Ultrasonographic measurement of patellar tendon thickness—a study of intra- and interobserver reliability

Soren T. Skou a,b,⁎, Johanne M. Aalkjaer c

a Orthopaedic Surgery Research Unit, Aalborg University Hospital Denmark
b Department of Health Science and Technology, Centre for Sensory-Motor Interaction, Aalborg University, Denmark
c University College of Northern Denmark, Aalborg, Denmark

Abstract

This article determines intra- and intertester reliability and measurement precision of ultrasonographic assessment of patellar tendon (PT) thickness. Eighteen healthy subjects were scanned three times in two sessions with a 45-min interval by two experienced examiners. Reliability and measurement precision were evaluated using intraclass correlation coefficient and limits of agreement. When measuring PT thickness, a mean of two measurements is recommended, and changes larger than 0.07 cm can be considered actual changes and not a result of measurement uncertainty.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Musculoskeletal ultrasonography (US) is a widely used, noninvasive, inexpensive, and sensitive tool for assessing soft tissue [1] that has developed to become part of a gold standard in the examination of tendons, including the patellar tendon (PT) [1,2].

Patellar tendinopathy is a prevalent condition that often causes prolonged disability [3]. The clinical diagnosis of patellar tendinopathy can be supported by US [4], since tendons with pathology appears more disorganized, hypoechogetic, and thicker than healthy tendons on US [5–7]. Furthermore, tendon thickness could be used when evaluating the effect of a given treatment protocol since a reduction in PT thickness on US has been observed following successful treatment of patellar tendinopathy [8] and since a correlation has been reported between a decrease in tendon thickness and a reduction in pain [9].

However, operator dependency is regarded as a major shortcoming in the use of US, and handling of the transducer and settings of the US machine could potentially influence the reliability of the measurement of the PT thickness [10]. Therefore, it is important to know precision and intra- and intertester reliability of the method because it would give the examiner the possibility to recognize if differences in measurements of PT thickness are due to actual changes or to measurement uncertainty. To the authors’ knowledge, this has not been investigated regarding US measurement of PT thickness when measuring the tendon once or more.

The purpose of this study was to determine intra- and intertester reliability and measurement precision of US assessment of PT thickness using one or the mean of multiple measurements.

2. Materials and methods

2.1. Participants

Eighteen healthy volunteers (9 women and 9 men, mean body mass index (S.D.) of 24.5 (2.6), mean age (S.D.) of 25.1 (3.1) without lower extremity symptoms participated. Exclusion criteria were current or prior pain or injury in PT or prior history of surgery in the lower extremity. Only the left PT of the participants was used in the statistical analysis in the study, since association between the left and right side of the same participant is expected [11,12]. Even though frequently applied, using both the left and right side from the same participant would violate the assumption of the statistical analysis [11,12].

The study was approved by the Danish Data Protection Agency and conducted in accordance with the Helsinki Declaration. Informed consent was obtained from all the participants on forms provided by the local Ethics Committee of The North Denmark Region.
2.2. Examiners

The US measurement was carried out by two experienced examiners using the US scanner applied in this study on a daily basis, thereby eliminating a possibly learning effect during the study. Examiner 1 had 6 years of experience using US in daily practice, while Examiner 2 had 2 years of experience using US. Both examiners had lower extremity disorders as their speciality and examined several knee patients each day using US. None of the examiners were involved in the study, and they were instructed and trained in the test setup before the start of the study to ensure that they followed the study protocol and to reduce the risk of bias.

2.3. Study protocol

US was performed using a SonoSite S-MSK scanner (SonoSite, Inc., Bothell, WA, USA) with a HFL38x 13–6 MHz linear transducer.

The measurements were carried out using the recommendations from the European Society of Musculoskeletal Radiology [13]. The participants were positioned in a supine position on an examination couch; the knee was kept in a flexion of 30° obtained by placing a pillow underneath the popliteal space, which stretches the extensor mechanism and thereby a possible anisotropy related to the concave profile that the quadriceps and PTs assume in full extension is avoided [13].

The transducer was placed centrally on the PT in a longitudinal direction (Fig. 1). To avoid affecting the underlying tissue and thereby possibly affecting the US measurement, care was taken not to apply unnecessary pressure with the transducer. To standardize the measurements, the settings of the US scanner were kept constant during all measurements. The scan depth was set to 1.8 cm, and three successive measurements of each left PT were made. Between the three measurements, the transducer was moved away from the knee and then repositioned. The measurement of PT thickness was made using the built-in software of the scanner 1 cm from apex patella, since the thickness of PT is more even at this location (Fig. 2).

The schedule for US measurement was randomly generated in blocks of four participants (the last block had six), and it was randomly decided whether Examiner 1 or 2 did the first measurement in both sessions. To blind the examiners, only the left knee of the participants was visible due to a cloth screen placed between the lower and upper body and a white cloth placed over the legs of the participants. To maintain the blinding, the examiner left the room between each participant while a new participant was positioned. Given the fact that the participants had similar skin color, this setup made the identification of the participants impossible for the examiners. Both Examiners 1 and 2 were blinded to the measurements obtained by the other examiner. To examine a possible learning effect during the measurements (i.e., increasing reliability), the randomly generated blocks of participants received a consecutive identification number.

After approximately 45 min, the participants underwent a second session of US measurement of the PT thickness using the same protocol as in the first session of US measurement. Care was taken to replicate the setup of the first session of US measurement to ensure the comparability between the sessions. The relative short time between the first and second session of US measurement was chosen, since time-of-day and physical activity might affect PT thickness [14–18]. The participants were asked to refrain from physical activity between the first and second session of US measurement.

2.4. Statistical analysis

Data were normally distributed, confirmed by visual inspection of Q-Q plots. All analyses were done using IBM SPSS Statistics Version 19.0. Only the left PT was used in the analyses, and every sequential measurement of PT thickness was treated as a separate variable. Paired t test was applied to test for differences in PT thickness between the first and second session for Examiners 1 and 2. Intratester reliability for Examiner 1 and Examiner 2, respectively, was evaluated using a two-way mixed model, consistency type intraclass correlation coefficient (ICC). The reliability of a single US measurement was evaluated by using the first measurement from the first session and the first measurement in the second session and the “single measure” output from SPSS. The reliability when using the mean of two measurements was evaluated by using the first and second measurement from the first and second session and the “average measures” output from SPSS. The reliability when using the mean of three measurements was evaluated by using the first, second, and third measurement from the first and second sessions and the “average measure” output from SPSS.

Fig. 1. The bar on the image indicates were the transducer was placed on the PT (colours are not required in figure 1).

Fig. 2. Longitudinal scan of the PT thickness (indicated by “B”) 1 cm from apex patella.
Intertester reliability was evaluated using a parallel model two-way random effects, absolute agreement, for the first measurement (single measure output in SPSS) and for the mean of two and three measurement (average measures output in SPSS) from the first session. LOA values can be classified as low: 0.20–0.49; moderate: 0.50–0.69; high: 0.70–0.89; or very high: 0.90–1.00 [19].

Measurement precision was evaluated using 95% limits of agreement (LOA) and LOA as a percentage of the mean (LOA %). LOA is presented as the difference between the mean difference and the upper and lower LOA to make the result transferable to a clinical context.

3. Results

The results from the US measurement for the left PT are shown in Table 1.

Paired t test showed no significant differences in PT thickness between the first and second session for Examiners 1 and 2. No consistent signs of learning effect were found for neither Examiner 1 nor 2 when plotting the difference in thickness between the first and second session against the consecutive identification number of the randomly generated blocks of participants.

3.1. Intrasession reliability

The ICC with 95% confidence interval (95% CI) for Examiner 1 using one measurement was 0.84 (0.63–0.94). It increased to 0.94 (0.84–0.98) when using the mean of two measurements but not further when using the mean of three measurements (0.95 (0.86–0.98)).

The ICC with 95% CI for Examiner 2 using one measurement was 0.70 (0.35–0.88). It increased to 0.89 (0.71–0.96) when using the mean of two measurements but not further when using the mean of three measurements (0.89 (0.71–0.96)) (Table 2).

3.2. Intertester reliability

The ICC with 95% CI for Examiners 1 and 2 using one measurement was 0.70 (0.18–0.89). It increased to 0.78 (0.23–0.93) when using the mean of two measurements and that using the mean of three measurements (0.75 (0.18–0.92)) (Table 2).

3.3. Measurement precision

Bland–Altman plots were inspected visually and indicated no systematic difference over the range of measurements.

LOA (%) for the intratester reliability was 0.08 cm (21.6%) for Examiner 1 and 0.09 cm (26.1%) for Examiner 2 using one measurement. It decreased to 0.07 cm (18.3%) for Examiner 1 and 0.07 cm (21.1%) for Examiner 2 when using the mean of two measurements but not further for neither Examiner 1 (0.07 cm (17.6%)) nor Examiner 2 (0.07 cm (21.4%)) when using the mean of three measurements (Table 3).

LOA (%) for the intertester reliability was 0.09 cm (24.7%) using one measurement, 0.10 cm (27.0%) using the mean of two measurements, and 0.10 cm (28.9%) using the mean of three measurements (Table 3).

### Table 1

<table>
<thead>
<tr>
<th>Thickness of PT in cm</th>
<th>Men (n=9) Mean (SD)</th>
<th>Women (n=9) Mean (SD)</th>
<th>Total Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examiner 1</td>
<td>0.42 (0.08)</td>
<td>0.35 (0.06)</td>
<td>0.38 (0.07)</td>
</tr>
<tr>
<td>Examiner 2</td>
<td>0.37 (0.06)</td>
<td>0.32 (0.05)</td>
<td>0.34 (0.05)</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Inter- and intrasession reliability</th>
<th>ICC One measurement ICC (95% CI)</th>
<th>Mean of two measurement ICC (95% CI)</th>
<th>Mean of three measurement ICC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intratester, Examiner 1</td>
<td>0.84 (0.63–0.94)</td>
<td>0.94 (0.84–0.98)</td>
<td>0.95 (0.86–0.98)</td>
</tr>
<tr>
<td>Intratester, Examiner 2</td>
<td>0.70 (0.35–0.88)</td>
<td>0.89 (0.71–0.96)</td>
<td>0.89 (0.71–0.96)</td>
</tr>
<tr>
<td>Intertester</td>
<td>0.70 (0.18–0.89)</td>
<td>0.78 (0.23–0.93)</td>
<td>0.75 (0.18–0.92)</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Results from the ultrasonographic measurement of the left PT thickness</th>
<th>LOA [cm]</th>
<th>One measurement LOA %</th>
<th>Mean of two measurements LOA %</th>
<th>Mean of three measurements LOA %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examiner 1</td>
<td>0.08</td>
<td>21.6</td>
<td>18.3</td>
<td>17.6</td>
</tr>
<tr>
<td>Examiner 2</td>
<td>0.09</td>
<td>26.1</td>
<td>21.1</td>
<td>21.4</td>
</tr>
<tr>
<td>Intertester</td>
<td>0.09</td>
<td>24.7</td>
<td>27.0</td>
<td>28.9</td>
</tr>
</tbody>
</table>

*LOA is presented as the difference between the mean difference and the upper and lower LOA.

4. Discussion

This was the first study evaluating intra- and intertester reliability and measurement precision of US measurement of PT thickness applying more than one measurement of the tendon. The study found that using a mean of two measurements of the PT thickness increased the intratester (from ICC 0.84 and 0.70 to 0.94 and 0.89) and intertester reliability (from ICC 0.70 to 0.78) while the addition of a third measurement did not increase the reliability further. Furthermore, this study showed that measurement precision for the intratester reliability increased to 18.3% for Examiner 1 and to 21.1% for Examiner 2 when using the mean of two measurements but not further when using the mean of three measurements and that measurement precision for the intertester reliability was 24.7% for one measurement and decreased to 27.0% for the mean of two and to 28.5% for the mean of three measurements.

The findings in the present study is in agreement with intratester (ICC 0.67–0.97) [20–28] and intertester (ICC 0.76–0.96) [24–29] reliability and measurement precision (LOA or the comparable measure, standard error of measurement (S.E.M.) [30]) of intratester (1.8–53%) [21,22,24–27,31–33], and intertester (15.8–49%) [24–27,34] reliability previously reported regarding US measurement of tendon and muscle thickness.

Black et al. [29] found correlation coefficients of US measurement of hypoechoic areas in patellar tendinopathy (Pearson r =0.87), which are seemingly comparable to the findings in the present study. However, since Pearson correlation coefficient is not an assessment of agreement but only of the association between the measurements of the two examiners [35,36], a direct comparison between the findings would be incorrect. O’Connor et al. [31] examined the variability in PT diameter in healthy participants and found an LOA % of 27% for the intertester reliability, which is consistent with the finding in the present study. O’Connor et al. also found an LOA% of 24 and 32% for the intratester reliability, which is comparable to the measurement precision for one measurement in the present study. However, O’Connor et al. did not evaluate reliability, and the study only evaluated the PT once, not giving the possibility to evaluate measurement precision of more than one measurement.

Since the reliability and the measurement precision improved when using the mean of two measurements of the PT thickness in the present study, it could be recommended to use the mean of two measurements when evaluating PT thickness. Previous studies have
recommended the use of a mean of three measurements when measuring the thickness of tendons and muscles [22,24]. Rathleff et al. [22] found that S.E.M. decreased when using the mean of two and even more when using the mean of three US measurements of the plantar fascia (PF) thickness. Koppenhaver et al. [22] found that S.E.M. decreased when using the mean of two and three measurements but not further when using the mean of four measurements of the transversus abdominis (TrA) and lumbar multifidus (LM) muscles. An explanation for the nonexistent improvement in reliability and measurement precision when measuring the PT three times instead of two could be the fact that the PF, the TrA, and the LM are deeper structures than the PT, thereby probably more influenced by sonographic attenuation of the overlying soft tissue affecting the quality of the US imaging.

Put into clinical context, the findings of the present study could be used when evaluating the effects of a treatment protocol. Fredberg et al. [8] found that the diameter of the PT was significantly reduced after 1 week (reduction of 0.09 cm), 3 weeks (reduction of 0.17 cm), and 6 months (reduction of 0.14 cm) following a steroid injection. Since only changes of above 0.09 cm can be considered actual changes when only one measurement of the PT is performed, it cannot be ruled out that the effects after 1 week was measurement uncertainty while the effects after 3 weeks and 6 months can be considered actual changes. Even though both examiners were experienced US examiners, their intraterest reliability measurement precision differed. Examiner 1 was the most experienced examiner and had the highest reliability and measurement precision. Previous studies [24,27] have also found differences in reliability comparing examiners with different degrees of experience in US assessment. This highlights the importance of standardization of and continuous training in US assessment [10].

The study is limited by its sample size since the width of the confidence intervals may reflect the relative small sample. However, by construction, the reported estimates are optimal in light of the sample.

In conclusion, US measurement of PT thickness is a reliable method. When measuring PT thickness, a mean of two measurements is recommended. If the same examiner measures the same PT twice, changes larger than 0.07 cm can be considered actual changes and not a result of measurement uncertainty, while changes larger than 0.10 cm can be considered actual changes when two different examiners measures the same PT.

References