confidence Interval; CV=Coefficient of Variation; SEM=Standard Error of Measurement; MDC_{95} = Minimal Detectable Change; mm = millimeters.

Table 3
Inter-rater reliability of muscle thickness measurements.

Group	$ICC_{2,1}$ (95% CI)	$ICC_{2,3}$ (95% CI)	CV (%)	Bias (95% CI) (mm)	Limits of Agreement (mm)		
					Lower	Upper	(95% CI)
Total	0.69 (0.55; 0.81)	0.90 (0.83; 0.95)	3.56				
Experts (A + B)	0.71 (0.49; 0.84)	0.83 (0.66; 0.92)	3.68	−0.04 (−1.28; 1.19)	−3.13	3.05	(-5.52; 5.43)
Novices (C + D)	0.60 (0.33; 0.78)	0.75 (0.50; 0.88)	3.68	−0.58 (−1.61; 0.45)	−3.25	2.09	(-5.20; 4.04)

ICC=Interclass Correlation Coefficient; CI=Confidence Interval; CV=Coefficient of Variation; mm = millimeters.

thickness when performing repeated assessments on the same patient. In addition, the comparable LoA between groups suggest that, under standardized conditions, measurements obtained by different clinicians may be used with a similar level of confidence. However, despite the overall limited measurement error, some degree of variability remains between raters.

For this reason, longitudinal monitoring is most consistent when assessments are performed by the same operator. In clinical settings where, due to logistical or organizational factors, multiple clinicians must share patient monitoring, our results indicate that inter-rater reliability is significantly improved by averaging three separate measurements rather than relying on a single acquisition. This approach minimizes individual measurement error and provides a more stable foundation for clinical decision-making when inter-operator consistency is required.

The clinical utility of US ultimately depends on its ability to detect true physiological changes relative to measurement error. In this context, our study provides two clinically relevant reference values for RF thickness assessment. First, the pooled SEM of 0.48 mm indicates that measurement error is typically below 0.5 mm during repeated assessments, reflecting a high level of precision. Second, this corresponds to a pooled MDC_{95} of 1.34 mm, which represents the minimum change required to be confident that a true change in MT has occurred. This value establishes a practical benchmark that allows clinicians to isolate genuine physiological tissue changes from background noise. This threshold may assist clinicians in interpreting longitudinal changes, whether monitoring muscle atrophy following surgery or hypertrophy during rehabilitation.

4.4. Strengths and limitations

This study has several methodological strengths. First, the inclusion of both expert and novice physiotherapists provides significant insights into the learning curve and practical accessibility of MSK-US. Second, we minimized potential bias through a rigorous data collection process: both the ultrasound image acquisition and the subsequent image analysis were performed in a fully blinded manner and followed a randomized sequence to prevent any order effects. Finally, we employed a comprehensive statistical framework; by reporting relative reliability (ICC), relative variability (CV%), alongside absolute measurement error estimates, such as the SEM, MDC_{95}, systematic bias, and LoA, we offer a transparent look at the consistency and precision of these measurements, which is essential for any clinician tracking physiological changes over time.

Despite these strengths, several limitations should be considered. First, the sample size was relatively small, which may limit the statistical power and the precision of the estimates. Second, the use of young, healthy, and physically active participants likely provided favorable imaging conditions with clearly defined fascial boundaries. Consequently, these findings may not be directly generalizable to older adults, post-operative patients, sarcopenic individuals, or other clinical populations where altered tissue characteristics could increase measurement complexity. Third, the study assessed intra-day reliability only; therefore, the findings do not account for potential variability across different days. Finally, while this study evaluated image acquisition reliability across different operators, the image measurements themselves were performed by a single blinded investigator, which did not allow for the assessment of image measurement reliability.

5. Conclusions

This study provides preliminary evidence supporting the integration of RUSI into physiotherapy practice as an objective tool for muscle assessment. After standardized training, novice raters may achieve acceptable reliability under standardized conditions, supporting the technique's accessibility. Although this study focused on intra-day assessments, these findings provide preliminary evidence to guide measurement protocols. Despite the overall limited measurement error and acceptable agreement, some degree of variability between raters persists. For this reason, although different operators may obtain similar measurements, longitudinal monitoring is likely to be more consistent when assessments are performed by the same clinician to minimize variability. However, in settings where multiple operators must alternate, averaging three separate measurements significantly improves agreement and provides a more reliable basis for tracking physiological changes over time.

CRediT authorship contribution statement

Daniele Conte: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft. **Giulia Costantini:** Data curation, Formal analysis, Investigation, Visualization, Writing – review & editing. **Chiara Pinzini:** Conceptualization, Investigation, Methodology, Writing – review & editing. **Giuliana De Maio:** Conceptualization, Investigation, Methodology, Writing – review & editing. **Maria Parpinel:** Conceptualization, Supervision, Writing – review & editing. **Daniele Vidi:** Conceptualization, Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing.

Patient consent

This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained by the Institutional Review Board of the Department of Medicine at the University of Udine (Reference: Prot IRB 259/2025, Tit III, cl 13, fasc.24/2025). All participants provided written informed consent prior to their inclusion in the study.

Ethics

This study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained by the Institutional Review Board of the Department of Medicine at the University of Udine (Reference: Prot IRB 259/2025, Tit III, cl 13, fasc.24/2025). All participants provided written informed consent prior to their inclusion in the study.

Data availability statement

Data will be made available upon reasonable request to the corresponding author.

Funding

D.C. received a fellowship funded by the Autonomous Region of Friuli Venezia Giulia (Italy) and co-financed by the European Social Fund Plus (ESF+) through the Regional Programme 2021/2027 (Specific Programme No. 22/23, Line A; CUP G23C23001130008).

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbmt.2026.08.020.

References

Ahmad, A., Bandpei, M.A.M., Gilani, S.A., Munawar, A., Ahmed, I., Tanveer, F., 2017. Reliability of musculoskeletal ultrasound imaging to measure supraspinatus tendon thickness in healthy subjects. J. Phys. Ther. Sci. 29, 1394–1398. https://doi.org/10.1589/jpts.29.1394.

Caldwell, A.R., 2022. SimplyAgree: an R package and jamovi module for simplifying agreement and reliability analyses. J. Open Source Softw. 7, 4148. https://doi.org/10.21105/joss.04148.

Chaput, M., Palimenio, M., Farmer, B., Katsavelis, D., Bagwell, J.J., Turman, K.A., Wichman, C., Grindstaff, T.L., 2021. Quadriceps strength influences patient function more than single leg forward hop during late-stage ACL rehabilitation. Int. J. Sports Phys. Ther. 16, 145–155. https://doi.org/10.26603/001c.18709.

Chen, G., Taylor, P.A., Haller, S.P., Kircanski, K., Stoddard, J., Pine, D.S., Leibenluft, E., Brotman, M.A., Cox, R.W., 2018. Intraclass correlation: improved modeling approaches and applications for neuroimaging. Hum. Brain Mapp. 39, 1187–1206. https://doi.org/10.1002/hbm.23909.

Donner, A., Zou, G.Y., 2012. Closed-form confidence intervals for functions of the normal mean and standard deviation. Stat. Methods Med. Res. 21, 347–359. https://doi.org/10.1177/0962280210383082.

Franchi, M.V., Longo, S., Mallinson, J., Quinlan, J.I., Taylor, T., Greenhaff, P.L., Narici, M.V., 2018. Muscle thickness correlates to muscle cross-sectional area in the assessment of strength training-induced hypertrophy. Scand. J. Med. Sci. Sports 28, 846–853. https://doi.org/10.1111/sms.12961.

Hegde, D.D., Hadya Ananda, K., Vavachan, N., 2025. Effectiveness of quadriceps strength training in adults with knee osteoarthritis: a systematized review. Muscoskel. Care 23, e70134. https://doi.org/10.1002/msc.70134.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropract. Med. 15, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M., Streiner, D.L., 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J. Clin. Epidemiol. 64, 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002.

Mokkink, L.B., Eekhout, I., Boers, M., van der Vleuten, C.P.M., de Vet, H.C.W., 2023. Studies on reliability and measurement error of measurements in medicine – from design to statistics explained for medical researchers. Patient Relat. Outcome Meas. 14, 193–212. https://doi.org/10.2147/PROM.S398886.

Munro, B.H., 2005. Statistical Methods for Health Care Research, fifth ed. Lippincott Williams & Wilkins, Philadelphia, Pa London.

Pardo, E., El Behi, H., Boizeau, P., Verdonk, F., Alberti, C., Lescot, T., 2018. Reliability of ultrasound measurements of quadriceps muscle thickness in critically ill patients. BMC Anesthesiol. 18, 205. https://doi.org/10.1186/s12871-018-0647-9.

Parry, S.M., El-Ansary, D., Cartwright, M.S., Sarwal, A., Berney, S., Koopman, R., Annoni, R., Puthucheary, Z., Gordon, I.R., Morris, P.E., Denehy, L., 2015. Ultrasonography in the intensive care setting can be used to detect changes in the quality and quantity of muscle and is related to muscle strength and function. J. Crit. Care 30. https://doi.org/10.1016/j.jcrc.2015.05.024, 1151.e9–14.

Pinto-Ramos, J., Costa-Santos, C., Costa, F., Tavares, H., Cabral, J., Moreira, T., Brito, R., Barroso, J., Sousa-Pinto, B., 2022. Reliability of point-of-care ultrasound for measuring quadriceps femoris muscle thickness. Eur. J. Phys. Rehabil. Med. 58, 767–773. https://doi.org/10.23736/S1973-9087.22.07432-9.

Puthucheary, Z.A., Rawal, J., McPhail, M., Connolly, B., Ratnayake, G., Chan, P., Hopkinson, N.S., Phadke, R., Dew, T., Sidhu, P.S., Velloso, C., Seymour, J., Agley, C. C., Selby, A., Limb, M., Edwards, L.M., Smith, K., Rowlerson, A., Rennie, M.J., Moxham, J., Harridge, S.D.R., Hart, N., Montgomery, H.E., 2013. Acute skeletal muscle wasting in critical illness. JAMA 310, 1591–1600. https://doi.org/10.1001/jama.2013.278481.

Schneider, C.A., Rasband, W.S., Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671–675. https://doi.org/10.1038/nmeth.2089.

Şekir, U., Yalaki, U.C., Akova, B., 2022. Rectus femoris muscle thickness and cross-sectional area on ultrasonography may predict isometric and isokinetic knee extension strength: a cross-sectional study. Spor Hekimliği Dergisi 57, 21–30. https://doi.org/10.47447/tjsm.0585.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428. https://doi.org/10.1037//0033-2909.86.2.420.

Stausholm, M.B., da Silva, K.R., Inácio, P.A., de Sá Filho, A.S., Lopes-Martins, P.S.L., Bjordal, J.M., Leonardo, P.S., Lopes-Martins, R.A.B., 2024. Reliability of ultrasound assessment of the rectus femoris muscle thickness: intra-rater, inter-rater, and inter-day analysis accounting for age and sex. BMC Muscoskelet. Disord. 25, 916. https://doi.org/10.1186/s12891-024-08033-5.

Takahashi, Y., Fujino, Y., Miura, K., Toida, A., Matsuda, T., Makita, S., 2021. Intra- and inter-rater reliability of rectus femoris muscle thickness measured using ultrasonography in healthy individuals. Ultrasound J 13, 21. https://doi.org/10.1186/s13089-021-00224-8.

Valera-Calero, J.A., Sánchez-Jorge, S., Álvarez-González, J., Ortega-Santiago, R., Cleland, J.A., Fernández-de-Las-Peñas, C., Arias-Buría, J.L., 2020. Intra-rater and inter-rater reliability of rehabilitative ultrasound imaging of cervical multifidus muscle in healthy people: imaging capturing and imaging calculation. Musculoskelet Sci Pract 48, 102158. https://doi.org/10.1016/j.msksp.2020.102158.

Walter, S.D., Eliasziw, M., Donner, A., 1998. Sample size and optimal designs for reliability studies. Stat. Med. 17, 101–110. https://doi.org/10.1002/(sici)1097-0258(19980115)17:1%253C101::aid-sim727%253E3.0.co;2-e.

Whittaker, J.L., Ellis, R., Hodges, P.W., Osullivan, C., Hides, J., Fernandez-Carnero, S., Arias-Buria, J.L., Teyhen, D.S., Stokes, M.J., 2019. Imaging with ultrasound in physical therapy: what is the PT's scope of practice? A competency-based educational model and training recommendations. Br. J. Sports Med. 53, 1447–1453. https://doi.org/10.1136/bjsports-2018-100193.