Published in final edited form as:

Acad Radiol. 2009 November ; 16(11): 1424–1432. doi:10.1016/j.acra.2009.05.001.

Reliability of Quantitative Ultrasound Measures of the Biceps and Supraspinatus Tendons

Jennifer L. Collinger, PhD^{1,2,3}, Dany Gagnon, PhD, PT^{1,2,4}, Jon Jacobson, MD⁵, Bradley G. Impink, BS^{1,3}, and Michael L. Boninger, MD^{1,2,3}

¹ Human Engineering Research Laboratories, VA Pittsburgh Healthcare System, Pittsburgh, PA

² Department of Physical Medicine and Rehabilitation, University of Pittsburgh

³ Department of Bioengineering, University of Pittsburgh

⁴ School of Rehabilitation, Faculty of Medicine, University of Montreal

⁵ Department of Radiology, University of Michigan

Abstract

Rationale and Objectives—Ultrasound is a proven method for examining soft tissue structures including tendons, and recently quantitative ultrasound has become more prevalent in research settings. However, limited reliability data has been published for these new quantitative ultrasound measures. The main study objective is to quantify the reliability and measurement error of multiple quantitative ultrasound imaging protocols for the biceps and supraspinatus tendons.

Materials and Methods—Two examiners captured ultrasound images of the non-dominant long head of the biceps tendon and supraspinatus tendon from 15 able-bodied participants and five manual wheelchair users. Each examiner captured two images per subject under two different preparations which includes subject positioning and reference marker placement. Image processing (reading) was performed twice to compute nine quantitative ultrasound measures of greyscale tendon appearance using first-order statistics and texture analysis. Generalizability theory was applied to compute inter- and intra-rater reliability using the coefficient of dependability (Φ) for multiple study design protocols.

Results—Inter-rater reliability was generally low ($0.26 < \Phi < 0.82$), and we recommend that a single evaluator capture all images for quantitative ultrasound protocols. Most ($n = 14$ of 18) of the quantitative ultrasound measures exhibited at least moderate ($\Phi > 0.50$) intra-rater reliability for a single image, captured under one preparation, and read once.

Conclusion—By following a protocol designed to minimize measurement error, one can increase the reliability of quantitative ultrasound measures. We believe that an appropriately designed protocol will allow quantitative ultrasound to be used as an outcome measure to identify structural changes within the tendon.

Keywords

Ultrasound; Reliability; Tendon; Biceps; Supraspinatus

Correspondence address: Michael Boninger, MD, Human Engineering Research Laboratory, VA Pittsburgh Healthcare System, 7180 Highland Dr., Building 4, 151R-1, Pittsburgh, PA 15206; Phone: (412) 954-5287; Fax: (412) 954-5340; email: boninger@pitt.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

INTRODUCTION

Ultrasound is a well-established, non-invasive, method for examining soft tissue structures of the shoulder including rotator cuff tendons and the long head of the biceps tendon.

Traditionally, ultrasound imaging has been a valuable tool in clinical practice to qualitatively evaluate the integrity of musculoskeletal structures. Symptoms of pathology of the rotator cuff detected by ultrasound include a hypoechoic tendon appearance (1–3) and hypertrophy of the long biceps tendon (4–5). Biceps tendon inflammation often coexists with rotator cuff disease and may be a result of chronic inflammation and impingement (4) or bicipital tenosynovitis (6). Healthy tendons are known to have a well-organized, uniform, hyperechoic pattern of collagen along the long-axis of the tendon. Conversely, tendons with pathology have a more disorganized, diffuse, or hypoechoic appearance on ultrasound. In essence, tendon health is clinically evaluated by visually examining the greyscale image texture of the tendon.

Quantifying tendon greyscale texture has many potential benefits, particularly in research focused on reducing the risk of chronic tendinopathy.

To date, quantitative analysis of tendons on ultrasound has primarily been limited to measurements of tendon thickness or cross-sectional area (7). However, researchers have applied first-order statistics and texture analysis to other medical images to characterize micro-structure (8–10). One study classified muscle based on the concentration of contractile components using first-order greyscale statistics. (9). Another research group employed quantitative ultrasound techniques, including structural measurements and mean echogenicity, to discriminate between skeletal muscle in children with and without neuromuscular disease (11,12). Finally, a recently published study examined eight spatial frequency parameters, derived from two-dimensional Fourier analysis of the ultrasound images, that discriminated between subjects with and without Achilles tendinopathy with approximately 80% accuracy (13). These studies provide a basis for the application of quantitative ultrasound to understanding chronic musculoskeletal pathology development. Despite the growing interest for this well established technique in research settings, there have been very few attempts to establish the psychometric properties of quantitative ultrasound measures (14–16). None have evaluated the reliability and measurement error for a set of tendon image features under multiple testing conditions. This information is needed to develop effective measurement protocols in order to quantify acute or chronic tendon changes linked to cumulative trauma disorders of the upper limb.

The main objective of this study was to quantify the reliability and measurement error associated with quantitative ultrasound outcomes of the long head of the biceps and supraspinatus tendons among able-bodied subjects and long-term manual wheelchair users. Future research may focus on manual wheelchair users as they have a high incidence of upper limb musculoskeletal injury and could benefit from interventions designed to reduce risk exposure. The second objective of this study was to define a time-efficient, reliable, quantitative ultrasound measurement protocol that could be used in the future to quantify acute or chronic tendon changes. Nine outcome measures were computed to quantitatively describe the greyscale tendon appearance on ultrasound using image analysis techniques such as first-order statistics and texture analysis. We expected that inter-evaluator reliability would be lower than intra-evaluator reliability, as has been previously reported (14). Overall, it was expected that a standardized protocol in which one examiner records a single ultrasound image of the tendon after setting-up reference skin markers would result in a reliable ($\Phi > 0.75$) and precise (standard error of measurement $< 15\%$) quantitative ultrasound measurement outcomes.

MATERIALS AND METHODS

Participants

Fifteen able-bodied individuals (12 male, 3 female; age=43.8±13.1 years; height= 1.80±0.09 m; body mass=86.58±11.13 kg) and five manual wheelchair users (5 male; age=43.5±15.5 years; height= 1.80±0.12 m; body mass=84.64±18.22 kg) volunteered to participate in this reliability study, which was approved by our local review board. All five manual wheelchair users had a spinal cord injury and were an average of 15.5±10.1 years post-injury. All participants provided informed consent before entering the study and were analyzed as a single group. Subjects were eligible to participate if they were between 18 and 75 years old and if they were able to attend multiple ultrasound sessions. Participants were not screened for the presence of shoulder pain or pathology prior to participation.

Ultrasound Examination

Two examiners conducted ultrasound examinations of each participant. Both were trained in a specially developed quantitative ultrasound examination protocol and had approximately 3 years of experience. Study investigators met frequently to review and refine this quantitative ultrasound examination protocol, which is described in detail below. All quantitative ultrasound examinations were conducted using a Phillips HD11 1.0.6 ultrasound machine with a 5–12 MHz 50 mm linear array transducer (Philips Medical Systems, Bothell, WA). The machine settings were kept identical across all examinations performed for all subjects. Of note, image field depth was set to 4 cm and gain was set at 85 dB. All images were saved for later analysis.

Test—The examination of the non-dominant biceps tendon was performed with the subject sitting in an upright position with the upper arm in line with the trunk, the elbow flexed to 90°, the forearm supinated with the wrist in a neutral position and the hand resting on the ipsilateral thigh (Figure 1a). For the biceps tendon, the proximal end of the transducer was positioned such that the apex of the lesser tuberosity of the humerus was at the edge of the ultrasound image field of view and oriented to obtain a longitudinal view of the widest part of the tendon, while maximizing collagen fiber reflection. The biceps brachii tendon fibers were imaged perpendicular to the ultrasound beam to minimize anisotropy. The transducer location was traced on the skin and a steel “A-shaped” reference marker was taped to the skin at the distal end of the transducer footprint. The crossbars of the reference marker (Figure 2) create an interference pattern in the ultrasound image (Figure 3) which is used to define the tendon region of interest (ROI) used during image analysis. Once this initial set-up was completed (preparation #1), two consecutive images (images #1 and #2) of the long head of the biceps were collected while avoiding exerting undue pressure with the ultrasound head. Once the images were taken, the markers were removed and the skin was cleaned to erase all marks.

A similar protocol was followed for the supraspinatus tendon although the upper limb positioning was modified to optimize viewing. For this tendon, the subject placed his palm on his lower back, or wheelchair backrest, with the elbow pointing posteriorly (Figure 1b). The transducer was positioned to obtain a transverse image of the widest part of the supraspinatus tendon, with the rotator interval and cross-sectional view of the biceps tendon clearly in view. The ultrasound beam was maintained perpendicular to the supraspinatus tendon fibers so that the tendon appeared hyperechoic and the adjacent humerus head cortex was brightly reflective in order to avoid tendon anisotropy. A second reference marker was taped to the skin and two images (images #1 and #2) were collected under preparation #1.

Retest—After a rest period of approximately 30 minutes, participants underwent a second quantitative ultrasound examination (preparation #2) during which two additional images

(images #3 and #4) of the biceps and supraspinatus tendons were recorded. Care was taken to maintain the same standardized seated position, to keep the ultrasound machine settings constant and to replicate the exact measurement protocol during the test and retest examinations. No significant change in the quantitative ultrasound outcomes were expected during the initial and final tests as subjects were instructed to rest between these tests.

Image Analysis

Based on the selected dynamic range, each pixel in the ultrasound images represents a greyscale value ranging from 0 (black) to 255 (white). Collagen will reflect ultrasound waves back to the transducer and appear hyperechoic (closer to 255), while the waves pass through fluid which appears darker (closer to 0) on the resulting image. The ROI for each tendon was defined in relation to the center of the interference pattern created by the externally placed reference markers using a customized interactive Matlab function (The Mathworks, Natick, MA) as illustrated in Figure 3. All images were processed in Matlab twice (readings #1 and #2) by an evaluator who was blinded to the preparation and image number during analysis. The following features were calculated for the tendon ROI: tendon thickness, echogenicity, variance, skewness, kurtosis, entropy, contrast, homogeneity, and energy (17,18).

The upper and lower boundaries of the ROI were outlined manually and average tendon thickness was computed. Increased tendon thickness may be a result of chronic inflammation and may indicate the presence of rotator cuff pathology (4). To determine echogenicity, the mean pixel greyscale was computed from all pixels within the ROI. The collagen structure of a healthy tendon is organized parallel to the long axis and will have a brighter appearance as compared to a tendon with degeneration. The greyscale values of all pixels within the ROI were represented as a greyscale histogram from which first-order statistics were derived. The variance, skewness, kurtosis, and entropy describe the spread, symmetry, peakedness, and uniformity, respectively, of the greyscale histogram. While these image features have previously been used to describe ultrasound image texture, clinical interpretation of these features remains to be clarified once the psychometric properties are known. In general, a healthy tendon with highly aligned collagen fibers should have a striped appearance of alternating light and dark bands, while a tendon with degeneration would have a more uniform, darker appearance. Based on previous comparisons of muscle tissue (9), we would expect the greyscale histogram of a healthy tendon to be wider (increased variance), more symmetrical (less skewed), flatter (less kurtosis), and more heterogeneous (increased entropy).

Second-order statistics provide additional information about the texture of a ROI. This analysis considers the pixels of an image in pairs at a set distance (d) apart with a relative orientation angle (ϕ). For each histogram, a co-occurrence matrix describes the probability of a pixel pair with a defined spatial relationship (d, ϕ) having given greyscale level values (r, c) where r and c range from 0–255 (19). Using Matlab, texture coefficients (contrast, energy, and homogeneity) were derived from this co-occurrence matrix which describes the spatial dependence of the pixels in a ROI. Since a horizontal striped pattern within the ROI is expected due to the collagen organization within the tendon, the sum of texture values for $\phi = 90^\circ$ and $d=1:5$ was computed.

Contrast measures the intensity difference between a pixel and its neighbour over the entire image, and is equal to zero for a constant image and increases for a heterogeneous image. Energy is defined as the sum of squared elements along the diagonal of the co-occurrence matrix and is equal to 1 for a constant image and decreases with the presence of spatial greyscale texture. Homogeneity measures how close the distribution of elements in the co-occurrence matrix is to a diagonal matrix. Homogeneity equals 1 for a diagonal co-occurrence matrix and gets closer to zero as the spatial texture increases. Therefore a healthy tendon would have higher contrast, lower energy, and lower homogeneity than a tendon with signs of degeneration.

Statistical analysis

Generalizability theory is considered an extension of the intraclass correlation coefficient (ICC) and provides additional information about the sources of variance and effect of the experimental design (20). Based on the analysis of variance, the generalizability theory is divided into two parts: the generalizability study (G-study) and the dependability study (D-study). The G-study allows one to determine the magnitude of the variances attributed to specified sources of variance. The D-study relies on information generated from the G-study to determine the reliability of specific protocol designs.

First, we analyzed data from both evaluators to determine overall, inter-evaluator, inter-preparation, and inter-image reliability, measured as the dependability coefficient. Subject (S), Evaluator (E), Preparation (P), Image (I), and all possible interactions of these four facets were included as possible sources of variance. The dependability coefficient (Φ), ranges from 0 to 1, and is computed as the ratio between the inter-subject variance and the sum of the inter-subject variance and all possible sources of error (21). General interpretation guidelines suggest that a $\Phi < 0.50$ represents poor reliability, and a Φ between 0.50 and 0.75 indicates moderate reliability, while values greater than 0.75 confirms good reliability (22).

Since the evaluator contributed the most to the total variance in comparison to the other facets (Table 1), data from a single evaluator (Evaluator #1) was used for the remainder of the analysis. The reliability of three protocol designs was evaluated using the dependability coefficient and standard error of measurement (SEM). Absolute SEM represents the square root of the absolute error variance. Normalized SEM (SEM_{norm}), expressed as a unitless percentage, was calculated as $(SEM/overall\ mean) * 100$ to facilitate clinical interpretation. The overall mean reflects the mean of a given outcome averaged at the test ($n=4$) and retest ($n=4$) sessions for all participants. The analysis of variance and generalizability analysis were completed with an adapted version of the GENOVA statistical software, version 2.2 (JE Crick/National Board of Medical Examiner, Philadelphia, PA).

RESULTS

Inter-rater Reliability

Mean (SD) quantitative ultrasound results for both evaluators are presented in Table 1. Overall reliability for a study design employing a single evaluator capturing a single image during one preparation ($E=1$; $P=1$; $I=1$) is described, computed from a random D-study model which allows all sources of variance to contribute to measurement error. Three systematic facets of error were analyzed to specifically determine inter-evaluator, inter-preparation, and inter-image reliability. In each case, the facet of interest contributed to measurement error, while the other facets were fixed in the mixed D-study model. Inter-evaluator reliability was the lowest for all quantitative ultrasound measures for both tendons. While good ($\Phi > 0.75$) inter-evaluator reliability was achieved for biceps and supraspinatus tendon thickness, most measures showed moderate ($0.5 < \Phi < 0.75$; $n=4$) or poor ($\Phi < 0.5$; $n=12$) inter-evaluator reliability. The inter-preparation dependability coefficient, Φ , describes reliability between the test and re-test sessions, while inter-image Φ isolates reliability of quantitative ultrasound measures computed from two images captured during a single preparation. Inter-preparation reliability was generally lower than inter-image reliability. Inter-preparation Φ ranged from 0.528–0.908 for the 18 quantitative ultrasound measures, while inter-image Φ ranged from 0.463–0.962. Inter-preparation and inter-image reliability were moderate or good ($\Phi > 0.5$) for all ultrasound measures for both tendons, except supraspinatus kurtosis ($\Phi = 0.463$). No systematic differences in Φ were noted between the biceps and supraspinatus tendons. Tendon thickness was consistently the most reliable quantitative ultrasound measure.

Intra-rater Reliability

The test-retest dependability coefficient (Φ), standard error of measurement (SEM) for a 90% confidence interval, and normalized SEM_{norm} are summarized in Table 2. For each tendon, three experimental scenarios are presented. The first ($P=1$; $I=1$; $R=1$) describes a situation in which a single image is captured during a single preparation and is read only one time. Essentially, this compares a single measurement value to a hypothetical true value. Imaging of the biceps tendon with this experimental design would yield good ($\Phi>0.75$) reliability for tendon thickness (0.906) and homogeneity (0.764), moderate ($0.5<\Phi<0.75$) reliability for echogenicity (0.742), variance (0.614), skewness (0.533) entropy (0.616), contrast (0.646), and energy (0.709), and poor ($\Phi<0.5$) reliability for kurtosis (0.462). SEM_{norm} ranged from 2.28% (entropy) to 47.3% (skewness) for the biceps tendon. For the supraspinatus tendon, the dependability coefficients confirmed good ($\Phi>0.75$) reliability for tendon thickness (0.921) and echogenicity (0.754), moderate ($0.50<\Phi<0.75$) reliability for skewness (0.579), contrast (0.589), energy (0.618), and homogeneity (0.657), and poor ($\Phi<0.50$) reliability for variance (0.474), kurtosis (0.477), and entropy (0.484). SEM_{norm} fluctuated between 0.484% (entropy) and 163% (skewness). For both tendons, SEM_{norm} for skewness was twice as large as the second largest SEM_{norm} .

D-study measurement error estimates are presented for two additional experimental designs. The first ($P=1$; $I=1$; $R=2$) shows only a marginal improvement in reliability if an additional reading is performed. The second scenario ($P=1$; $I=2$; $R=1$) illustrates the effect of using the average outcome measure value from two images taken under a single preparation. Slightly larger improvements in reliability are observed using this experimental design.

DISCUSSION

The objective of this study was to quantify the reliability and measurement error of quantitative ultrasound measures of the biceps and supraspinatus tendons. Furthermore, this study also aimed to translate these results into recommendations for the development of a time-efficient and reliable quantitative ultrasound measurement protocol. This study represents the first investigation into the psychometric properties of quantitative ultrasound measures in the biceps and supraspinatus tendon.

Inter-evaluator Repeatability

As expected, inter-evaluator repeatability was generally low, which is in agreement with previous studies that suggest that ultrasound is an operator-dependent modality. Brushoj et al. (14) reported significant differences in Achilles tendon thickness and thickness measurements between observers, although cross-sectional area was statistically similar. No explanation of observer experience is provided. In the current study, biceps and supraspinatus tendon thickness measurements showed good dependability ($\Phi>0.75$) between evaluators who had approximately the same level of experience and followed a standardized protocol. However, other quantitative ultrasound measures exhibited only moderate or poor dependability. Ideally, research applications that seek to make comparisons between subjects should ensure that a single examiner performs all ultrasound scans.

Since evaluator error can easily be eliminated, it is desirable to quantify reliability assuming that only one evaluator conducted the ultrasound examinations. Using data from both evaluators, we also examined inter-preparation and inter-image dependability. Inter-preparation Φ describes the reliability of a single measurement compared to a hypothetical true score. We have developed an external reference marker and standardized positioning protocol, which essentially allows the preparation to be kept constant. Therefore, inter-image Φ describes reliability of two measurements made under the same preparation. For example, images could

be collected pre- and post-intervention while the reference marker remained in place. The development of a reference marker and standardized positioning protocol improves reliability and will give increased power to detect acute changes occurring within a tendon. All quantitative ultrasound measures exhibited moderate or good inter-preparation dependability ($\Phi > 0.5$) and 17 of the 18 ultrasound measures exhibited a similar level of inter-image dependability. It should be noted that the estimate of reliability presented in this study are more conservative than other measures of reliability, including ICC. Inter-evaluator Φ describes variability between two evaluators taking a single image during a single preparation, as opposed to comparing averaged data from all images and preparations. Similarly inter-preparation and inter-image reliability are computed for one evaluator, taking a single image during one preparation. The variance and measurement error is estimated using data from two evaluators capturing two images during each of two preparations.

Effect of Study Design

For all study designs presented in Table 2, greyscale variance, skewness, and kurtosis demonstrated the lowest reliability while tendon thickness and echogenicity were the most repeatable. First order statistics may be more sensitive to small changes in image appearance since no averaging is performed during computation. Ultrasound measures calculated as averages would be less sensitive to changes at the edge of the region of interest, which is most likely to be affected by small changes in probe tilt or orientation.

The values of SEM_{norm} listed in Table 2 provide a guideline for interpreting changes within a single subject as real or due to measurement error. Minimum detectable change (MDC) is linearly related to SEM and can be calculated as $1.65 * \sqrt{2} * SEM$ where 1.65 represents the two-sided tabled z value for the 90% confidence interval and $\sqrt{2}$ accounts for the variance of the measurements to be compared that were recorded at two distinct points in time. Therefore within a single individual, observed changes greater than the MDC can be considered significantly different. This may be useful in clinical applications tracking a single patient's progress or to stratify research subjects into groups based on who experienced significant change. MDC may be too conservative when examining differences between groups where sample variance may be a better indicator of the confidence interval for the observed mean. Due to the limited application of ultrasound to study acute musculoskeletal changes, it is difficult to know if these quantitative ultrasound measures are sensitive enough to detect tendon changes in response to an intervention. Research needs to investigate the responsiveness of tendons to physical activity.

Reliability estimates of two hypothetical experimental situations, averaging either two images or two readings, are presented in Table 2. Performing a second reading does not provide a meaningful improvement in reliability. Capturing a second image provides a marginal increase in the dependability coefficient and reduction in SEM. Therefore, when making comparisons within a single individual, capturing more than one image at each time point may provide a quantitative ultrasound value closer to a hypothetical true score. We believe that we limited error by using an external reference marker. If a protocol involves repeated measurements, we recommend using a reference marker that would remain in place as described in this study. This would reduce the source of variation due to preparation between ultrasound images captured before and after an intervention.

Ultrasound reliability studies have been primarily limited to tendon thickness or cross-sectional area. Brushoj et al. (14) reported within observer limits of agreement for Achilles tendon cross sectional area to be $\pm 1.25\text{mm}$ (19%). Achilles tendon diameter in the sagittal (thickness) and frontal (width) planes, calculated as the mean of two measurements, demonstrated within observer agreement of 0.6mm (13%) and 2.09mm (12%) respectively. O'Connor et al. (15) reported a 90% confidence interval of $\pm 23\%$ for supraspinatus thickness as measured by an

experienced examiner. In our study, the observed measurement error was only 6.7% for biceps tendon thickness and 4.5% for the supraspinatus tendon. Nielsen et al. (18) reported first-order greyscale statistics of the supraspinatus muscle on two different days in 8 subjects. Although specific values are not presented, the authors note that no statistically significant differences were found between the two different days for any of the first-order greyscale statistics. Due to the lack of detail, it is difficult to make direct comparison to the current study.

Limitations

The results of the current study are specifically based on a relatively small sample of subjects (N=20). As with any reliability study, these results describe context-specific reliability and can only serve as guidelines to other researchers. Five manual wheelchair users, along with 15 able-bodied subjects, were studied because of future applications to investigating upper limb injury development and prevention in wheelchair users. Subjects were not screened for shoulder pain or injury prior to participation as future studies will include individuals with both healthy and degenerated tendons. Varying levels of tendon health were informally observed among the subjects in this study. Healthy tendons often have better-defined borders and the collagen pattern is more easily visualized. Therefore, a reliability study of individuals with healthy tendons may result in inflated reliability estimates that would not translate to tendons with tendinopathy which may be more difficult to image.

Additionally, anisotropy can impact tendon appearance as measured by quantitative ultrasound. The scanning protocols were developed to minimize the effects of anisotropy. In particular, the biceps tendon was imaged along the longitudinal axis, which optimized imaging of the horizontal collagen pattern within the tendon. In this view, the tendon region of interest is easily viewed perpendicular to the ultrasound waves. The supraspinatus tendon was imaged in the transverse direction, such that the region of interest was located at the top of the curved humerus with a bright reflection from the underlying cortical surface. It is possible to achieve good reliability even if images are affected by anisotropy. Future studies should relate quantitative ultrasound measures to clinically graded tendon health in order to establish the validity of these features. This will also provide a gauge as to the influence of anisotropy and measurement error on quantitative ultrasound measures. If image analysis can be automated, quantitative measures of tendinosis may be useful in a clinical setting, in addition to potential research applications.

CONCLUSIONS

Quantitative ultrasound is a promising tool for quantitatively evaluating tendon appearance. Although the measured reliability for most outcomes was lower than we hypothesized ($\Phi > 0.75$), we are encouraged that most quantitative ultrasound measures exhibit at least moderate ($\Phi > 0.50$) reliability when images are captured by a single evaluator. We have developed an external reference marker and a subject positioning protocol to reduce the error related to multiple preparations. This minimizes the measurement error of within-subject repeated measurement study designs. We have described normalized standard error of measurement (SEM_{norm}) for three protocols which can serve as a guideline for interpreting results within an individual. Intra-rater reliability was greater than inter-rater reliability and therefore it is recommended that a single examiner perform ultrasound examinations, particularly if multiple exams are being performed for each individual. First-order statistics seems to be more susceptible to error than tendon thickness and echogenicity and therefore extra caution should be used when interpreting these parameters. The current results provide evidence that an appropriately designed protocol will allow quantitative ultrasound to be used as an outcome measure to identify early changes in the tendon.

Acknowledgments

Special thanks are extended to Denis Gravel, pht, PhD, School of Rehabilitation, University of Montreal for his assistance with the Generalizability Theory and with the validation of all results reported in the current manuscript. This material is based upon work supported by the Office of Research and Development, Rehabilitation Research & Development Service, Department of Veterans Affairs, Grant# B3142C, the National Institute of Health, Grant# R21HD054529, the National Science Foundation (NSF) Rehabilitation Engineering Research Center on Spinal Cord Injury, Grant# H133E070024, and an NSF graduate research fellowship. Dany Gagnon held a post-doctoral scholarship from the Fonds de la recherche en santé du Québec at the time of the study.

This material is based upon work supported by the Office of Research and Development, Rehabilitation Research & Development Service, Department of Veterans Affairs, Grant# B3142C, the National Institute of Health, Grant# R21HD054529, the National Science Foundation (NSF) Rehabilitation Engineering Research Center on Spinal Cord Injury, Grant# H133E070024, and an NSF graduate research fellowship.

REFERENCE LIST

1. Arslan G, Apaydin A, Kabaalioglu A, Sindel T, Luleci E. Sonographically detected subacromial/subdeltoid bursal effusion and biceps tendon sheath fluid: reliable signs of rotator cuff tear? *J Clin Ultrasound* 1999;27(6):335–9. [PubMed: 10395129]
2. Thain L, Adler R. Sonography of the rotator cuff and biceps tendon: Technique, normal anatomy, and pathology. *J Clin Ultrasound* 1999;27(8):446–58. [PubMed: 10477887]
3. Zanetti M, Hodler J. Imaging of degenerative and posttraumatic disease in the shoulder joint with ultrasound. *Eur J Radiol* 2000;35:119–25. [PubMed: 10963917]
4. Teefey S, Middleton W, Yamaguchi K. Musculoskeletal Ultrasound: Shoulder Sonography: State of the Art. *Radiol Clin North Am* 1999 Jul;37(4):767–85. [PubMed: 10442080]
5. Wallny T, Wagner U, Prange S, Schmitt O, Reich H. Evaluation of chronic tears of the rotator cuff by ultrasound: A NEW INDEX. *J Bone Joint Surg Br* 1999 Jul;81(B4):675–8. [PubMed: 10463744]
6. Hashimoto B, Kramer D, Wiitala L. Applications of musculoskeletal sonography. *J Clin Ultrasound* 1999;27(6):293–318. [PubMed: 10395126]
7. Leung JLY, Griffith JF. Sonography of Chronic Achilles Tendinopathy: A Case-Control Study. *J Clin Ultrasound* Jan 2008;36(1):27–32.
8. Alvarenga A, Pereira WCA, Infantsi AFC. Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images. *Med Phys* Feb 2007;34(2):379–387.
9. Nielsen PK, Jensen BR, Darvann T, Jorgensen K, Bakke M. Quantitative ultrasound tissue characterization in shoulder and thigh muscles—a new approach. *BMC Musculoskelet Disord* 2006;7:2. [PubMed: 16420695]
10. Sharma N, Ray AK, Sharma S, Shukla KK, Pradhan S, Aggarwal LM. Segmentation and classification of medical images using texture-primitive features: Application of BAM-type artificial neural network. *J Med Phys* 2008;33(3):119–26.
11. Scholten RR, Pillen S, Verrips A, Zwarts MJ. Quantitative ultrasonography of skeletal muscles in children: normal values. *Muscle Nerve* 2003 Jun;27(6):693–8. [PubMed: 12766980]
12. Pillen S, Scholten RR, Zwarts MJ, Verrips A. Quantitative skeletal muscle ultrasonography in children with suspected neuromuscular disease. *Muscle Nerve* 2003 Jun;27(6):699–705. [PubMed: 12766981]
13. Bashford GR, Tomsen N, Arya S, Burnfield JM, Kulig K. Tendinopathy discrimination by use of spatial frequency parameters in ultrasound B-mode images. *IEEE Trans Med Imag* May 2008;27(5):608–15.
14. Brushoj C, Henriksen BM, Albrecht-Beste E, Holmich P, Larsen K, Bachmann Nielsen M. Reproducibility of ultrasound and magnetic resonance imaging measurements of tendon size. *Acta Radiologica* 2006;9:954–9. [PubMed: 17077048]
15. O'Connor PJ, Grainger AJ, Morgan SR, Smith KL, Waterton JC, Nash AFP. Ultrasound assessment of tendons in asymptomatic volunteers: a study of reproducibility. *Eur Radiol* 2004;14:1968–73. [PubMed: 15309497]

16. Ying M, Yeung E, Li B, Li W, Lui M, Tsoi C-W. Sonographic evaluation of the size of the achilles tendon: the effect of exercise and dominance of the ankle. *Ultrasound Med Biol* 2003;29(5):637–42. [PubMed: 12754062]
17. Haralick R, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Trans Syst Man Cybern* 1993 Nov;SMC-3(6):610–21.
18. Nielsen PK, Jensen BR, Darvann T, Jorgensen K, Bakke M. Quantitative ultrasound image analysis of the supraspinatus muscle. *Clin Biomech* 2000;15(Suppl 1):S13–S16.
19. MATLAB Reference Documentation. The MathWorks, Inc; Version 2008a
20. Shavelson, RJ.; Webb, NM. Generalizability theory: A primer. Newbury Park, CA: Sage; 1991.
21. Brennan R, Kane M. An index of dependability for mastery tests. *J Education and Management* 1977;14:277–289.
22. Portney, LG.; Watkins, MP. Foundations of clinical research: Application to practice. Vol. 2. Stamford: Appleton & Lange; 2000.

Figure 1.
Subject positioning for imaging of the biceps (A) and supraspinatus (B) tendons.

Figure 2. Transducer and reference marker position relative to the long head of the biceps tendon (LHBT). The crossbars create an interference pattern visible on the ultrasound image.

Figure 3.

Marker interference pattern and region of interest of the long head of the biceps tendon and supraspinatus tendon. d_{constant} was defined for each subject as the distance from the midpoint of the interference pattern to the midpoint of the tendon for the first analyzed image.

Table 1

Mean (SD) quantitative ultrasound outcome measures computed for each evaluator. The coefficient of dependability (Φ) describes overall reliability derived from a random D-study model (E=1; P=1; I=1) and intra-evaluator, inter-preparation, and inter-image reliability, computed mixed D-study models (E=1; P=1; I=1).

Ultrasound Outcome Measures	Tendon	Evaluator #1		Evaluator #2		Overall Φ^*	Inter-evaluator $\Phi^\dagger$	Inter-preparation $\Phi^\ddagger$	Inter-image $\Phi^\S$
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)				
Thickness	Biceps	4.02 (0.86)	4.03 (0.89)			.804	.819	.872	.941
	Supraspinatus	4.87 (0.78)	4.78 (0.92)			.804	.814	.908	.962
Echogenicity	Biceps	114.33 (18.85)	103.84 (17.66)			.396	.397	.686	.836
	Supraspinatus	105.97 (18.08)	111.04 (23.12)			.605	.659	.771	.829
Variance	Biceps	2052.4 (669.7)	2064.8 (713.6)			.310	.330	.660	.806
	Supraspinatus	1306.4 (359.1)	1536.7 (575.8)			.368	.402	.631	.728
Skewness	Biceps	0.56 (0.32)	0.82 (0.34)			.261	.355	.569	.677
	Supraspinatus	0.17 (0.33)	0.12 (0.36)			.339	.360	.551	.678
Kurtosis	Biceps	3.66 (0.97)	4.03 (1.27)			.385	.453	.664	.684
	Supraspinatus	3.66 (0.47)	3.47 (0.52)			.252	.259	.535	.463
Entropy	Biceps	7.11 (0.27)	7.06 (0.35)			.312	.312	.636	.768
	Supraspinatus	6.88 (0.22)	6.99 (0.26)			.335	.376	.528	.669
Contrast	Biceps	6.02 (1.72)	5.86 (2.08)			.328	.348	.625	.707
	Supraspinatus	4.08 (0.92)	4.35 (1.17)			.504	.589	.666	.700
Energy	Biceps	0.58 (0.24)	0.68 (0.29)			.270	.297	.752	.826
	Supraspinatus	0.71 (0.22)	0.65 (0.21)			.506	.518	.649	.629
Homogeneity	Biceps	3.46 (0.21)	3.54 (0.24)			.360	.400	.698	.825
	Supraspinatus	3.65 (0.18)	3.61 (0.20)			.581	.655	.723	.736

* Overall reliability for a study design with E=1, P=1, I=1 computed with a random D-study model that includes all possible sources of variation

† Inter-evaluator reliability computed with a mixed D-study design with fixed facets (P and I)

‡ Inter-preparation reliability computed with a mixed D-study model with fixed facets (E and I)

§ Inter-image reliability computed with a mixed D-study model with fixed facets (E and P)

Table 2
Dependability coefficient (Φ), standard error of measurement (SEM) calculated for various D-study designs using a single value (design: P=1; I=1; R=1) or a mean value (all other designs) obtained by Evaluator #1.

Outcome Measures	D-study measurement error estimations for multiple study designs											
	Biceps Tendon			Supraspinatus Tendon			Biceps Tendon			Supraspinatus Tendon		
	P=1	I=1	R=1	P=1	I=1	R=1	P=1	I=1	R=1	P=1	I=1	R=1
Thickness												
Φ	0.906			0.910			0.919			0.921		
SEM	0.271			0.265			0.250			0.221		
SEM _{norm} (%)	6.727			6.582			6.206			4.547		
Echogenicity												
Φ	0.742			0.743			0.793			0.754		
SEM	10.610			10.591			9.190			9.985		
SEM _{norm} (%)	9.288			9.272			8.046			9.405		
Variance												
Φ	0.614			0.616			0.673			0.474		
SEM	480.747			478.994			423.072			328.352		
SEM _{norm} (%)	23.442			23.357			20.630			25.382		
Skewness												
Φ	0.533			0.535			0.677			0.579		
SEM	0.265			0.264			0.196			0.260		
SEM _{norm} (%)	47.302			47.097			34.917			163.133		
Kurtosis												
Φ	0.462			0.463			0.568			0.477		
SEM	0.887			0.886			0.716			0.443		
SEM _{norm} (%)	24.214			24.180			19.559			12.035		
Entropy												
Φ	0.616			0.620			0.681			0.484		
SEM	0.196			0.194			0.170			0.198		
SEM _{norm} (%)	2.756			2.731			2.389			0.484		
Contrast												
Φ	0.646			0.651			0.708			0.589		
SEM	1.162			1.148			1.006			0.725		
SEM _{norm} (%)	19.260			19.033			16.674			17.655		
Energy												
Φ	0.709			0.714			0.796			0.618		
SEM	0.150			0.148			0.118			0.162		
SEM _{norm} (%)	25.884			25.551			20.470			22.973		
Homogeneity												
Φ	0.764			0.768			0.826			0.657		
SEM	0.113			0.112			0.094			0.124		
SEM _{norm} (%)	3.272			3.237			2.704			3.392		

P= Number of time markers were affixed to skin and upper limb positioned

I= Number of ultrasound image recorded

R= Number of reading completed for each ultrasound image recorded