Inter-Rater Reliability and Measurement Error of Sonographic Muscle Architecture Assessments

Niklas König, MSc, Michael Cassel, MD, Konstantina Intziegianni, MSc, Frank Mayer, MD

Objectives—Sonography of muscle architecture provides physicians and researchers with information about muscle function and muscle-related disorders. Inter-rater reliability is a crucial parameter in daily clinical routines. The aim of this study was to assess the inter-rater reliability of sonographic muscle architecture assessments and quantification of errors that arise from inconsistent probe positioning and image interpretation.

Methods—The medial gastrocnemius muscle of 15 healthy participants was measured with sagittal B-mode ultrasound scans. The muscle thickness, fascicle length, superior pennation angle, and inferior pennation angle were assessed. The participants were examined by 2 investigators. A custom-made foam cast was used for standardized positioning of the probe. To analyze inter-rater reliability, the examinations of both raters were compared. The impact of probe positioning was assessed by comparison of foam cast and freehand scans. Error arising from picture interpretation was assessed by comparing the investigators’ analyses of foam cast scans independently. Reliability was expressed as the intraclass correlation coefficient (ICC), inter-rater variability (IRV), Bland-Altman analysis (bias ± limits of agreement [LoA]), and standard error of measurement (SEM).

Results—Inter-rater reliability was good overall (ICC, 0.77–0.90; IRV, 9.0%–13.4%; bias ± LoA, 0.2 ± 0.2–1.7 ± 3.0). Superior and inferior pennation angles showed high systematic bias and LoA in all setups, ranging from 2.0° ± 2.2° to 3.4° ± 4.1°. The highest IRV was found for muscle thickness (13.4%). When the probe position was standardized, the SEM for muscle thickness decreased from 0.1 to 0.05 cm.

Conclusions—Sonographic examination of muscle architecture of the medial gastrocnemius has good to high reliability. In contrast to pennation angle measurements, length measurements can be improved by standardization of the probe position.

Key Words—gastrocnemius muscle; inter-rater variability; muscle architecture; musculoskeletal ultrasound; sonography

The geometric construction of a muscle is defined by the location of fascicles and aponeuroses, by which the muscle interacts with the tendon to its insertion at the bone. Muscle architecture can be described by muscle thickness, fascicle length, and the angle between fascicles and aponeuroses (Figure 1). The capacity of a muscle to create forces directly depends on its architectural characteristics. For example, the larger the fascicle length, the greater the contraction length and therefore its impact on the velocity potential of the muscle. The pennation angle also influences force development capacity, as muscles with larger pen-
nation angles have more fibers in parallel. Thus, the fascicle length and pennation angle predict the functional capacity of a particular muscle.

Muscle function can be increased by strengthening exercises and decreased by disuse or aging. As muscle tissue is highly adaptable, alteration of muscle function and concomitant muscle architecture will become noticeable early. It was shown that 23 days of muscle disuse caused declines in strength, fascicle length, and pennation angle of 21%, 7%, and 7%, respectively. In the same manner, 90 days of bed rest caused decreases in fascicle length and pennation angle of 10% and 13%. Vice versa, it was shown that strength training also altered muscle thickness, pen- nanation angles, and fascicle length. An immediate response and long-term adaptation were found in muscle architecture after exhausting training sessions. Furthermore, the relationship between muscle architecture and functional performance was obvious in competitive sprinting, since the sprinters’ best time was determined by their muscles’ fascicle length. Considering that training and disuse cause changes in muscle architecture, it can be concluded that muscle architecture contains information about the maximal force capacity and training history of a muscle.

In clinical practice, sonography of muscle structures is one of the most frequent and cost-effective methods used by physicians. It allows diagnosis of structural lesions and muscle function as well as assessment of hypertrophy and deconditioning. Also, sonography can be used to assess muscle size and thus direct identification of structural asymmetry. If conventional strength measurements are not possible, sonography might be a useful complementary technique for the prediction of muscle strength and function. Additionally, in patients with tendon ruptures, it was shown that assessment of muscle architecture can identify the severity of the injury. Moreover, sonographic examinations of muscles can help prevent injuries. It has been shown that the pennation angle determines the amount of stress on the muscle-tendon complex, and higher pennation angles and shorter fascicles have been found to be risk factors for muscle tears. Therefore, sonographic examinations of muscle architecture allow for a more comprehensive understanding of muscle-related disorders and injury mechanisms and consequently might have an impact on clinical decision making.

To obtain meaningful information about patient-specific situations, the validity of sonographic evaluation has to be viewed in daily clinical routines. As the same patient is not always examined by the same clinician, different judgments of 2 or more observers have to be considered. This influence is quantified by the inter-rater reliability, which also allows for interpretation of the maximum precision of the tool. Measurement of inter-rater reliability provides information on whether the assessed difference in a single patient is true or due to measurement error. In sonographic evaluation of muscle architecture, 2 main sources of error arise. First, proper positioning of the probe has to be taken into account. Klímstra et al defined
probe placement as the probe-skin interface, whereas probe orientation refers to the rotation of the probe in the transverse and frontal planes. In contrast, in this study, the probe was held manually by the investigators, which does not allow for differentiation between probe placement and orientation. Therefore, the term “positioning” is used here to describe errors that result from any form of inconsistent probe application. A second source of error is image interpretation by the investigators to identify the geometric structure. Previous reports showed excellent reliability, with coefficients of variation of less than 10%.29,30 However, these studies were conducted in highly standardized laboratory settings, which did not allow transfer of the reliability of measurements to a clinical context.

The aim of this study was to assess the inter-rater reliability of sonographic muscle architecture measurements of the medial gastrocnemius muscle, to describe the magnitude of error that arises from inconsistent probe positioning, and to evaluate the measurement error by varying interpretation of the images collected.

**Materials and Methods**

**Participants**
Fifteen recreationally active participants (9 female: mean ± SD, 28 ± 3 years, 170 ± 7 cm, and 62 ± 8 kg; and 6 male: 29 ± 5 years, 184 ± 6 cm, and 81 ± 5 kg) participated in the study. Participants were excluded from the study if they reported any acute or chronic musculoskeletal injury of the lower limb. Participants were only included after signing written informed consent. The study was approved by the local Ethics Committee.

**Measurement Procedure**
The muscle architecture of the left medial gastrocnemius was assessed during rest. The participants lay prone with the knee fully extended and the ankle 90° flexed. To ensure the correct position of the ankle, the foot was secured in a dynamometer (Con-Trex MJ, Physiomed Elektromedizin AG, Schnaittach/Laipersdorf, Germany), and 90° flexion was adjusted. Subsequently, sagittal B-mode ultrasound scans (Vivid q; GE Healthcare, Tirat Carmel, Israel) of the gastrocnemius muscle belly were conducted with a 7.5-MHz continuous linear ultrasound array (4–13 MHz). Presets were standardized at a frequency of 13 MHz and a depth of 3 cm. In this investigation, the location of the probe with respect to the leg length was standardized, whereas the probe rotation around the sagittal-transverse axis varied between investigators. To standardize the location of the probe on the muscle belly, the distance from the lateral knee joint to the lateral malleolus of the left leg was measured, and subsequently the probe was placed at one-third of this distance distally from the knee joint space.12

To allow inter-rater reliability assessment, all participants were measured by 2 investigators independently (N.K. and M.C.). The investigators held the probe manually at the predefined location to ensure that both aponeuroses and fascicles were clearly visible in the scan. To allow quantification of the independent influence of probe positioning, a custom-made foam cast was used to hold the ultrasound probe on the predefined location (Figure 2). In this setup, the probe was held by the foam cast, which was fixed on the leg with Velcro straps. Attachment of the foam cast was only conducted by a single investigator (N.K.). For each participant, 3 scans of the muscle architecture were conducted by each investigator (N.K., M.C., and foam cast), resulting in a total of 135 images (3 [muscle] × 3 [investigators] × 15 [participants]), which were stored in the Digital Imaging and Communications in Medicine format and analyzed offline using ImageJ version 1.46k23 software (National Institutes of Health, Bethesda, MD).

**Data Analysis**
The muscle thickness, fascicle length, superior pennation angle, and inferior pennation angle were measured to analyze the muscle architecture. Muscle thickness was measured as the mean of the distance between the upper and lower aponeuroses at the most proximal and distal parts. These values were also considered to identify whether the aponeuroses were in parallel, which was necessary to accurately calculate the fascicle length. As the fascicle length is longer than the probe, the fascicles extended the field of view on sonography; therefore, it was necessary to...
were assigned to the participants according to the random-number code, which did not allow identification of the images, the investigators were only aware of the random-number code that was also noted on the case file. During testing, images were saved according to number codes in the master file. All 3 images of each muscle scan were analyzed, and the mean value was considered for statistical analysis.

**Statistical Analysis**

The results on inter-rater reliability are presented as the intraclass correlation coefficient (ICC, 2.1), inter-rater variability (IRV, percent), and Bland-Altman analysis with bias and limits of agreement (LoA). Inter-rater variability provides information about the relative difference between 2 raters, where lower levels indicate higher reliability. Bias is the mean absolute difference between raters and is supposed to assess systematic error; values closer to 0 indicate higher reliability. The LoA are supposed to be measures of random error between raters, where narrower limits are preferable. To receive valid LoA, the data were checked for heteroscedasticity or homoscedasticity by applying a Pearson moment correlation, with the variables “difference” and “mean” of 2 measurements. The $\alpha$ error for the Pearson moment correlation was set at 5%, and a significant correlation indicated heteroscedasticity. If heteroscedasticity was present in a data set, the data were log transformed before the Bland-Altman analysis was conducted. Additionally, the standard error of measurement (SEM) is presented to provide a measure that can be directly interpreted by clinicians, as the SEM preserves the unit of the measurement. It was calculated as follows:

$$SEM = SD \sqrt{1 - ICC}.$$

The ICCs, Pearson moment correlations, and log transformations were calculated with SPSS version 19 software (IBM Corporation, Armonk, NY), whereas the remaining calculations were performed in Microsoft Excel 2010 (Microsoft Corporation, Redmond, WA).

**Results**

**Inter-Rater Reliability**

The mean values ± standard deviations for the assessed muscle architecture parameters by the raters are presented in Table 1. The inter-rater reliability of muscle architecture measurements showed ICC values for the 4 measurements that ranged between 0.77 and 0.90 (Table 2). Muscle thickness showed the highest IRV of 13.4%. Assessment of the superior and inferior pennation angles had IRV of 10.7% and 9.2%, respectively. The lowest IRV was shown for fascicle length, at 9%. The SEMs of the angle, muscle thickness, and fascicle length measurements were 1.0°, 0.1 cm, and 0.4 cm, respectively.
Positioning Error
The different probe positions during the foam cast and free-hand scans showed an influence on the superior pennation angle, with an ICC of 0.62 and IRV of 22%. A systematic bias was found for both angles (superior pennation angle, 3.4°; and inferior pennation angle, 2.2°) and LoA for the superior and inferior pennation angles (±4.1° and ±3.6°, respectively; Figure 3). However, the inferior pennation angle showed a good ICC of 0.86 and IRV of 10.9%. Except for the superior pennation angle, the SEM only changed marginally between the probe positioning and inter-rater setups (muscle thickness, 0.1 cm; superior pennation angle, 2.1°; inferior pennation angle, 1.4°; and fascicle length, 0.4 cm).

Interpretation Error
When a common probe position was used but the scans were interpreted by different raters, reliability became excellent for length measurements. Muscle thickness and fascicle length showed ICCs of 0.96 and 0.93, respectively. Inter-rater variability of muscle thickness dropped to 3.6%, and that of fascicle length dropped to 5.9%. The SEMs for muscle thickness and fascicle length were 0.05 and 0.2 cm, respectively. The reliability of angle measurements did not change in the same manner, with SEMs for the superior and inferior pennation angles of 1.3° and 1.7°, respectively.

Discussion
The first aim of this study was to evaluate the inter-rater reliability of muscle architecture assessments in a clinical setting. Although the transducer was manually held during collection of scans, and 2 investigators in this study had only little experience with sonographic assessments, results show that muscle architecture can be assessed with good to high reliability. Compared to the literature, the values presented seem to be valid. In their study on medial gastrocnemius muscles of healthy male participants, Narici et al30 found average muscle thickness, pennation angle (mean of the superior and inferior pennation angles) and fascicle length values of 2.0 ± 0.2 cm, 17.3° ± 2.6°, and 5.1 ± 0.4 cm, respectively. In another study on medial gastrocnemius muscles of healthy young male and female participants, a mean inferior pennation angle of 17.4° ± 2.5° was reported.34 Thus, it is concluded that muscle architecture measurements can be assessed reliably and validly within a clinical context.

Generally, inter-rater setup assessment of muscle thickness showed the lowest reliability compared to the other 3 parameters. Furthermore, muscle thickness was more robust in interpretation compared to the positioning setup. This finding might have been due to the different pressure applied by the investigators during probe place-
ment on the skin and is supported by the post hoc comparison of the individual scans, which showed consistently that one rater had a narrower distance between the aponeuroses. Previous agreement on practical aspects of how to conduct the ultrasound scans probably would have avoided this problem and could have led to increased inter-rater reliability of muscle thickness measurements. Thus, the coefficient of variation for muscle thickness measurements in this study was higher (14.5%) compared to results ranging between 2.1% and 3.1% in a laboratory study with more standardized probe placement. Consequently, it is recommended to find common agreement between all investigators about practical aspects of probe application. In particular, the applied pressure during probe placement should be adapted, since it might cause different tissue deformation and potentially result in low reliability of muscle thickness measurements.

The second and the third aims of this study were to evaluate error that arises from probe positioning and image interpretation. In comparison of the positioning and interpretation setups, it is apparent that inconsistent probe positioning influences length measurements, whereas angle measurements generally reveal wider LoA and therefore increased levels of random error. The LoA across all 3 setups relative to the mean for the superior and inferior pennation angles were 18% and 17%, respectively, compared to the 9% relative LoA for muscle thickness. It appears that angle measurements are affected by positioning and interpretation error and are generally less reliable. Furthermore, the reliability of length measurements can be improved when positioning error is removed, which is reflected in SEMs for muscle thickness of 0.1 and 0.05 cm in the positioning and interpretation setups as well as drastically reduced LoA (Figure 3). Thus, it can be concluded that length measurements are less affected by error arising from interpretation of sonograms, which is supported by Klimstra et al., as they found a main effect on probe rotation for muscle thickness and fascicle length but none for assessment of pennation angles in their laboratory study. Compared to this finding, angle measurements remained nearly unchanged when the impact of the positioning was removed, with mean variability of 16.3% and 13.8% for the positioning and interpretation setups, respectively. Furthermore, it appears that angle measurements are highly sensitive to image interpretation; thus, more raters involved in the analysis would cause the reliability to decrease. Hence, careful probe placement can help increase the precision of sonographic length measurements, whereas determination of pennation angles is affected by error from various sources.

The results on measurement error have to be interpreted according to their clinical relevance to decide on their applicability in daily routines. Especially, these results provide information on which differences can be assumed to be real within the examination of a single participant. The SEMs in the inter-rater comparison were 1.0° for angle measurements and 0.1 and 0.4 cm for muscle thickness and fascicle length, respectively. By removing the impact of probe positioning using the standard foam cast, the SEMs for length measurements decreased to 0.05 and 0.2 cm for muscle thickness and fascicle length, respectively. Therefore, it can be speculated that all changes of less than 1.0° for angle measurements, 0.05 cm for muscle thickness, and 0.2 cm for fascicle length measurements have to be interpreted carefully, as they are within the SEMs. Vice versa, changes that appear greater than the limits can be considered valid alterations of muscle architecture. Although the values presented here seem to be very small, this information has to be viewed relative to the actual change potential of a muscle. According to the data, an effect of training or deconditioning of a muscle can only be regarded as valid if the muscle thickness is increased or decreased by greater than 10%. However, the potential of muscles to increase in thickness with training has been reported to range from 5% to 12%. To assess valid alterations in pennation angles, the difference has to be greater than 1.0° (SEM of angle measurements). Considering a muscle’s mean pennation angle of 17°, this difference represents a necessary change of 6%. In the literature, the pennation angle was reported to change by about 7% after 23 days of limb suspension. As the adaptation potential of muscles is rather small compared to the SEM during sonographic assessments, it remains a matter of debate whether sonographic assessments of muscle architecture are precise enough to track small changes that occur in individual patients in clinical practice. Especially, as the results of this study were limited to measurements conducted on very lean, healthy, active participants, sonographic measurements in those with higher subcutaneous fat or pathologic changes in muscle tissue might be less reliable. Also, Raj et al. argued that sonographic assessment of muscle architecture might not be sensitive enough on the individual level but is still sufficient to evaluate differences on the group level. Therefore, more data on individual adaptations of muscles after training or deconditioning are necessary to reveal the applicability of sonographic assessments of muscle architecture for the use in the clinical context.

In conclusion, with a minimum of standardization, sonographic architecture measurements of the medial gastrocnemius muscle can be conducted by different inves-
Figure 3. Bland-Altman plots for muscle thickness (MT), superior pennation angle (SPA), inferior pennation angle (IPA), and fascicle length (FL) for the positioning and interpretation setups. Note that due to heteroscedasticity in the interpretation setup, fascicle length values were log transformed before Bland-Altman analysis.
tigators with at least good reliability in a cohort of active asymptotic individuals. The reliability of length measurements can be improved by standardizing the probe position. Assessment of pennation angles is biased by relatively high systematic and random error that arises from image interpretation and probe positioning. Further investigations are required to determine whether sonographic assessments of muscle architecture are sensitive enough to track adaptations in individual muscles.

References


