Original Article

Quantitative ultrasonography of the foot muscles: a comprehensive perspective on reliability

Nicolas Haelewijn[1#^], Jean-Louis Peters-Dickie[1,2#^], Roel de Ridder[3^], Kevin Deschamps[1,4^], Christine Detrembleur[2^], Sébastien Lobet[2,5^], Valentien Spanhove[3^]

[1]Department of Rehabilitation Sciences, Musculoskeletal Rehabilitation Research Group, KU Leuven, Brugge, Belgium; [2]Neuromusculoskeletal Lab (NMSK), Secteur des Sciences de la Santé, Institut de Recherche Expérimentale et Clinique, Université Catholique de Louvain, Brussels, Belgium; [3]Department of Rehabilitation Sciences, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium; [4]Division of Podiatry, Haute Ecole Leonard De Vinci, Brussels, Belgium; [5]Haemostasis and Thrombosis Unit, Division of Hematology, Cliniques universitaires Saint-Luc, Université catholique de Louvain (UCLouvain), Brussels, Belgium

Contributions: (I) Conception and design: All authors; (II) Administrative support: R de Ridder, K Deschamps, N Haelewijn, V Spanhove; (III) Provision of study materials or patients: R de Ridder, K Deschamps; (IV) Collection and assembly of data: N Haelewijn, JL Peters-Dickie, V Spanhove; (V) Data analysis and interpretation: N Haelewijn, JL Peters-Dickie, V Spanhove, K Deschamps, R de Ridder; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Prof. Kevin Deschamps, PhD. Department of Rehabilitation Sciences, Musculoskeletal Rehabilitation Research Group, KU Leuven, Spoorwegstraat 12, 8200 Brugge, Belgium; Division of Podiatry, Haute Ecole Leonard De Vinci, 1200 Brussels, Belgium. Email: kevin.deschamps@kuleuven.be.

Background: Quantitative ultrasound imaging is a popular technique to assess the structural properties of the intrinsic and extrinsic foot muscles. Although several studies examined test-retest reliability, specific gaps remain in assessing inter-rater reliability, particularly distinguishing between image acquisition and muscle measurement. Additionally, these studies utilized equipment that may not be generalizable across both clinical and research settings and often involved small sample sizes without prior sample size calculations. This study aimed to investigate test-retest reliability as well as global and measurement-based inter-rater reliability (MIRR) using a low-end ultrasound device to measure intrinsic and extrinsic foot muscle sizes.

Methods: This prospective reliability study included 21 active individuals. Five intrinsic muscles [abductor hallucis (AbH), flexor digitorum brevis (FDB), flexor hallucis brevis (FHB), quadratus plantae (QP), abductor digiti minimi (AbDM)], and three extrinsic muscles [peroneal (PER), flexor digitorum longus, tibialis anterior (TA)] were scanned. Three investigators independently acquired images on two occasions and measured cross-sectional area (CSA) and thickness in September and October 2023. Participants were assessed either at the Musculoskeletal Research Group laboratory (University of Leuven, Bruges) or in the Rehabilitation Sciences laboratory (Ghent University hospital). Test-retest (same investigator, one week in between), global inter-rater (each investigator measures own image set) and MIRR (three investigators measure one image set) was performed following intra-class correlation, standard error of the measurement (SEM) and coefficient of variation.

Results: Test-retest reliability showed intraclass-correlation coefficients of 0.60–0.88 for the FDB and 0.38–0.73 for the TA. SEM ranged from 0.16 to 0.41 cm^2 (CSA) and from 0.05 to 0.31 cm (thickness) for the intrinsic, while they ranged from 0.19 to 1.13 cm^2 and from 0.12 to 0.44 cm for the extrinsic muscles. Global

^ ORCID: Nicolas Haelewijn, 0000-0002-7510-7844; Jean-Louis Peters-Dickie, 0000-0002-5922-2711; Roel de Ridder, 0000-0001-9723-103X; Kevin Deschamps, 0000-0001-6559-6237; Christine Detrembleur, 0000-0003-0776-3820; Sébastien Lobet, 0000-0002-3829-6850; Valentien Spanhove, 0000-0001-8996-4583.

inter-rater correlation coefficients varied between 0.4 and 0.8 for the AbH and FDB. Measurement based inter-rater correlation coefficient varied between 0.50 and 0.96 for AbH, FDB, TA and PER muscles. SEM ranged from 0.14 to 0.89 cm^2 (CSA) and from 0.07 to 0.24 cm (thickness) for the intrinsic, while they ranged from 0.29 to 0.85 cm^2 (CSA) and from 0.12 to 0.51 cm (thickness) for the extrinsic muscles. Coefficients of variations were between 4% and 34%. For test-retest, they were consistently ≤10% for AbH thickness, FDB CSA, FHB and TA. FDB coefficients of variation were ≤10% across all inter-rater reliabilities.

Conclusions: Most muscles demonstrated moderate to excellent test-retest reliability using a portable ultrasound device, supporting its generalizability. However, the greater variability in global inter-rater reliability suggests substantial variation during image acquisition. The same clinician should perform pre-intervention and follow-up assessments to minimize errors. If different clinicians are involved, caution is needed when comparing measurements.

Keywords: Foot muscles; foot core; abductor hallucis (AbH); ultrasound; reliability

Submitted Jul 27, 2024. Accepted for publication Nov 19, 2024. Published online Dec 30, 2024.
doi: 10.21037/qims-24-1309
View this article at: https://dx.doi.org/10.21037/qims-24-1309

Introduction

With each foot strike in locomotion, elastic energy is stored and released through deformation of the medial foot arch (1), regulated by the intrinsic (IFM) and extrinsic (EFM) foot muscles (2,3). The IFM are small muscles with short moment arms that act as local foot stabilizers, helping to maintain foot posture and control foot deformation (4). Additionally, they work alongside the plantar aponeurosis to regulate foot stiffness during running (5). The EFM, on the other hand, originate from the lower leg and attach to the foot via long tendons, making them global foot movers (4).

IFM and EFM size and strength can be decreased in various disorders such as hallux valgus and diabetes (6,7), and in older people (8). Therefore, researchers and health care professionals aim to quantitatively evaluate IFM and EFM strength. However, isolating the contributions of each muscle in global toe and ankle strength is impossible with direct assessments such as dynamometry (9,10). Consequently, imaging techniques are typically employed as indirect means of quantitative assessment, with magnetic resonance imaging as the golden standard (11,12). However, it is costly and generally poorly accessible. Alternatively, two-dimensional ultrasound imaging (USI) is frequently used for assessing the structural properties of both IFM and EFM (13-16). The measurements obtained with USI are highly similar to those obtained with magnetic resonance imaging, with correlations superior to 0.9 (17).

Several studies showed good reliability when quantitatively evaluating cross-sectional area (CSA) and dorso-plantar thickness of various IFM (11,13,17,18) and EFM (18-25). Nonetheless, some aspects need to be considered in quantitative ultrasound: Firstly, USI reliability is susceptible to variation in how the image is acquired and then how the measurement is performed. This applies to conditions where there is a repeated measure design including individual therapists separately; however, the situation becomes even more complex when USI evaluations are performed across different observers. Therefore, inter-rater reliability can be fragmented into two aspects: the measurement-based inter-rater reliability (MIRR) and the global inter-rater reliability (GIRR). The MIRR emphasizes the reliability of the image measurement process itself, regardless of who acquired the images. The MIRR has, to our knowledge, only been studied for the tibialis anterior (TA) (20). The GIRR replicates clinical contexts, where health care professionals independently acquire and measure images. To our knowledge, GIRR has been reported for the abductor hallucis (AbH), flexor digitorum brevis (FDB), quadratus plantae (QP), TA and peroneal (PER) muscles (23,26-28), but is lacking regarding the abductor digiti minimi (AbDM), flexor hallucis brevis (FHB) and flexor digitorum longus (FDL). Secondly, it is important to note that the majority of these studies utilized high-end USI devices (12,23,26), a factor that limits the generalizability of their findings to lower-end devices. The cost of USI equipment has been stated to directly relate to the attained images resolution and quality (29). This therefore indicates that a high-resolution ultrasonography machine produces higher quality images, which are more easily interpreted, especially on small

structures, such as the IFM. Moreover, time constraints during image acquisition are rarely discussed in literature. Nevertheless, this represents a significant challenge which clinicians daily face, that could affect the reliability of USI.

Thirdly, it is worth highlighting that only a single study mentioned an a priori sample size calculation (30), and that some of the aforementioned studies adopted small sample sizes of approximately 10 participants, raising concerns about the generalizability of their results (19,22-25). As last consideration, intra-class correlation coefficients (ICC) and standard error of the measurement (SEM) are common reliability metrics. However, there are several ways to calculate ICC and SEM, depending on the study design of the reliability study and the future use of the device in actual practice (31). Some of the published USI reliability studies used inappropriate calculations (11,17,19,20,22-24,27,28,30,32), which can lead to overestimations (31).

While USI is a promising tool for assessing IFM and EFM size, attention must be paid to study design, equipment quality and sample size to ensure reliable and generalizable results. Therefore, the objectives of this reliability study were: to determine the test-retest reliability of IFM and EFM measurements using a low-end USI device, and to report the inter-rater reliability spanning from MIRR and GIRR. We present this article in accordance with the GRRAS reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1309/rc) (33).

Methods

Participants

The sample size was estimated based on the lookup table published by Borg [2022] (34). AbH and PER CSA were the primary quantitative outcomes of this study, because they are respectively the biggest foot and lower leg muscles assessed in the frame of this study. Based on previous ICC for these muscles ranging between 0.90 and 0.99, an ICC confidence interval width of 0.15 (19,26,30,35), a significance level of 5% and a power of 80% and 10% drop out, a sample size of 21 participants was deemed sufficient.

A convenience sample (non-probabilistic method) of 21 asymptomatic recreationally active adults was recruited at our two university campuses. "Recreationally active" was defined as the practice of at least 20 minutes of physical activity a day, at least three times a week. Exclusion criteria were foot deformities, neurological, musculoskeletal, or systemic diseases, pregnancy, lower limb injury or surgery in the 6 months prior participation, and if participants reported recent ankle or foot pain.

Each participant signed the informed consent form. This study was approved by the Ethical Committees of UZ/KU Leuven (No. S67722) and UZ Ghent (No. B6702023000319). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Measures and procedures

In this study, each participant was assessed twice either in Bruges at the Musculoskeletal Research Group laboratory (University of Leuven) or in the university hospital of Ghent at the Rehabilitation Sciences laboratory (Ghent University), exactly 7 days apart, respectively session 1 and session 2 (*Figure 1*). During session 1, demographic data were collected including age, weight, height, sex, shoe size and dominant foot. Participants completed the Foot and Ankle Outcome Score (FAOS) and Baecke questionnaire. The FAOS is a self-reported questionnaire assessing the functional status related to ankle and foot conditions (36,37). Each subscale is scored from 0 to 100, with 0 indicating extreme problems and 100 indicating no problems. The Baecke questionnaire assesses physical activity levels in individuals. This questionnaire gathers information on various domains of physical activity, including work-related activities, sports, and leisure-time activities (38,39). Furthermore, the six items of the Foot-Posture-Index score were determined by agreement between the investigators (40), and navicular drop was calculated between sit to bipedal stance by one investigator (N.H.) (41). This allowed to describe foot posture as this may have an effect on foot muscle sizes (42).

During the two sessions, images of IFM and EFM of the dominant foot were collected using a low-end portable USI system with a 5–12 MHz 40 mm broadband linear array probe (MicrUs EXT-1H, Telemed UAB, Lithuania). Three musculoskeletal physiotherapists (N.H., J.L.P.D. and V.S.) independently acquired muscle images with a time constraint of 20 to 30 minutes during the months of September and October 2023. The three investigators had varying USI and clinical experiences: V.S. (30 years, MSc, PhD, 3-year clinical experience) had 3 years of scanning experience while N.H. (25 years, MSc, 1-year clinical experience) and J.L.P.D. (26 years, MSc, 1-year clinical experience) had both 1-year scanning experience. The three investigators conducted several pilot testing sessions together to ensure consistency in the protocol (*Figure 1*).

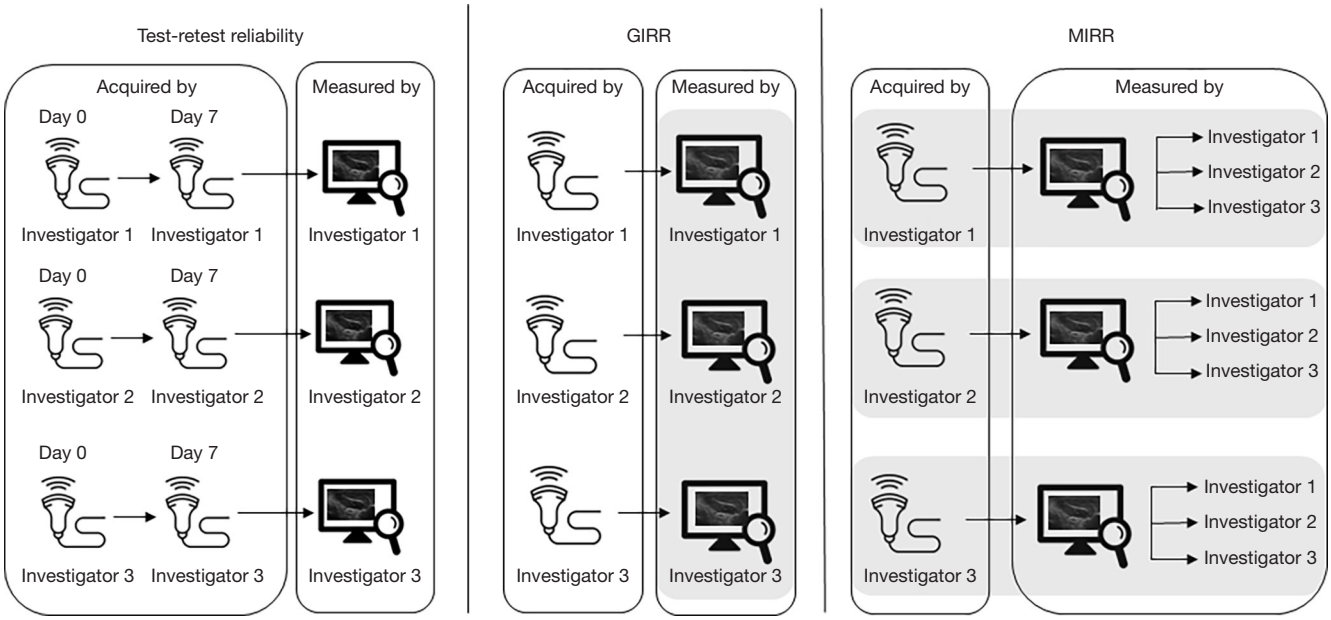

Figure 1 Study set up depicting the acquisition and measurement processes to obtain the three reliability types. Test-retest reliability was calculated for each investigator between day 0 and day 7. The grey zone indicates the measurements used to calculate GIRR and MIRR: GIRR compares the measurements done the investigators on their own images and MIRR compares for each of the three image sets the measurements of the three investigators. GIRR, global inter-rater reliability; MIRR, measurement-based inter-rater reliability.

The scanning protocol was adapted from previous studies and conducted with the participants lying on their back (i.e., relaxed and non-weight bearing position) (19,22,43). A summary of probe positions and sample images is available in *Figure 2*. Each investigator independently acquired transverse (short-axis) and longitudinal (long-axis) images of five IFM (AbH, FDB, FHB, QP, AbDM) and three EFM (PER, FDL and TA). Transverse imaging of the FHB was not possible due to the indefinite appearance of this muscle in the frontal plane. Longitudinal imaging of the TA was not possible as the muscle did not consistently fit within the USI display. The depth, focal point, frequency, and gain were adjusted for each participant and each muscle to optimize quality of image capturing. In case of doubt, subjects were instructed to perform specific movements to elicit muscular contraction, aiding the investigator in identifying muscle borders. These contractions were brief and performed without resistance to minimize the risk of fluid shifts significantly affecting muscle size. For each muscle, its relaxed state was captured in a cine-loop lasting up to 5 seconds. The investigator then saved a single image, before scanning the next muscle view. Muscle size measurement was performed in the days following image acquisition to avoid recall bias. During this step, investigators were blinded to participant information and to the measurements taken by other investigators. For the GIRR, each investigator individually measured muscle sizes using their own set of acquired images. For the MIRR, muscle measurement was performed by the three investigators on the image set acquired by one investigator. The image sets were taken from the second session and were measured using still images in Image J software (National Institute for Health, Bethesda, MD, USA). Previous studies used this software to measure muscle CSA and thickness from ultrasound images and demonstrated excellent reliability (19,22,44). Transverse-view and longitudinal-view images were respectively used to obtain CSA (cm^2) and thickness (cm). Thickness was averaged from three measurements at the thickest part of the muscle on the image.

Statistical analysis

Statistical analyses were performed in SPSS for Windows v27 (IBM SPSS Inc. Chicago, IL, USA). Means and standard deviations were reported for age, height, weight, shoe size, navicular drop, FAOS, Baecke questionnaire and muscle measurements. Foot posture index scores were interpreted

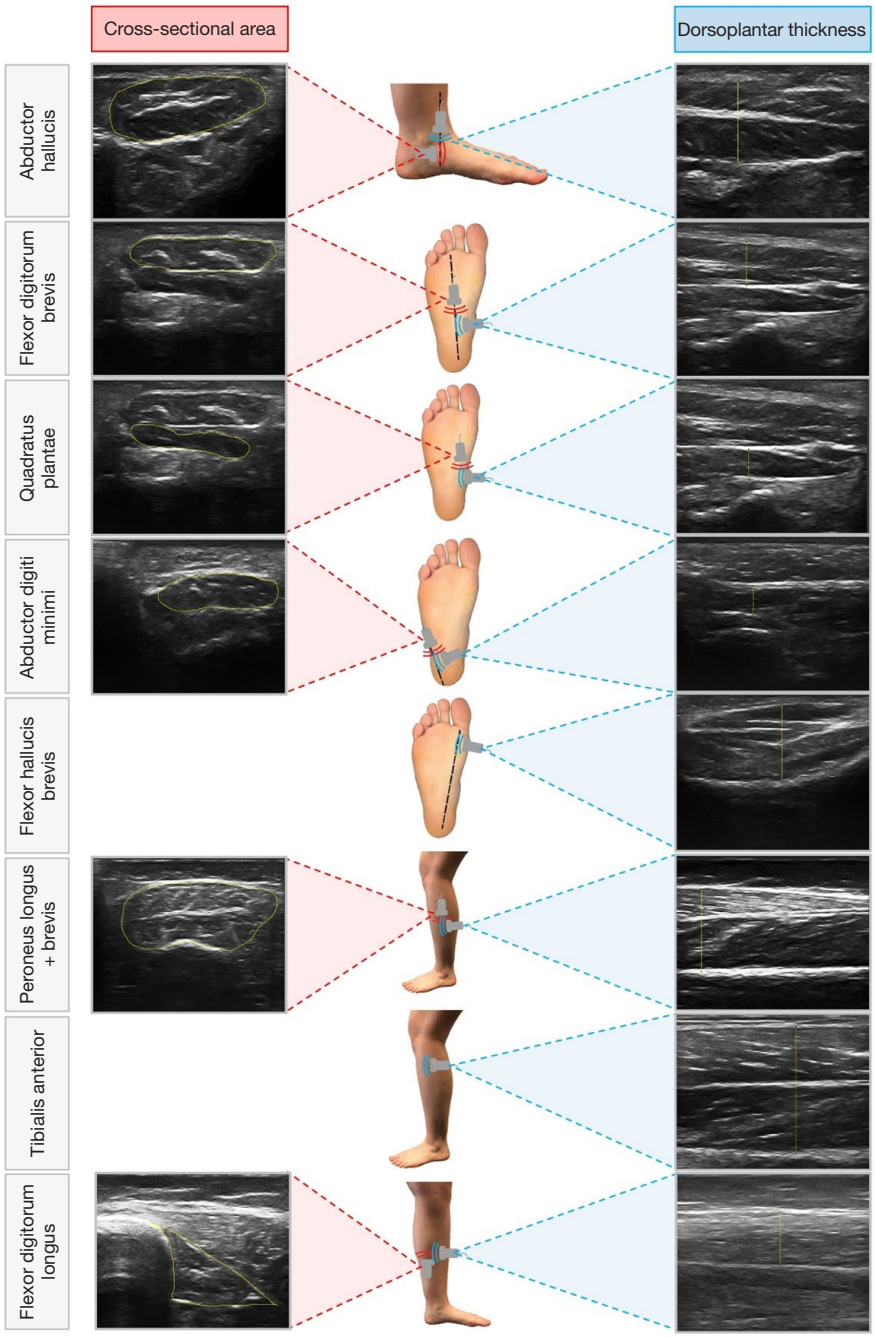

Figure 2 Summary of probe positions with sample images. Probe position during acquisition of transverse (red) and longitudinal (blue) images of the intrinsic and extrinsic foot muscles. The following landmarks were used during the acquisition: abductor hallucis was scanned on a vertical line passing in front of the medial malleolus; FDB was scanned on a line passing through the third toe and the medial tubercle of the calcaneus; QP was scanned more medial than the scanning zone of FDB and near the talo-calcaneo-navicular joint; abductor digiti minimi was scanned between lateral tuberosity of the calcaneus and tuberosity of the 5th metatarsal; flexor hallucis brevis was scanned on a line parallel to the muscle orientation, i.e., medio-plantar face of the first metatarsal bone; peroneus brevis and longus were scanned together at the mid-distance between the fibular head and the lateral malleolus; tibialis anterior was scanned at 20% of the distance between the fibular head and the two malleolus; flexor digitorum longus was scanned at mid-distance between the medial tibial plateau and the medial malleolus. FDB, flexor digitorum brevis; QP, quadratus plantae.

Table 1 Participant characteristics (n=21)

Variable	Values
Age (years)	25.9±2.47
Sex (male/female)	10/11
Body weight (kg)	70.68±12.29
Height (m)	1.76±0.09
Shoe size (European)	40.2±3.08
Dominant foot, right	18 (85.7)
Foot Posture Index (–12 to 12 score)	4±2.5
Neutral	11 (52.4)
Pronated	5 (23.8)
Extremely pronated	0
Supinated	5 (23.8)
Extremely supinated	0
Navicular drop (mm)	5.2±1.1
Baecke questionnaire	10.02±1.72
Foot and Ankle Outcome Score	96.16±3.35

Data are presented as mean ± standard deviation or number of participants (%).

as neutral (between 0 and 5), pronated (higher than 5), extremely pronated (higher than 9), supinated (lower than 0) and extremely supinated (lower than –4) (40,45).

ICC and corresponding 95% confidence intervals were calculated. According to McGraw and Wong convention, test-retest reliability was calculated as single measure, absolute agreement, two-way mixed model while GIRR and MIRR were calculated using single measure, absolute agreement, two-way random effects model. These use the same formula and referred to as $ICC_{2,1}$ in Shrout and Fleiss convention (Eq. [1]), but the interpretation of their results is different (31):

$$\frac{MSR - MSE}{MSR(k-1)MSE + \frac{k}{n}(MSC - MSE)} \quad [1]$$

Where MSR is the mean square rows, MSE is the mean square error, MSC is the mean square columns, k is the number of raters (=3) and n the number of participants (=21). The scanning resulted in three sets of images (one per investigator). Test-retest reliability was considered between the two sessions held one week apart. For inter-rater, only the images of the second session were used. GIRR was calculated with each investigator measuring muscle sizes on their own images and MIRR was obtained with the three investigators interpreting muscle sizes on USI images on one image set. Therefore, three test-retest and three MIRR were obtained for each muscle. ICC were interpreted as follow: <0.5 poor, ≥0.5 to <0.75 moderate, ≥0.75 to <0.9 good, ≥0.9 excellent (31). SEM, agreement ($SEM_{agreement}$) were then calculated (Eq. [2]) (46):

$$SEM_{agreement} = \sqrt{\left(\sigma^2_{Ip} + \sigma^2_{residual}\right)} \quad [2]$$

Where σ^2 is the variance, "I_P" the interaction between the investigators and the participants and "residual" refers to the error.

The coefficient of variation (CV) was calculated for the three reliability types. The CV of the "*i*"th triplet of measurements (i.e., the measurements of the three investigators) is given by (Eq. [3]) (47):

$$CV = \frac{\sum CV_i}{n} \times 100\%, \text{ where } CV_i = \frac{SD_i}{\bar{x}_i} \quad [3]$$

Where n is the number of participants (=21), CV the coefficient of variation, $\bar{x}$ the mean of the measurements, SD the standard deviation and i refers to one "*i*"th triplet of measurements.

Results

The 21 participants were scanned for the five IFM and three EFM by the three investigators on two occasions, as per protocol. This included 10 participants recruited from the Bruges campus of KU Leuven University and 11 participants from the University Hospital of Ghent.

Descriptive characteristics are available in *Table 1*. Participants were young adults, with an even distribution between males and females. Five participants had a pronated foot (23.8%), five participants (23.8%) had a supinated foot. The remaining eleven participants had a neutral foot (52.4%). All were physically active and pain free, as confirmed by the Baecke and FAOS scores. No significant differences in participant demographics were observed between the two scanning sites, ensuring consistency in the sample.

Test-retest reliability

CSA

Test-retest reliability (*Table 2*) between sessions 1 and 2 was moderate to good for the three investigators for AbH (ICC,

Table 2 Test-retest reliability with mean muscle size

Muscle type	Name	Measure	Investigator 1				Investigator 2				Investigator 3			
			Size, mean (SD)	ICC (95% CI)	CV (%)	SEM	Size, mean (SD)	ICC (95% CI)	CV (%)	SEM	Size, mean (SD)	ICC (95% CI)	CV (%)	SEM
Intrinsic foot muscles	AbH	CSA, cm^2	2.4 (0.69)	0.78 (0.53–0.9)	10	0.29	2.5 (0.69)	0.86 (0.42–0.95)	10	0.27	2.65 (0.73)	0.53 (0.15–0.78)	12	0.41
		Thickness, cm	1.08 (0.2)	0.68 (0.36–0.86)	9	0.11	1.17 (0.22)	0.78 (0.54–0.9)	6	0.11	1.22 (0.18)	0.51 (0.1–0.77)	8	0.14
	FDB	CSA, cm^2	2.03 (0.47)	0.88 (0.74–0.95)	7	0.16	2.2 (0.5)	0.85 (0.67–0.94)	6	0.18	2.02 (0.42)	0.81 (0.58–0.92)	7	0.19
		Thickness, cm	1.04 (0.22)	0.6 (0.23–0.82)	11	0.15	1.02 (0.19)	0.8 (0.57–0.92)	6	0.09	1.04 (0.18)	0.79 (0.54–0.91)	7	0.09
	QP	CSA, cm^2	0.88 (0.26)	0.68 (0.37–0.86)	19	0.17	1.24 (0.55)	0.75 (0.48–0.89)	18	0.26	1.44 (0.33)	0.6 (0.23–0.82)	15	0.24
		Thickness, cm	0.6 (0.13)	0.87 (0.7–0.94)	7	0.05	0.69 (0.18)	0.51 (0.1–0.77)	16	0.13	0.83 (0.14)	0.74 (0.46–0.89)	7	0.07
	AbDM	CSA, cm^2	1.52 (0.46)	0.28 (0–0.6)	18	0.36	1.57 (0.34)	0.18 (0–0.56)	16	0.31	1.04 (0.22)	0.56 (0.19–0.79)	10	0.19
		Thickness, cm	0.7 (0.23)	0.66 (0.33–0.85)	15	0.13	0.9 (0.23)	0.29 (0–0.64)	14	0.17	0.92 (0.21)	0.82 (0.6–0.92)	9	0.09
	FHB	Thickness, cm	1.78 (0.35)	0.06 (0–0.47)	10	0.31	1.55 (0.22)	0.79 (0.54–0.91)	5	0.11	1.49 (0.21)	0.5 (0.1–0.76)	8	0.16
Extrinsic foot muscles	PER	CSA, cm^2	5.35 (1.4)	0.32 (0–0.66)	17	1.13	3.97 (1.01)	0.83 (0.64–0.93)	9	0.44	4.29 (1.05)	0.79 (0.53–0.91)	7	0.46
		Thickness, cm	2.45 (0.58)	0.21 (0–0.58)	14	0.44	1.77 (0.35)	0.55 (0.15–0.79)	9	0.22	2.02 (0.41)	0.79 (0.55–0.91)	7	0.17
	FDL	CSA, cm^2	1.75 (0.61)	0.18 (0–0.56)	20	0.46	1.76 (0.4)	0.58 (0.23–0.8)	9	0.25	1.55 (0.35)	0.71 (0.41–0.87)	10	0.19
		Thickness, cm	1.12 (0.36)	0.75 (0.48–0.89)	14	0.18	1.06 (0.17)	0.6 (0.23–0.81)	9	0.12	1.74 (0.23)	0.02 (0–0.45)	12	0.26
	TA	Thickness, cm	2.55 (0.47)	0.38 (0–0.69)	10	0.36	2.42 (0.32)	0.67 (0.35–0.85)	7	0.21	2.34 (0.37)	0.73 (0.45–0.88)	6	0.19

Mean and SD muscle sizes were calculated from the second day of scanning. Negative lower ends of ICC confidence intervals were adjusted to zero. SD, standard deviation; ICC, intraclass correlation coefficients; CI, confidence interval; CV, coefficient of variation; SEM, standard error of the measurement; AbH, abductor hallucis; FDB, flexor digitorum brevis; QP, quadratus plantae; AbDM, abductor digiti minimi; FHB, flexor hallucis brevis; PER, peroneal; FDL, flexor digitorum longus; TA, tibialis anterior; CSA, cross-sectional area.

0.53–0.86), FDB (ICC, 0.81–0.88), and QP (ICC, 0.6–0.75). Estimated ICC were good in two investigators for PER (0.79–0.83). Estimated ICC were poor to moderate in all three investigators for AbDM and FDL. CV were lower or equal to 10% for the three investigators for FDB, and the other measurements had CV lower than 20%. SEM of the IFM ranged from 0.16 to 0.36 cm^2 for rater 1, from 0.18 to 0.31 cm^2 for rater 2, and from 0.19 to 0.41 cm^2 for rater 3. SEM of the EFM ranged from 0.46 to 1.13 cm^2 for rater 1, from 0.25 to 0.44 cm^2 for rater 2, and from 0.19 to 0.46 cm^2 for rater 3.

Thickness

Test-retest reliability (*Table 2*) between sessions 1 and 2 was moderate to good in the three investigators for AbH (ICC, 0.51–0.78), FDB (ICC, 0.6–0.8), and QP (ICC, 0.51–0.87). Estimated ICC were reported to be poor in at least one investigator for AbDM, FHB, PER, and TA. CV were lower or equal to 10% for the three investigators for AbH, FHB and TA, and the other measurements had CV lower than 16%. SEM of the IFM ranged from 0.05 to 0.31 cm for rater 1, from 0.09 to 0.17 cm for rater 2, and from 0.07

Table 3 GIRR and MIRR

Muscle type	Name	Measure	GIRR			MIRR (image set investigator 1)			MIRR (image set investigator 2)			MIRR (image set investigator 3)		
			ICC (95% CI)	CV (%)	SEM	ICC (95% CI)	CV (%)	SEM	ICC (95% CI)	CV (%)	SEM	ICC (95% CI)	CV (%)	SEM
Intrinsic foot muscles	AbH	CSA, cm^2	0.74 (0.54–0.87)	12	0.37	0.88 (0.77–0.94)	8	0.24	0.96 (0.91–0.98)	5	0.14	0.80 (0.64–0.91)	10	0.89
		Thickness, cm	0.4 (0.14–0.65)	12	0.16	0.67 (0.44–0.83)	7	0.12	0.89 (0.79–0.95)	4	0.07	0.76 (0.58–0.88)	5	0.1
	FDB	CSA, cm^2	0.8 (0.61–0.91)	9	0.21	0.81 (0.66–0.91)	7	0.21	0.92 (0.82–0.96)	6	0.15	0.93 (0.87–0.97)	5	0.33
		Thickness, cm	0.69 (0.48–0.85)	9	0.3	0.69 (0.48–0.85)	8	0.13	0.67 (0.45–0.84)	6	0.1	0.49 (0.23–0.73)	8	0.14
	QP	CSA, cm^2	0.28 (0.03–0.56)	31	0.41	0.32 (0.06–0.6)	26	0.35	0.51 (0.22–0.75)	23	0.34	0.54 (0.28–0.76)	21	0.32
		Thickness, cm	0.17 (0–0.44)	22	0.47	0.51 (0.24–0.74)	15	0.14	0.32 (0.07–0.59)	16	0.13	0.03 (0–0.29)	18	0.18
	AbDM	CSA, cm^2	0.12 (0–0.36)	26	0.43	0.91 (0.76–0.96)	8	0.14	0.46 (0.19–0.7)	15	0.26	0.46 (0.2–0.7)	15	0.22
		Thickness, cm	0.3 (0.06–0.58)	22	0.58	0.61 (0.35–0.8)	15	0.14	0.50 (0.24–0.73)	16	0.2	0.53 (0.28–0.75)	16	0.17
	FHB	Thickness, cm	0.1 (0–0.37)	13	0.29	0.47 (0.21–0.71)	6	0.24	0.76 (0.55–0.89)	5	0.11	0.6 (0.36–0.8)	9	0.24
Extrinsic foot muscles	PER	CSA, cm^2	0.48 (0.11–0.75)	18	0.99	0.67 (0.32–0.85)	15	0.85	0.81 (0.62–0.92)	12	0.5	0.76 (0.41–0.9)	11	0.58
		Thickness, cm	0.37 (0.05–0.66)	18	0.45	0.65 (0.38–0.83)	10	0.33	0.50 (0.24–0.73)	8	0.51	0.82 (0.67–0.91)	7	0.2
	FDL	CSA, cm^2	0.32 (0.06–0.6)	20	0.39	0.35 (0.08–0.62)	20	0.5	0.24 (0–0.54)	18	0.47	0.59 (0.35–0.79)	13	0.29
		Thickness, cm	0 (0–0.12)	34	0.46	0.14 (0–0.41)	22	0.31	0.06 (0–0.27)	30	0.41	0.05 (0–0.27)	28	0.45
	TA	Thickness, cm	0.35 (0.09–0.62)	11	0.33	0.7 (0.5–0.85)	7	0.23	0.88 (0.77–0.95)	4	0.12	0.75 (0.51–0.89)	5	0.2

Mean and SD muscle sizes were calculated from the second day of scanning. Negative lower ends of ICC confidence intervals were adjusted to zero. GIRR, global inter-rater reliability; MIRR, measurement-based inter-rater reliability; ICC, intraclass correlation coefficients; CI, confidence interval; CV, coefficient of variation; SEM, standard error of the measurement; AbH, abductor hallucis; FDB, flexor digitorum brevis; QP, quadratus plantae; AbDM, abductor digiti minimi; FHB, flexor hallucis brevis; PER, peroneal; FDL, flexor digitorum longus; TA, tibialis anterior; CSA, cross-sectional area.

to 0.16 cm for rater 3. SEM of the EFM ranged from 0.18 to 0.44 cm for rater 1, from 0.12 to 0.22 cm for rater 2, and from 0.17 to 0.26 cm for rater 3.

Inter-rater reliability

CSA

GIRR (*Table 3*) was good for FDB (ICC =0.80) and moderate for AbH (ICC =0.74). Inter-rater reliability of the other muscles was poor. The CV were lower than 10% for FDB, and lower or equal to 20% for AbH, PER, and FDL. It reached 30% for QP. SEM of the IFM ranged from 0.21 to 0.43 cm^2, while SEM of EFM ranged from 0.39 to 0.99 cm^2.

MIRR (*Table 3*) was good to excellent for the three image sets for AbH (ICC, 0.80–0.96) and FDB (ICC, 0.81–0.93), and moderate to good for PER (ICC, 0.67–0.81). MIRR was poor to moderate in the three image sets for QP (ICC, 0.32–0.54) and FDL (ICC, 0.24–0.59). CV was lower or

equal to 10% for AbH and FDB. It was lower or equal to 20% for AbDM, PER and FDL. SEM of the IFM ranged from 0.14 to 0.89 cm^2, while SEM of EFM ranged from 0.29 to 0.85 cm^2.

Thickness

GIRR (*Table 3*) was moderate for FDB (ICC =0.69). It was poor for the other muscles. The CV was lower than 10% for FDB, and lower than 20% for AbH, FHB, PER, TA. It reached 30% for FDL. SEM of the IFM ranged from 0.16 to 0.58 cm, while SEM of EFM ranged from 0.29 to 0.45 cm.

MIRR (*Table 3*) was moderate to good for AbH (ICC, 0.67–0.89), PER (ICC, 0.50–0.82), and TA (ICC, 0.70–0.88). MIRR was poor to moderate for FDB (ICC, 0.49–0.69), QP (ICC, 0.03–0.51) and AbDM (ICC, 0.50–0.61). Poor reliability was found for FDL (ICC, 0.05–0.14). CV was lower than 10% for the three image sets for AbH, FDB, FHB, PER, and TA. It was lower than 20% for AbDM and QP. SEM of the IFM ranged from 0.12 to 0.18 cm, while SEM of EFM ranged from 0.11 to 0.33 cm.

Discussion

Quantitative ultrasound is frequently used in research and health-care contexts to measure the sizes of IFM and EFM, including in patients with hallux valgus (6,48), painful pes planus (49), chronic ankle instability (15), and diabetic neuropathy (7,50). This study aimed to determine its test-retest and inter-rater reliability in comparable settings using a low-end USI device. For the majority of the muscles, we demonstrated moderate to good test-retest reliability, poor to good GIRR, and poor to excellent MIRR.

Test-retest reliability was moderate to good for the three investigators for the CSA and thickness of AbH, FDB and QP in our study. For the other measures, at least two of the three investigators of our study had moderate or good ICC, except for AbDM CSA where two investigators had poor reliability. These findings thus differ from previous studies which consistently reported good to excellent test-retest reliability for IFM and EFM sizes (17,18,21,22,27,28,30), and moderate to excellent reliability for TA thickness (20,21). Still, at least two of the three investigators of our study had moderate or good ICC for a given structure, except for AbDM CSA where two investigators had poor reliability. To our knowledge, this is the first report of IFM or EFM test-retest reliability involving two or more raters who all performed both the image acquisition and the measurement. Indeed, in previous test-retest studies that involved 2 or 3 raters, the investigators did the image acquisition once and then performed the measurement twice on that image (20,26), or were limited to pathological populations (24). Selecting three raters is also in accordance with guidelines published by Koo & Li (31).

MIRR was good to excellent for the three image sets for the CSA of AbH and FDB, and at least moderate for the thickness of AbH, AbDM, TA, and PER CSA and thickness. This aligns with a previous study reporting moderate to excellent MIRR for TA thickness (ICC, 0.64–0.99) (20).

Moderate to good GIRR was found for AbH CSA, FDB CSA, and FDB thickness. Previous GIRR results have varied from moderate to excellent for IFM and EFM, with moderate GIRR reported for AbH, FDB, and QP thickness (23,26). As expected, GIRR was generally worse compared to MIRR. Differences between investigators become more pronounced when both acquisition and measurement are done individually, suggesting that scanning is more dependent on the investigator, while measurements in Image J are easier to standardize due to agreed guidelines and well-defined muscle borders. This standardization is facilitated by several factors: the controlled environment in which image evaluation is often performed minimizes external variability; software tools can automate measurement processes, ensuring consistent results; and defined protocols for evaluation provide clear steps that can be universally applied. The availability of reference images for comparison also aids standardization, as raters can align their measurements against established benchmarks. Additionally, training and calibration focused on evaluation techniques can help standardize the process across different raters, while consensus processes allow for collaborative assessment of images, further enhancing reliability. Clinicians should therefore be cautious when comparing muscle sizes acquired by different peers. Similarly to MIRR, previous GIRR results have been reported to vary from moderate to excellent for IFM and EFM. More particularly, Battaglia *et al.* reported moderate GIRR for the thicknesses of AbH, FDB and QP (26). Our study is the first to investigate GIRR for FHB, AbDM, and FDL, showing poor reliability which needs to be interpreted cautiously until confirmation by future studies using similar low-end USI devices.

While ICC is a relative estimate of reliability providing insights on individuals maintaining their position in a sample over repeated measurement, CV and SEM are absolute estimates of reliability which express the variety of repeated measurements for an individual (47). According to

Hopkins *et al.*, CV equal or lower than 10% are considered reasonable (51). Therefore, the CV we obtained align well with the moderate to good test-retest, MIRR and GIRR ICC of AbH and FDB. Although an acceptable CV was generally associated with a moderate or good ICC, this was not always true as for example test-retest ICC of FHB and TA were poor for one of the investigators with a CV equal to 10%. CV related to GIRR were overall not acceptable, whereas CV of MIRR reached 5% for many structures.

We also calculated $SEM_{agreement}$, an outcome quantifying the uncertainty associated with USI measurements. SEM values indicate how representative the measured size is of the true size, with smaller SEM indicating that the measured size is more likely to be representative of the true size, increasing confidence in clinical decisions. In this study, $SEM_{agreement}$ ranged from 0.12 to 0.89 cm^2 for CSA and from 0.05 to 0.58 cm for thickness of IFM. For EFM, SEM values ranged from 0.19 to 1.13 cm^2 for CSA and from 0.11 to 0.51 cm for thickness. These values are comparable to some previous studies for test-retest (27), and inter-rater reliability (24,27). Although others reported impressive SEM values, lower than a millimetre (17,20,24,25,30). Consistent with ICC results, SEM values were generally higher for GIRR than for MIRR, indicating greater variability in more clinically relevant settings.

Overall the majority of structures with moderate to excellent test-retest reliability highlights its generalizability, despite the use of a low-end USI device. However, the results obtained for the two perspectives on inter-rater reliability should raise the attention of people using USI in the intrinsic and extrinsic muscles of the foot. GIRR was indeed generally worse compared to MIRR, indicating greater variability in more clinically relevant settings. Importantly, the poor GIRR observed is not attributable to varying levels of scanning or clinical experience, as there was no consistency in one rater performing better than the other. This highlights that errors can be performed during the image acquisition because of the difficulties that clinicians face during this process. For example, identifying the thickest muscle part of a muscle involves subjectivity. On the other hand, measurement in Image J is easier to standardize due to agreed guidelines and often well-defined muscle borders on the image. Two clinical applications related to orthopaedic interventions arise from these findings: First, the pre-intervention and follow-up assessments of muscle atrophy or hypertrophy in the frame of an orthopaedic intervention should be performed by the same clinician in order to minimize the sources of errors. Second, if the follow-up is performed by another clinician than the pre-intervention evaluation, caution is needed when comparing muscle sizes acquired by these peers.

The mean muscle sizes for AbH, FDB, and FHB align with reference values from a recent systematic review by Haelewijn *et al.* encompassing a large participant pool. This emphasises our consistency in the use of the ultrasound protocols. For example, our AbH CSA (2.40, 2.50 and 2.65 cm^2) was similar to the reference value (2.43 cm^2) reported in a cohort of 781 subjects (13). Likewise, mean TA thickness (2.34, 2.42, 2.55 cm) also closely matched literature values (2.49 cm), while mean AbDM CSA and thickness are similar to the values reported by Swanson *et al.* (17), despite another study finding higher values (22). AbDM thickness (21,22), QP CSA (13,17), and FDL thickness values are in the lower range of what could be expected from the literature (25), while mean QP thickness values were lower to references values (13). PER CSA and thickness values are higher compared to those reported elsewhere (approximatively 4 cm^2 and 1.4 cm, respectively) (19,21,35). FDL CSA values face high discrepancy in previous articles (19,20,43), and our results are comparable to those of Johnson *et al.* (25). However, most of the aforementioned EFM size estimates are based on few cases (i.e., 10 to 28 participants), which may not be enough to result in valid norms (19-21,25), or with PER separated into longus and brevis (35), which makes comparison difficult.

Differences between our findings and previous studies may be due to our attempt to emulate real clinical settings. Firstly, the utilization of a low-cost USI device represents a notable difference from the methodologies used in previous studies (17,19,22,23,26,27). These devices typically provide lower image resolution and quality compared to high-end systems (52). Secondly, our study imposed a limited scanning time of 20–30 minutes per individual due to our pre-defined schedule, mimicking real-life clinical practice. This time constraint may have resulted in the acquisition of suboptimal images, as prior studies often do not specify such limits and may allow for longer examination times to optimize image quality. Also, although we performed a cine loop, we only registered a single image per muscle for evaluation. In contrast, prior studies may have had the opportunity to select multiple or higher-quality images for measurement, reducing variability. Lastly, we did not calculate the average of several independent measurements, which usually results in higher reliability (31).

Our method to calculate ICC and SEM may also have influenced reliability estimates. Indeed, the calculation

of ICC and SEM vary by model, type, and definition. According to Koo and Li (31), test-retest ICC should be calculated with "two-way mixed effects" model and "absolute agreement" definition. For inter-rater ICC, the model depends on the study design and clinical applications, while "absolute agreement" definition is preferred. Similarly, $SEM_{agreement}$ should be preferred over $SEM_{consistency}$, because the exact muscle size matters more than the consistent differences between measurements (31). This was our rational to report "single measures, absolute agreement" ICC and SEM, hypothesized to better reflect clinical practice (31), while some previous studies used multiple measures (25,26,28) or consistency (19,23) models.

Our results should be interpreted in the light of the study. Sample size calculation was based on the hypothesis that the ICC results would be superior to 0.9. However, the reliabilities associated with our primary outcomes were lower, and although the selected sample size allowed to expect confidence interval widths of less than 0.15, this was achieved in only two results. Hence, the desired power of this study is not guaranteed. Besides, although the investigators performed extended study piloting, they had approximately 1 year in USI experience, probably influencing the results. Despite these limitations, this study enhances our understanding of the generalizability of low-end USI in measuring IFM and EFM sizes.

Conclusions

This study investigated the reliability of USI for measuring IFM and EFM across both clinical and research settings, using a low-end device. Our findings indicate moderate to good test-retest reliability, poor to good GIRR, and poor to excellent MIRR. While test-retest reliability was generally favourable, GIRR tended to be lower, suggesting greater variability between investigators, particularly in image acquisition. Our results emphasize the need for caution when comparing muscle sizes acquired by different investigators, as variability in image acquisition can definitely impact reliability. Clinicians should be mindful of these limitations when using USI for muscle size assessment and should receive appropriate training. This training should transcend the common sources of technical error associated to the ultrasound device itself (depth, focus, frequency, intensity) by making clinicians aware of the intricacies at both the scanning and measurement stages. Future studies mimicking clinical settings are an essential next step in confirming generalizability of USI.

Acknowledgments

The authors would like to acknowledge Dr. Stefanie De Buyser from the Biostatistics Unit of Ghent University (Belgium) for her valuable help in the selection of the statistical methods. A part of this work has already been presented as a poster during the International Foot and Ankle Biomechanics congress, 2023 ("Test-retest and inter-rater reliability of extrinsic and intrinsic foot muscles using 2D ultrasound", first author N.H.).

Funding: A part of this work was supported by the French Community of Belgium as a FRIA grant (No. 40021590 to J-L.P-D.).

Footnote

Reporting Checklist: The authors have completed the GRRAS reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-1309/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1309/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Each participant signed the informed consent form. This study was approved by the Ethical Committees of UZ/KU Leuven (No. S67722) and UZ Ghent (No. B6702023000319).

References

1. Wager JC, Challis JH. Elastic energy within the human plantar aponeurosis contributes to arch shortening during the push-off phase of running. J Biomech 2016;49:704-9.
2. Kelly LA, Cresswell AG, Racinais S, Whiteley R, Lichtwark G. Intrinsic foot muscles have the capacity to control deformation of the longitudinal arch. J R Soc Interface 2014;11:20131188.
3. Farris DJ, Birch J, Kelly L. Foot stiffening during the push-off phase of human walking is linked to active muscle contraction, and not the windlass mechanism. J R Soc Interface 2020;17:20200208.
4. McKeon PO, Hertel J, Bramble D, Davis I. The foot core system: a new paradigm for understanding intrinsic foot muscle function. Br J Sports Med 2015;49:290.
5. Kelly LA, Lichtwark G, Cresswell AG. Active regulation of longitudinal arch compression and recoil during walking and running. J R Soc Interface 2015;12:20141076.
6. Moulodi N, Azadinia F, Ebrahimi-Takamjani I, Atlasi R, Jalali M, Kamali M. The functional capacity and morphological characteristics of the intrinsic foot muscles in subjects with Hallux Valgus deformity: A systematic review. Foot (Edinb) 2020;45:101706.
7. Severinsen K, Obel A, Jakobsen J, Andersen H. Atrophy of foot muscles in diabetic patients can be detected with ultrasonography. Diabetes Care 2007;30:3053-7.
8. Mickle KJ, Angin S, Crofts G, Nester CJ. Effects of Age on Strength and Morphology of Toe Flexor Muscles. J Orthop Sports Phys Ther 2016;46:1065-70.
9. Soysa A, Hiller C, Refshauge K, Burns J. Importance and challenges of measuring intrinsic foot muscle strength. J Foot Ankle Res 2012;5:29.
10. Tourillon R, Gojanovic B, Fourchet F. How to Evaluate and Improve Foot Strength in Athletes: An Update. Front Sports Act Living 2019;1:46.
11. Latey PJ, Burns J, Nightingale EJ, Clarke JL, Hiller CE. Reliability and correlates of cross-sectional area of abductor hallucis and the medial belly of the flexor hallucis brevis measured by ultrasound. J Foot Ankle Res 2018;11:28.
12. Franettovich Smith MM, Elliott JM, Al-Najjar A, Weber KA 2nd, Hoggarth MA, Vicenzino B, Hodges PW, Collins NJ. New insights into intrinsic foot muscle morphology and composition using ultra-high-field (7-Tesla) magnetic resonance imaging. BMC Musculoskelet Disord 2021;22:97.
13. Haelewijn N, Peters Dickie JL, Staes F, Vereecke E, Deschamps K. Current evidence regarding 2D ultrasonography monitoring of intrinsic foot muscle properties: A systematic review. Heliyon 2023;9:e18252.
14. Arima S, Maeda N, Komiya M, Tashiro T, Fukui K, Kaneda K, Yoshimi M, Urabe Y. Morphological and Functional Characteristics of the Peroneus Muscles in Patients with Lateral Ankle Sprain: An Ultrasound-Based Study. Medicina (Kaunas) 2022;58:70.
15. Fraser JJ, Koldenhoven R, Hertel J. Ultrasound Measures of Intrinsic Foot Muscle Size and Activation Following Lateral Ankle Sprain and Chronic Ankle Instability. J Sport Rehabil 2021;30:1008-18.
16. Leigheb M, de Sire A, Colangelo M, Zagaria D, Grassi FA, Rena O, Conte P, Neri P, Carriero A, Sacchetti GM, Penna F, Caretti G, Ferraro E. Sarcopenia Diagnosis: Reliability of the Ultrasound Assessment of the Tibialis Anterior Muscle as an Alternative Evaluation Tool. Diagnostics (Basel) 2021;11:2158.
17. Swanson DC, Sponbeck JK, Swanson DA, Stevens CD, Allen SP, Mitchell UH, George JD, Johnson AW. Validity of ultrasound imaging for intrinsic foot muscle cross-sectional area measurements demonstrated by strong agreement with MRI. BMC Musculoskelet Disord 2022;23:146.
18. Kazemi K, Saadi F, Javanshir K, Goharpey S, Shaterzadeh Yazdi MJ, Miraali SS, Nassadj G. Reliability of musculoskeletal ultrasonography for peri-ankle muscles in subjects with unilateral chronic ankle instability. J Bodyw Mov Ther 2021;27:565-72.
19. Crofts G, Angin S, Mickle KJ, Hill S, Nester CJ. Reliability of ultrasound for measurement of selected foot structures. Gait Posture 2014;39:35-9.
20. Hagoort I, Hortobágyi T, Vuillerme N, Lamoth CJC, Murgia A. Age- and muscle-specific reliability of muscle architecture measurements assessed by two-dimensional panoramic ultrasound. Biomed Eng Online 2022;21:15.
21. Willemse L, Wouters EJM, Pisters MF, Vanwanseele B. Intra-assessor reliability and measurement error of ultrasound measures for foot muscle morphology in older adults using a tablet-based ultrasound machine. J Foot Ankle Res 2022;15:6.
22. Mickle KJ, Nester CJ, Crofts G, Steele JR. Reliability of ultrasound to measure morphology of the toe flexor muscles. J Foot Ankle Res 2013;6:12.
23. Özgül B, Starbuck C, Polat MG, Abdeen R, Nester C. Inter and intra-examiner reliability of musculoskeletal ultrasound scanning of Anterior Talofibular Ligament and ankle muscles. J Ultrasound 2023;26:137-46.

24. Boulard C, Mathevon L, Arnaudeau LF, Gautheron V, Calmels P. Reliability of Shear Wave Elastography and Ultrasound Measurement in Children with Unilateral Spastic Cerebral Palsy. Ultrasound Med Biol 2021;47:1204-11.
25. Johnson AW, Stoneman P, McClung MS, Van Wagoner N, Corey TE, Bruening DA, Hunter TD, Myrer JW, Ridge ST. Use of Cine Loops and Structural Landmarks in Ultrasound Image Processing Improves Reliability and Reduces Error in the Assessment of Foot and Leg Muscles. J Ultrasound Med 2020;39:1107-16.
26. Battaglia PJ, Mattox R, Winchester B, Kettner NW. Non-Weight-Bearing and Weight-Bearing Ultrasonography of Select Foot Muscles in Young, Asymptomatic Participants: A Descriptive and Reliability Study. J Manipulative Physiol Ther 2016;39:655-61.
27. Franettovich Smith MM, Hides JA, Hodges PW, Collins NJ. Intrinsic foot muscle size can be measured reliably in weight bearing using ultrasound imaging. Gait Posture 2019;68:369-74.
28. Miyachi R, Kanazawa Y, Fujii Y, Ohno N, Miyati T, Yamazaki T. Reliability of lower leg muscle thickness measurement along the long axis of the muscle using ultrasound imaging, in a sitting position. J Phys Ther Sci 2022;34:515-21.
29. Kane D, Grassi W, Sturrock R, Balint PV. A brief history of musculoskeletal ultrasound: 'From bats and ships to babies and hips'. Rheumatology (Oxford) 2004;43:931-3.
30. Fraser JJ, Mangum LC, Hertel J. Test-retest reliability of ultrasound measures of intrinsic foot motor function. Phys Ther Sport 2018;30:39-47.
31. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15:155-63.
32. Cameron AF, Rome K, Hing WA. Ultrasound evaluation of the abductor hallucis muscle: Reliability study. J Foot Ankle Res 2008;1:12.
33. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96-106.
34. Borg DN, Bach AJE, O'Brien JL, Sainani KL. Calculating sample size for reliability studies. PM R 2022;14:1018-25.
35. Lobo CC, Morales CR, Sanz DR, Corbalán IS, Marín AG, López DL. Ultrasonography Comparison of Peroneus Muscle Cross-sectional Area in Subjects With or Without Lateral Ankle Sprains. J Manipulative Physiol Ther 2016;39:635-44.
36. Sierevelt IN, Zwiers R, Schats W, Haverkamp D, Terwee CB, Nolte PA, Kerkhoffs GMMJ. Measurement properties of the most commonly used Foot- and Ankle-Specific Questionnaires: the FFI, FAOS and FAAM. A systematic review. Knee Surg Sports Traumatol Arthrosc 2018;26:2059-73.
37. Sierevelt IN, van Eekeren IC, Haverkamp D, Reilingh ML, Terwee CB, Kerkhoffs GM. Evaluation of the Dutch version of the Foot and Ankle Outcome Score (FAOS): Responsiveness and Minimally Important Change. Knee Surg Sports Traumatol Arthrosc 2016;24:1339-47.
38. Baecke JA, Burema J, Frijters JE. A short questionnaire for the measurement of habitual physical activity in epidemiological studies. Am J Clin Nutr 1982;36:936-42.
39. Hertogh EM, Monninkhof EM, Schouten EG, Peeters PH, Schuit AJ. Validity of the modified Baecke questionnaire: comparison with energy expenditure according to the doubly labeled water method. Int J Behav Nutr Phys Act 2008;5:30.
40. Redmond AC, Crosbie J, Ouvrier RA. Development and validation of a novel rating system for scoring standing foot posture: the Foot Posture Index. Clin Biomech (Bristol) 2006;21:89-98.
41. Cote KP, Brunet ME, Gansneder BM, Shultz SJ. Effects of Pronated and Supinated Foot Postures on Static and Dynamic Postural Stability. J Athl Train 2005;40:41-6.
42. Angin S, Mickle KJ, Nester CJ. Contributions of foot muscles and plantar fascia morphology to foot posture. Gait Posture 2018;61:238-42.
43. Angin S, Crofts G, Mickle KJ, Nester CJ. Ultrasound evaluation of foot muscles and plantar fascia in pes planus. Gait Posture 2014;40:48-52.
44. McCreesh K, Egan S. Ultrasound measurement of the size of the anterior tibial muscle group: the effect of exercise and leg dominance. Sports Med Arthrosc Rehabil Ther Technol 2011;3:18.
45. Aquino MRC, Avelar BS, Silva PL, Ocarino JM, Resende RA. Reliability of Foot Posture Index individual and total scores for adults and older adults. Musculoskelet Sci Pract 2018;36:92-5.
46. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.
47. Shechtman O. The coefficient of variation as an index of measurement reliability. In: Doi SAR, Williams GM. editors. Methods of Clinical Epidemiology. Springer; 2013:39-49.

48. Taş S, Çetin A. Mechanical properties and morphologic features of intrinsic foot muscles and plantar fascia in individuals with hallux valgus. Acta Orthop Traumatol Turc 2019;53:282-6.
49. Zhang X, Pauel R, Deschamps K, Jonkers I, Vanwanseele B. Differences in foot muscle morphology and foot kinematics between symptomatic and asymptomatic pronated feet. Scand J Med Sci Sports 2019;29:1766-73.
50. Henderson AD, Johnson AW, Rasmussen LG, Peine WP, Symons SH, Scoresby KA, Ridge ST, Bruening DA. Early-Stage Diabetic Neuropathy Reduces Foot Strength and Intrinsic but Not Extrinsic Foot Muscle Size. J Diabetes Res 2020;2020:9536362.
51. Hopkins WG. Measures of reliability in sports medicine and science. Sports Med 2000;30:1-15.
52. Blaivas M, Brannam L, Theodoro D. Ultrasound image quality comparison between an inexpensive handheld emergency department (ED) ultrasound machine and a large mobile ED ultrasound system. Acad Emerg Med 2004;11:778-81.

Cite this article as: Haelewijn N, Peters-Dickie JL, de Ridder R, Deschamps K, Detrembleur C, Lobet S, Spanhove V. Quantitative ultrasonography of the foot muscles: a comprehensive perspective on reliability. Quant Imaging Med Surg 2025;15(1):203-216. doi: 10.21037/qims-24-1309