ARNOLD Y.L. WONG, PT, MPhil[1] • **ERIC PARENT,** PT, PhD[1,2] • **GREG KAWCHUK,** DC, PhD[1]

Reliability of 2 Ultrasonic Imaging Analysis Methods in Quantifying Lumbar Multifidus Thickness

SUPPLEMENTAL VIDEO ONLINE

The lumbar multifidus (LM) muscle plays a crucial role in maintaining segmental lumbar stability and preventing recurrence of low back pain (LBP).[17,27,42] Studies have shown that patients with acute or chronic LBP have localized LM atrophy and reduced voluntary LM muscle contraction ipsilateral to the side of symptoms.[14,17,38] Unfortunately, the resolution of LBP is not accompanied with spontaneous recovery in the function and morphology of the LM muscle, unless appropriate exercises are prescribed.[16,27] To evaluate the function of the LM muscle and to monitor the progress of treatment, a cost-effective, noninvasive clinical assessment tool is needed.

Brightness-mode (B-mode) rehabilitative ultrasound imaging (RUSI) has been advocated as a valid noninvasive tool to evaluate the size and dynamic behavior of trunk muscles[40] and to provide real-time feedback to LBP patients during motor control training.[24,35] Research has demonstrated measurement agreement of LM muscle thickness by RUSI and by magnetic resonance imaging.[15] Indwelling electromyography studies have also validated the correlation between the myoelectric activity of the LM and the change in geometry of the LM as measured by RUSI.[21,36] Clinically, the role of RUSI biofeedback in facilitating LM muscle recruitment in patients with LBP during back-stabilizing exercises has been demonstrated.[13,16,24,35,40]

To measure LM muscle thickness

STUDY DESIGN: Reliability study.

OBJECTIVES: To compare the within- and between-day intrarater reliability of rehabilitative ultrasound imaging (RUSI) using static images (static RUSI) and video clips (video RUSI) to quantify multifidus muscle thickness at rest and while contracted. Secondary objectives were to compare the measurement precision of averaging multiple measures and to estimate reliability in individuals with and without low back pain (LBP).

BACKGROUND: Although intrarater reliability of static RUSI in measuring multifidus thickness has been established, using video RUSI may improve reliability estimates, as it allows examiners to select the optimal image from a video clip. Further, multiple measurements and LBP status may affect RUSI reliability estimates.

METHODS: Static RUSI and video RUSI were used to quantify multifidus muscle thickness at rest and during contraction and percent thickness change in 27 volunteers (13 without LBP and 14 with LBP). Three static RUSI images and 3 video RUSI video clips were collected in each of 2 sessions 1 to 4 days apart. Reliability and precision were assessed using intraclass correlation coefficients, standard error of measurement, minimal detectable change, bias, and 95% limits of agreement.

RESULTS: Using an average of 2 measures yielded optimal measurement precision for static RUSI and video RUSI. Based on the average of 2 measures obtained under the same circumstance, there was no significant difference in the reliability estimates between static RUSI and video RUSI across all testing conditions. Reliability point estimates (intraclass correlation coefficient model 3,2) of multifidus thickness were 0.99 for within-day comparisons and ranged from 0.93 to 0.98 for between-day comparisons. The within- and between-day intraclass correlation coefficients (model 3,2) of percent thickness change ranged from 0.97 to 0.99 and from 0.80 to 0.90, respectively. The exploratory analysis showed no significant difference in the reliability estimates between asymptomatic and LBP participants across most testing conditions.

CONCLUSION: Both RUSI methods yielded high reliability estimates for multifidus muscle measurements. Using an average of 2 measures obtained optimal measurement precision. Overall, video RUSI is a reliable surrogate for static RUSI for multifidus muscle measurements and has the additional advantage of requiring shorter data collection time. *J Orthop Sports Phys Ther 2013;43(4):251-262. Epub 7 December 2012. doi:10.2519/jospt.2013.4478*

KEY WORDS: *low back pain, lumbar multifidus, reproducibility, RUSI, ultrasonography*

[1]Department of Physical Therapy, University of Alberta, Edmonton, Alberta, Canada. [2]Glenrose Rehabilitation Hospital, Edmonton, Alberta, Canada. Arnold Wong is supported by the Alberta Innovates-Health Solutions Graduate Studentship. Greg Kawchuk is supported by the Canadian Research Chair Program. The protocol of this study was approved by the Health Research Ethics Board at the University of Alberta. The authors certify that they have no affiliations with or financial involvement in any organization or entity with a direct financial interest in the subject matter or materials discussed in the manuscript. Address correspondence to Dr Gregory N. Kawchuk, Department of Physical Therapy, University of Alberta, 3-48 Corbett Hall, Edmonton, Alberta, Canada T6G 2G4. E-mail: greg.kawchuk@ualberta.ca

with RUSI, examiners have traditionally had to either manually hold an ultrasound transducer and simultaneously use on-screen calipers to measure real-time dimensions of the LM muscle or save static images of the LM muscle for subsequent offline analysis. This complex procedure requires either a highly skilled examiner[40] or 2 examiners working together to collect data.[23,34] The difficulties of using this technique have now been solved by acquiring ultrasound videos of the LM muscle and holding an ultrasound transducer with an external mechanical arm.[18]

Recording ultrasound videos from the rest state to a contracted state may be well suited for offline analysis. The ability to watch the videos offline may allow the examiner to select better images, which can enhance reliability and decrease time during the patient encounter. Additionally, shorter image-acquisition time may benefit patients who have difficulty with particular testing postures. In the future, B-mode RUSI using video clips (video RUSI) may be used as a surrogate for motion-mode (M-mode) ultrasound imaging to record the onset of muscle activity or fatigue, given its ability to monitor changes in muscle thickness over time.

Previously, Dickx et al[7] adopted video RUSI to evaluate the intrarater reliability of LM thickness measurement in 6 asymptomatic individuals and reported a high reliability point estimate (intraclass correlation coefficient $[ICC]_{1,1}$ = 0.93). However, they did not compare this reliability estimate with that obtained by the conventional RUSI using static images (static RUSI) or specify whether the estimate was obtained in a single session or between 2 sessions.

Given the high reliability of static RUSI in measuring LM thickness,[20,35,37] it is conceivable that video RUSI for LM muscle measurements would also have high reliability. Unfortunately, no study has directly compared the reliability estimates between static RUSI and video RUSI, or between individuals with and without LBP. As LM muscle thickness of an individual is usually measured by the same examiner over time in many clinical and research situations, it is essential to compare the within- and between-day intrarater reliability of LM muscle measurements using video RUSI and static RUSI in individuals with and without LBP. Further, because the measurement precision of LM muscle is improved by multiple measures,[23,25,37] it is imperative to identify the optimal number of measures for clinical practice and research after taking the additional offline measurement effort into account.

Given the above, the primary objective of this study was to compare the within- and between-day intrarater reliability of LM muscle thickness measurements obtained by (1) static RUSI and video RUSI, and (2) a single measure or an average of 2 or 3 measures. The secondary objective was to perform an exploratory analysis to compare the intrarater reliability of LM muscle measurements obtained from participants with and without LBP. We hypothesized that the reliability of LM muscle thickness measurements would not be significantly different between RUSI methods, or between participants with and without LBP. We also hypothesized that there would be an optimal number of measures to improve measurement precision.

METHODS

Participants

Thirty-one volunteers, 18 to 60 years of age, with or without current nonspecific LBP, were recruited by posters, emails, and various listservs on campus. Nonspecific LBP was defined as pain or discomfort between the 12th-rib costal margin and above the gluteal folds, with or without leg pain.[16,31] Participants were excluded if they had prior lumbar surgery, pregnancy, congenital spinal disorders, inability to lie prone for at least 20 minutes, scoliosis with a rib-height difference of greater than 2 cm on forward flexion, spondylolisthesis, a history of severe trauma, current arm and shoulder pain, or medical red flag conditions such as cancer, cauda equina syndrome, spinal infection, fracture, or systemic disease. Participants provided written consent, and the rights of the participants were protected. The study protocol was approved by the Health Research Ethics Board at the University of Alberta.

Examiner

A physiotherapist with 7 years of clinical experience conducted the measurement procedure. The examiner underwent 10 hours of RUSI training and had hands-on experience on 6 individuals prior to the study.

Procedures

The study was a repeated-measures design, involving 2 visits that were 1 to 4 days apart, at the same time of day. After signing consent forms, participants completed self-report measures, including a demographic information sheet. An 11-point numeric pain rating scale (NPRS) was used to measure current pain, and least and worst pain in the last 24 hours.[10,19] A body pain diagram was used to indicate the location and area of current symptoms.[39] The Modified Oswestry Low Back Pain Disability Index (ODI) was used to assess self-perceived disability.[11] Upon completion of the self-report measures, the examiner performed a standardized physical examination to assess the eligibility of participants and to identify the side of each participant that was symptomatic for LBP, which determined the side to be imaged. For asymptomatic participants or participants with symmetrical LBP, the imaged side was chosen randomly.

Participants were prone, with pillows placed under the abdomen to ensure that the lumbosacral junction angle was less than 10°, as measured by an inclinometer.[7,20,21,32,38] Participants were then instructed to perform the contralateral arm lifting task (CALT), to induce automatic submaximal contralateral LM muscle contraction.[21] Participants grasped a handheld weight, then raised

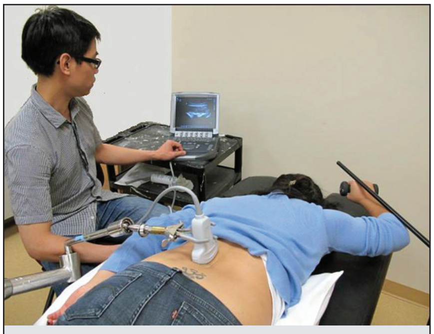

FIGURE 1. Experimental setup of ultrasound imaging.

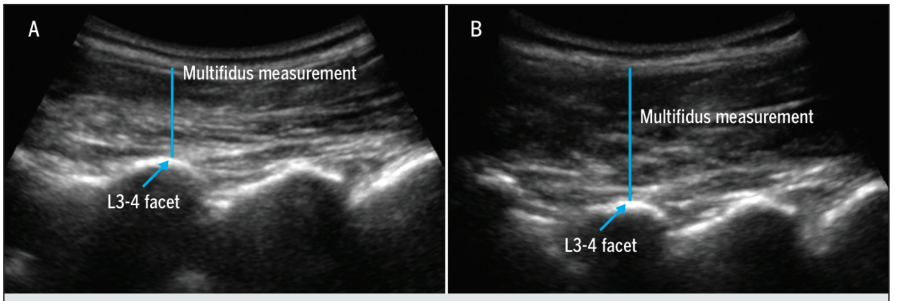

FIGURE 2. Ultrasound images of the multifidus (A) at rest and (B) during contraction.

the loaded arm off the examination table (with shoulders in 120° of abduction and the elbow in 90° of flexion), until their forearm touched a horizontal bar 5 cm above their original arm position (**FIGURE 1**).[18] They were instructed to maintain contact with the bar for 3 seconds. The handheld weight was designed to load the LM muscle to approximately 30% of maximum voluntary isometric contraction[21] according to individual body mass: participants with a body mass of less than 68.2 kg used a 0.68-kg weight, those whose body mass was between 68.2 and 90.9 kg used a 0.9-kg weight, and those with a body mass above 90.9 kg used a 1.36-kg weight. To minimize the effect of respiration on LM muscle thickness in our experimental setup, participants were instructed to hold their breath at the end of normal exhalation during each CALT. The examiner evaluated the participants' breathing throughout the CALT practice and data collection to prevent participants from using a Valsalva maneuver,[18] which may affect lumbar biomechanics.

Data Collection

An M-Turbo ultrasound machine (SonoSite, Inc, Bothell, WA) and a 60-mm, 2- to 5-MHz curvilinear transducer were used to acquire B-mode static images and video clips of the LM muscle. The examiner identified the L3 spinous process of each participant by palpation and marked it for reference. In cases of uncertainty, the spinal level was verified by ultrasound. The examiner then placed the transducer parasagittally along the participant's spine, with the midpoint over the L3 spinous process, sliding it laterally and tilting it medially until the examiner could identify the image of the zygapophyseal joints, multifidus muscle, and thoracolumbar fascia.[37] When the best image appeared on the screen, the transducer was held in position by an adjustable transducer support system on the left side of the participant (**FIGURE 1**) to reduce probe movements during the acquisition of the images at rest and during contraction.[41] The support system consisted of a multiarticulated arm that could be locked into position with compressed air, using a foot switch (SPIDER Limb Positioner; Tenet Medical Engineering, Calgary, Canada).[18] The angle between the transducer and the axis perpendicular to the participant's skin in the coronal plane was measured with a goniometer to guide the placement of the transducer in the second session.

In each session, participants performed 6 repetitions of the CALT, while the examiner captured 3 RUSI video clips and 3 pairs of RUSI static images (3 at rest and 3 during the CALT). Each video clip lasted for 15 seconds and covered the sequential change of the LM muscle from its relaxed to its contracted state and its return to the relaxed state. The sequence of static RUSI and video RUSI acquisition was randomized to minimize the order effect pertinent to fatigue and potential learning of participants. One minute of rest was interspersed between CALTs to avoid muscle fatigue. Participants rated their CALT-related pain intensity at the end of data collection.

Upon completion of data collection in the first session, the examiner placed a transparency sheet on the participant's back and traced the location of the transducer, the participant's body landmarks, and any permanent skin blemishes onto the sheet.[32,40] The mark corresponding to the transducer position was then cut out. For the second session, the transparency sheet was placed onto the participant's back to guide the accurate relocation of the transducer.

Measurements

The RUSI video clips and static images were analyzed offline using ImageJ Version 1.440 software (National Institutes of Health, Bethesda, MD). Prior to the measurement of LM muscle thickness, 4 static images of the LM muscle were selected from each trial video clip: 2 images of the resting LM and 2 images of the contracted LM. The image of the contracted muscle was chosen when the LM muscle reached maximal thickness.[7] A total of 24 static images (12 resting and 12 contracted LM images in 2 sessions) were collected from each of the participant's 6 video clips. All static images of the resting muscle were taken prior to the LM contraction from the same video. These images, together with the images acquired by the static RUSI (6 resting and 6 contracted LM images in 2 sessions), were used for LM muscle measurement.

Because 2 images of the LM were measured from each video clip at each activation state, the average of the 2 LM thickness values at each state was used in reliability analysis. This method was chosen to maximize the information available from video RUSI while minimizing processing time.

The LM thickness at the L3-4 and L4-5 lumbar levels was measured from the tip of the respective zygapophyseal joint to the inside edge of the echogenic connective tissue overlying the LM muscle (**FIGURE 2**, **ONLINE VIDEO**).[37] The ImageJ measurement procedure has been outlined elsewhere.[23] The examiner used an automatic measurement function (control-M) in ImageJ and covered the measurement output on the computer screen to blind the previous and current thickness values during measurement.

Statistical Analyses

Statistical analyses were conducted using SPSS Version 17.0 software (SPSS Inc, Chicago, IL). The baseline characteristics of the 2 groups were compared. Continuous and binary variables were tested by independent *t* tests and chi-square tests, respectively.

ICCs with 95% confidence intervals (CIs) were used to examine the intrarater reliability of LM thickness measurements using the 2 RUSI methods in our participants during a single session and between 2 sessions.[1] The dependent variables for ICC calculation included the resting thickness, contracted thickness, and percent thickness change [$(\text{thickness}_{contracted} - \text{thickness}_{relaxed})/\text{thickness}_{relaxed} \times 100$]. The reliability estimates of each RUSI method across different measurement conditions were calculated by $ICC_{3,k}$, where k = 1, 2, 3. The standard error of measurement (SEM) under each condition was calculated [pooled standard deviation $\times \sqrt{(1 - ICC)}$] to quantify the measurement precision.[8,29] Minimal detectable change at the 95% confidence level (MDC_{95}) was calculated ($1.96 \times \sqrt{2} \times$ SEM) to indicate the minimal change in thickness or in percent thickness change that could be observed with 95% confidence that the observed difference exceeded measurement error.[12] Further, relative MDC_{95} was calculated by dividing the MDC_{95} by its corresponding mean muscle thickness or mean percent thickness change to allow comparisons with other reliability studies. The a priori significance level for all tests was set at .05.

The Bland and Altman[2,4] 95% limits of agreement, bias, and 95% CI for the bias[3] were calculated to quantify the agreement between the 2 RUSI methods in LM thickness measurements.[30] Bias is equivalent to the mean difference in LM thickness measurements, as measured by static RUSI and video RUSI on the same participant. The 95% CI for the bias was calculated from the standard error of the mean difference in LM thickness measurements between the 2 methods.[3] If the 95% CI for the bias included zero, it indicated that there was no statistically significant difference in measurements between the 2 RUSI methods.

The improvement in measurement precision resulting from a single measure, or an average of the first 2 or 3 measures of both RUSI methods, was analyzed by comparing the corresponding SEMs. The reported reliability estimates in this paper were selected based on the balance between the improvement in precision and the additional measurement effort.

To compare the reliability of LM muscle measurements using RUSI in individuals with and without LBP, we conducted an exploratory analysis to investigate whether the reliability estimates of LM muscle measurements at various lumbar levels were affected by LBP status. Participants were classified into asymptomatic and symptomatic groups based on their reported LBP symptoms. Differences in the reliability estimates between the groups were reported when the 95% CIs of the ICC did not overlap.

The change in pain level across different conditions (during the CALT and between sessions) and between-session ODI scores were compared. If the change in pain level on the NPRS was greater than 2.4,[28] or if the change in ODI score was greater than 12%,[11] the data from the corresponding participant were excluded from the data analysis.

TABLE 1 DESCRIPTIVE STATISTICS OF PARTICIPANTS AT BASELINE*

	All Participants (n = 27)	Asymptomatic Participants (n = 13)	LBP Participants (n = 14)	Group Comparison
Age, y	29.6 ± 10.0	26.2 ± 5.5	32.7 ± 12.2	*t* = –1.8, *P* = .09
Sex (male), %	29.6	31.0	28.6	χ^2 = 0.02, *P* = .90
Body mass index, kg/m^2	22.7 ± 3.5	22.6 ± 3.1	22.8 ± 3.9	*t* = 0.1, *P* = .90
Modified ODI, %	5.6 ± 8.4	0.0 ± 0.0	10.7 ± 9.0	*t* = –4.0, *P*<.01
NPRS (0-10)	1.4 ± 1.8	0.1 ± 0.3	2.6 ± 1.9	*t* = –4.8, *P*<.01
Testing side (right), %	44.4	53.8	35.7	χ^2 = 0.9, *P* = .34
Prior history of LBP, n†	13	1	12	χ^2 = 5.8, *P*<.05

Abbreviations: LBP, low back pain; NPRS, numeric pain rating scale; ODI, Modified Oswestry Low Back Pain Disability Index.
**Values are mean ± SD unless otherwise indicated.*
†Prior history of LBP is defined as patients with a history of LBP who might or might not need to see physicians.

RESULTS

THIRTY-ONE PARTICIPANTS WERE RECRUITED in this study. Images from 4 participants were excluded. One participant's images were excluded because the investigator could not identify the LM boundaries. Images from 2 participants were excluded because the ultrasound transducer slid on their skin during muscle contraction. Additionally, 1 participant's results were excluded because no pillow was put underneath the participant's abdomen in the second visit.

The descriptive statistics of par-

TABLE 2 Changes in SEM Using the Average of 2 or 3 Measures of 2 Ultrasound Imaging Methods With Reference to a Single Measure

	Static RUSI (n = 27)			Video RUSI (n = 27)		
	Single Measure	2 Measures*†	3 Measures*‡	Single Measure	2 Measures*†	3 Measures*‡
Within-day intrarater SEM (day 1)						
L3-4 level						
Rest, mm	0.3	0.2 (21.1)	0.2 (42.1)	0.5	0.3 (29.4)	0.3 (41.2)
Contracted, mm	0.5	0.3 (32.2)	0.3 (42.2)	0.4	0.3 (17.1)	0.2 (43.9)
Percent change	3.5	2.4 (22.3)	2.1 (40.7)	3.1	2.1 (23.8)	1.8 (41.0)
L4-5 level						
Rest, mm	0.3	0.2 (25.0)	0.2 (41.7)	0.4	0.3 (44.2)	0.3 (44.2)
Contracted, mm	0.7	0.3 (35.4)	0.4 (41.7)	0.5	0.3 (37.3)	0.3 (41.2)
Percent change	3.2	1.4 (46.2)	1.9 (41.7)	2.4	1.7 (29.9)	1.4 (41.7)
Average decrease, %		30.4	41.7		30.3	42.2
Within-day intrarater SEM (day 2)						
L3-4 level						
Rest, mm	0.5	0.3 (34.4)	0.3 (43.8)	0.5	0.3 (45.1)	0.3 (43.1)
Contracted, mm	0.5	0.3 (40.0)	0.3 (41.8)	0.5	0.3 (44.9)	0.3 (44.9)
Percent change	3.6	2.3 (26.6)	2.2 (41.0)	3.0	2.2 (29.8)	1.8 (41.0)
L4-5 level						
Rest, mm	0.4	0.3 (36.4)	0.3 (43.2)	0.4	0.3 (30.6)	0.2 (42.9)
Contracted, mm	0.5	0.3 (33.3)	0.3 (40.5)	0.4	0.3 (40.0)	0.2 (37.1)
Percent change	2.8	1.9 (36.7)	1.7 (41.8)	2.3	1.9 (23.1)	1.3 (42.3)
Average decrease, %		34.6	42.0		35.6	41.9
Between-day intrarater SEM						
L3-4 level						
Rest, mm	1.1	1.0 (9.0)	1.0 (8.1)	0.7	0.6 (7.5)	0.6 (10.5)
Contracted, mm	1.0	0.8 (9.9)	0.7 (11.1)	0.6	0.6 (7.8)	0.6 (4.7)
Percent change	4.9	4.2 (14.1)	4.3 (13.9)	4.7	3.4 (27.2)	3.5 (24.2)
L4-5 level						
Rest, mm	1.1	1.1 (6.1)	1.0 (9.7)	0.8	0.8 (7.3)	0.7 (14.6)
Contracted, mm	1.0	0.8 (19.2)	0.8 (24.2)	0.6	0.6 (12.7)	0.5 (15.9)
Percent change	5.0	4.8 (5.6)	4.6 (9.5)	3.8	3.2 (14.8)	3.0 (21.6)
Average decrease, %		10.7	12.8		12.9	15.3

Abbreviations: SEM, standard error of measurement; static RUSI, rehabilitative ultrasound imaging static-image method; video RUSI, rehabilitative ultrasound imaging video-clip method.

*Values are mean (percent decrease from 1 measure), calculated as [1 – (SEM obtained from an average of multiple measures/SEM of a single measure] × 100%.

†Overall mean decrease in SEM resulting from using the average of 2 measures [(30.4 + 34.6 + 10.7 + 30.3 + 35.6 + 12.9)/6 = 25.8%].

‡Overall mean decrease in SEM resulting from using the average of 3 measures [(41.7 + 42.0 + 12.8 + 42.2 + 41.9 + 15.3)/6 = 32.6%]. Mean improvement in precision by using 3 measures rather than 2 measures, 6.8%.

ticipants at baseline are tabulated in **TABLE 1**. Though the asymptomatic and symptomatic groups did not significantly differ in age, gender, and body mass index, they were significantly different in ODI, NPRS, and prior history of LBP. No participant data were excluded due to a change in NPRS greater than 2.4 or an ODI score greater than 12% between visits or during the CALT.

Improvement of Reliability Estimates by Averaging Multiple Measures

The effect of multiple trials on measurement precision (reduction in SEMs) was first outlined to guide the choice of the reported reliability estimates in this paper. The reduction in SEMs as a result of averaging multiple trials using the static RUSI and video RUSI data is presented in **TABLE 2**. Overall, compared to the SEM using a single measurement across different testing conditions, SEMs decreased by an average of 25.8% when using a mean of 2 measures, and by 32.6% when using a mean of 3 measures (**TABLE 2**). Given the small improvement (6.8%) in precision by using an average of 3 measures compared to a 25.8% improvement by using 2 measures, the reliability results computed from the av-

TABLE 3 Within-Day Intrarater Reliability at Day 1*

	All Participants (n = 27)		Asymptomatic Participants (n = 13)		Participants With LBP (n = 14)	
	Mean ± SD	$ICC_{3,2}$ (95% CI)	Mean ± SD	$ICC_{3,2}$ (95% CI)	Mean ± SD	$ICC_{3,2}$ (95% CI)
L3-4 level						
Static RUSI						
Rest, mm	21.5 ± 3.8	0.99 (0.99, 1.00)	21.2 ± 4.7	0.99 (0.99, 1.00)	21.7 ± 3.1	0.99 (0.98, 1.00)
Contracted, mm	28.9 ± 4.1	0.99 (0.99, 1.00)	28.9 ± 5.6	0.98 (0.91, 0.99)	28.9 ± 2.3	0.98 (0.95, 0.99)
Percent change†	35.9 ± 13.3	0.97 (0.93, 0.99)	37.4 ± 10.0	0.95 (0.83, 0.98)	34.6 ± 15.9	0.98 (0.92, 0.99)
Video RUSI						
Rest, mm	21.6 ± 3.7	0.99 (0.98, 1.00)	21.2 ± 4.4	0.99 (0.96, 1.00)	22.0 ± 3.0	0.99 (0.96, 1.00)
Contracted, mm	29.1 ± 4.1	0.99 (0.99, 1.00)	29.1 ± 5.5	0.99 (0.99, 1.00)	29.1 ± 2.3	0.98 (0.94, 0.99)
Percent change†	35.6 ± 11.8	0.97 (0.93, 0.97)	37.7 ± 7.6	0.96 (0.88, 0.99)	33.7 ± 14.7	0.97 (0.91, 0.99)
L4-5 level						
Static RUSI						
Rest, mm	26.1 ± 3.9	0.99 (0.99, 1.00)	26.1 ± 4.7	0.99 (0.98, 1.00)	26.1 ± 3.2	0.99 (0.98, 1.00)
Contracted, mm	32.8 ± 4.4	0.99 (0.99, 1.00)	32.6 ± 6.1	0.99 (0.99, 1.00)	33.0 ± 2.1	0.98 (0.94, 0.99)
Percent change†	26.6 ± 12.0	0.99 (0.97, 0.99)	25.1 ± 7.2	0.96 (0.85, 0.99)	28.1 ± 15.3	0.99 (0.98, 1.00)
Video RUSI						
Rest, mm	26.5 ± 4.2	0.99 (0.99, 1.00)	26.2 ± 3.9	0.99 (0.99, 1.00)	26.7 ± 3.5	0.99 (0.98, 1.00)
Contracted, mm	32.9 ± 4.2	0.99 (0.99, 1.00)	32.6 ± 5.7	0.99 (0.99, 1.00)	33.2 ± 2.1	0.98 (0.94, 0.99)
Percent change†	25.2 ± 10.6	0.98 (0.94, 0.99)	24.6 ± 5.0	0.82 (0.43, 0.94)	25.7 ± 14.2	0.99 (0.98, 1.00)

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; LBP, low back pain; static RUSI, rehabilitative ultrasound imaging static-image method; video RUSI, rehabilitative ultrasound imaging video-clip method.
**Only day 1 within-day intrarater reliability was reported in this table because reliability estimates from both days were not statistically significantly different from one another.*
†Percent change in thickness.

erage of the first 2 measures are reported in this paper.

Reliability Estimates and Agreement of Muscle Measurements by Static RUSI and Video RUSI

The within-day intrarater reliability estimates ($ICC_{3,2}$) of all measurement parameters (resting and contracted muscle thickness and percent thickness change) using a given RUSI method on day 1 were not statistically different from the corresponding estimates on day 2. Specifically, the 95% CIs of estimates of either RUSI method on day 1 overlapped with the corresponding 95% CIs on day 2. Therefore, only the within-day-estimate data of day 1 are presented in **TABLE 3**.

Static RUSI The within-day reliability point estimates ($ICC_{3,2}$) of LM muscle thickness measurements (at rest and during contraction) using static RUSI were 0.99. The within-day static RUSI estimates of percent thickness change ranged from 0.97 to 0.99 (**TABLE 3**). Depending on the muscle condition (rest or contraction), the between-day reliability point estimates ($ICC_{3,2}$) of thickness measurements using static RUSI ranged from 0.93 to 0.97, and those of percent thickness change ranged from 0.80 to 0.87 (**TABLE 4**). With reference to the relative MDC_{95} values, the within-day relative MDC_{95} values of LM thickness measurements and percent thickness change using static RUSI ranged from 2.7% to 3.3% and from 14.3% to 18.7%, respectively (**TABLE 5**). The between-day relative MDC_{95} values of static RUSI for LM thickness measurements ranged from 6.7% to 13.1%, whereas those for percent thickness change were between 31.1% and 46.6% (**TABLE 5**).

Video RUSI All the within-day reliability point estimates of thickness measurements using video RUSI were 0.99. The within-day estimates of percent thickness change using video RUSI ranged from 0.97 to 0.98 (**TABLE 3**). The between-day reliability estimates of muscle thickness measurements using video RUSI ranged from 0.97 to 0.98. The between-day estimates of percent thickness change using video RUSI ranged from 0.89 to 0.90 (**TABLE 4**). Regarding the relative MDC_{95} values using video RUSI, the within-day relative MDC_{95} values of LM thickness measurements ranged from 2.6% to 4.2%, and those of percent thickness change ranged between 16.0% and 18.3% (**TABLE 5**). Likewise, the between-day relative MDC_{95} values of video RUSI for LM thickness measurements and percent thickness change ranged from 4.5% to 8.0% and from 25.1% to 32.7%, respectively (**TABLE 5**).

TABLE 4 Between-Day Intrarater Reliability

	All Participants (n = 27)		Asymptomatic Participants (n = 13)		Participants With LBP (n = 14)	
	Mean ± SD	$ICC_{3,2}$ (95% CI)	Mean ± SD	$ICC_{3,2}$ (95% CI)	Mean ± SD	$ICC_{3,2}$ (95% CI)
L3-4 level						
Static RUSI						
Rest, mm	21.4 ± 3.8	0.93 (0.85, 0.97)	21.1 ± 4.5	0.98 (0.93, 0.99)	21.8 ± 3.0	0.81 (0.47, 0.94)
Contracted, mm	29.2 ± 3.9	0.97 (0.92, 0.99)	29.1 ± 5.3	0.98 (0.96, 1.00)	29.5 ± 1.8	0.88 (0.63, 0.96)
Percent change*	37.9 ± 11.9	0.87 (0.74, 0.94)	38.7 ± 10.7	0.86 (0.59, 0.95)	37.1 ± 13.6	0.90 (0.68, 0.97)
Video RUSI						
Rest, mm	21.5 ± 3.8	0.97 (0.94, 0.99)	21.2 ± 4.5	0.99 (0.97, 1.00)	22.0 ± 3.0	0.93 (0.79, 0.98)
Contracted, mm	29.4 ± 4.0	0.98 (0.95, 0.99)	29.2 ± 5.4	0.99 (0.96, 0.99)	29.7 ± 1.6	0.89 (0.66, 0.97)
Percent change*	37.5 ± 10.8	0.90 (0.79, 0.96)	38.8 ± 7.8	0.79 (0.47, 0.93)	36.2 ± 13.5	0.94 (0.82, 0.98)
L4-5 level						
Static RUSI						
Rest, mm	25.9 ± 4.0	0.93 (0.85, 0.97)	26.2 ± 4.7	0.97 (0.91, 0.99)	25.6 ± 3.3	0.84 (0.55, 0.95)
Contracted, mm	33.0 ± 4.4	0.97 (0.92, 0.99)	32.5 ± 5.8	0.98 (0.92, 0.99)	33.5 ± 2.0	0.89 (0.67, 0.97)
Percent change*	28.3 ± 10.8	0.80 (0.61, 0.91)	24.4 ± 5.5	0.55 (0.05, 0.84)	32.5 ± 13.5	0.83 (0.52, 0.95)
Video RUSI						
Rest, mm	26.2 ± 4.1	0.97 (0.92, 0.98)	26.3 ± 4.9	0.98 (0.94, 0.99)	26.1 ± 3.1	0.93 (0.76, 0.98)
Contracted, mm	33.1 ± 4.2	0.98 (0.96, 0.99)	33.1 ± 5.7	0.99 (0.96, 1.00)	33.6 ± 1.8	0.94 (0.82, 0.98)
Percent change*	27.2 ± 9.9	0.89 (0.77, 0.95)	24.7 ± 5.1	0.60 (0.08, 0.92)	29.6 ± 13.0	0.94 (0.77, 0.98)

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; LBP, low back pain; static RUSI, rehabilitative ultrasound imaging static-image method; video RUSI, rehabilitative ultrasound imaging video-clip method.
**Percent change in thickness.*

The 95% CIs of reliability estimates obtained by static RUSI and video RUSI overlapped with one another across all testing conditions (**TABLES 3** and **4**). Further, bias estimates between the 2 RUSI methods were small and not significantly different from zero for all thickness comparisons except for resting thickness at the L4-5 level on day 1, in which video RUSI measurements were 0.4 mm larger on average than static RUSI measurements (**TABLE 6**). The limits of agreement, biases, and the corresponding 95% CIs are presented in **TABLE 6**.

The Reliability Estimates of Different Participants

Asymptomatic Group In this group, the within-day reliability estimates ($ICC_{3,2}$) of muscle thickness measurements as obtained by the 2 RUSI methods ranged from 0.98 to 0.99. Regarding the reliability estimates of percent thickness change, all $ICC_{3,2}$ values ranged between 0.95 and 0.99, except for the within-day point estimate at the L4-5 level, which was 0.82 (**TABLE 3**). Similarly, the between-day reliability estimates of muscle thickness measurements ranged from 0.97 to 0.99, whereas such estimates for percent thickness change ranged from 0.55 to 0.86 (**TABLE 4**).

Symptomatic Group The within-day reliability estimates ($ICC_{3,2}$) of thickness measurements and percent thickness change measured by the 2 RUSI methods in this group ranged from 0.98 to 0.99 and from 0.97 to 0.99, respectively (**TABLE 3**). The between-day estimates for LM thickness measurements ranged from 0.81 to 0.94, and those for percent thickness change ranged from 0.83 to 0.94 (**TABLE 4**).

DISCUSSION

Our study compared the within- and between-day intrarater reliability of LM muscle thickness and percent thickness change obtained (1) by static RUSI and video RUSI, (2) by a single measure or an average of 2 or 3 measures of static RUSI and video RUSI, and (3) between individuals with and without LBP. Our results showed no significant difference between the 2 RUSI methods for corresponding reliability estimates. Given that an ICC value above 0.7 is considered to be the minimal acceptable reliability for research[26] and 0.90 is the acceptable reliability for individual comparisons or clinical use,[33] both RUSI methods demonstrated excellent intrarater reliability of muscle thickness measurements for within- and between-day comparisons. Our findings were consistent with those of other studies.[7,21,23,25,35,37] The present RUSI methods, however, yielded higher within- and between-day reliability estimates of percent thickness change than those of previous studies.[22,23] As in previous research,[23] using an average of 3 measures yielded

TABLE 5 RELATIVE MDC_{95} WITH REFERENCE TO THE CORRESPONDING MEAN RESTING OR CONTRACTED MULTIFIDUS MUSCLE THICKNESS OR MEAN PERCENT THICKNESS CHANGE VALUES

	All Participants (n = 27)			Asymptomatic Participants (n = 13)			Participants With LBP (n = 14)		
	Mean	MDC_{95}	Relative MDC_{95}	Mean	MDC_{95}	Relative MDC_{95}	Mean	MDC_{95}	Relative MDC_{95}
Within-day, L3-4 level, day 1									
Static image									
Rest, mm	21.5	0.7	3.3%	21.2	0.7	3.3%	21.7	0.7	3.2%
Contracted, mm	28.9	0.9	3.1%	28.9	0.9	3.1%	28.9	0.8	2.8%
Percent change*	35.9	6.7	18.7%	37.4	6.5	17.4%	34.6	7.0	20.2%
Video clip									
Rest, mm	21.6	0.9	4.2%	21.2	0.9	4.2%	22.0	0.4	1.8%
Contracted, mm	29.1	0.9	3.1%	29.1	0.8	2.7%	29.1	0.9	3.1%
Percent change*	35.6	5.7	16.0%	37.7	4.1	10.9%	33.7	6.9	20.5%
Within-day, L4-5 level, day 1									
Static image									
Rest, mm	26.1	0.7	2.7%	26.1	0.9	3.4%	26.1	0.3	1.1%
Contracted, mm	32.8	0.9	2.7%	32.6	1.1	3.4%	33.0	0.8	2.4%
Percent change*	26.6	3.8	14.3%	25.1	3.9	15.5%	28.1	3.8	13.5%
Video clip									
Rest, mm	26.5	0.7	2.6%	26.2	0.7	2.7%	26.7	1.0	3.7%
Contracted, mm	32.9	0.9	2.7%	32.6	0.9	2.8%	33.2	0.8	2.4%
Percent change*	25.2	4.6	18.3%	24.6	5.9	24.0%	25.7	3.5	13.6%
Between-day, L3-4 level									
Static image									
Rest, mm	21.4	2.8	13.1%	21.1	1.9	9.0%	21.8	3.6	16.5%
Contracted, mm	29.2	2.0	6.8%	29.1	2.3	7.9%	29.5	1.7	5.7%
Percent change*	37.9	11.8	31.1%	38.7	11.3	29.2%	37.1	12.9	34.8%
Video clip									
Rest, mm	21.5	1.7	7.9%	21.2	1.3	6.1%	22.0	2.1	9.5%
Contracted, mm	29.4	1.6	5.4%	29.2	1.8	6.2%	29.7	1.5	5.1%
Percent change*	37.5	9.4	25.1%	38.8	9.8	25.3%	36.2	8.9	24.6%
Between-day, L4-5 level									
Static image									
Rest, mm	25.9	3.0	11.6%	26.2	2.1	8.0%	25.6	3.6	14.1%
Contracted, mm	33.0	2.2	6.7%	32.5	2.4	7.4%	33.5	1.8	5.4%
Percent change*	28.3	13.2	46.6%	24.4	10.2	41.8%	32.5	15.7	48.3%
Video clip									
Rest, mm	26.2	2.1	8.0%	26.3	1.2	4.6%	26.1	2.4	9.2%
Contracted, mm	33.1	1.5	4.5%	33.1	1.8	5.4%	33.6	1.2	3.6%
Percent change*	27.2	8.9	32.7%	24.7	9.0	36.4%	29.6	9.1	30.7%

Abbreviations: LBP, low back pain; MDC_{95}, minimal detectable change at the 95% confidence interval.
**Percent change in thickness.*

little improvement in measurement precision compared to 2 measures. There was no significant difference in the reliability estimates between asymptomatic and LBP participants across most testing conditions.

Although prior studies have investigated the within- and between-day reliability estimates of static RUSI in measuring LM thickness, most of them were limited by (1) a small sample size,[37] (2) measuring muscle thickness at rest and not while contracted,[37] (3) failure to report the repeatability of percent thickness change over time,[6,12,23,24] or (4) not comparing individuals with and without LBP. Our study aimed to address these limitations and to quantify the intrarater reliability of static RUSI and video RUSI by recruiting individuals with and without LBP. To minimize error due to

TABLE 6 Intermethods 95% LOA, Bias, and 95% CIs of Bias for Comparing Static RUSI to Video RUSI Measurements

	Mean ± SD	Bias*	95% CI of Bias	95% LOA
Day 1, L3-4 level				
Rest, mm	21.5 ± 3.7	−0.2	−0.4, 0.0	−1.6, 1.2
Contracted, mm	29.0 ± 4.1	−0.2	−0.4, 0.0	−1.4, 0.1
Percent change†	35.8 ± 12.2	0.3	−1.2, 1.8	−7.4, 7.9
Day 2, L3-4 level				
Rest, mm	21.4 ± 3.6	−0.2	−0.6, 0.2	−2.4, 2.0
Contracted, mm	29.4 ± 3.8	−0.2	−0.6, 0.2	−2.0, 1.6
Percent change†	38.4 ± 10.9	0.4	−2.0, 2.7	−10.9, 11.7
Day 1, L4-5 level				
Rest, mm	26.3 ± 3.8	−0.4‡	−0.6, −0.2	−1.8, 1.0
Contracted, mm	32.9 ± 4.2	−0.1	−0.5, 0.3	−2.1, 1.9
Percent change†	26.0 ± 10.8	1.6	−0.4, 3.5	−8.3, 11.4
Day 2, L4-5 level				
Rest, mm	26.0 ± 4.1	−0.3	−0.7, 0.1	−2.1, 1.5
Contracted, mm	33.1 ± 4.2	−0.2	−0.4, 0.0	−1.4, 1.0
Percent change†	28.1 ± 10.2	0.6	−1.5, 2.6	−9.4, 10.5

Abbreviations: CI, confidence interval; LOA, limits of agreement; static RUSI, rehabilitative ultrasound imaging static-image method; video RUSI, rehabilitative ultrasound imaging video-clip method.
**Static image minus video.*
†Percent change in thickness.
‡Statistically significant bias.

manual probe handling, we employed a transducer support system to hold the ultrasound transducer instead of involving 2 examiners working together to collect data, as in other studies.[23,34]

Static RUSI Versus Video RUSI

No statistically significant difference in reliability estimates of LM muscle thickness or percent thickness change was noted between static RUSI and video RUSI, as their corresponding 95% CIs overlapped in all comparisons. The close association between these methods was further supported by the absence of significant bias in the measured LM muscle thickness in all conditions except for the day 1 resting LM muscle thickness at the L4-5 level. However, the 95% CI for the bias (−0.6, −0.2 mm) of day 1 resting muscle thickness at the L4-5 level was smaller than the corresponding within-day intrarater MDC_{95} values (0.7-0.9 mm) of the 2 methods at day 1. This implies that this statistically significant bias was smaller than the measurement errors. In short, our results suggest that static RUSI and video RUSI are interchangeable for LM muscle thickness measurement.

In addition, our static RUSI and video RUSI reliability point estimates were comparable to the corresponding estimates reported in previous research.[7,21,25,37] Regarding the reliability estimates of LM thickness, our reliability estimates using video RUSI agreed with the Dickx et al[7] study, whereas our within- and between-day point estimates using static RUSI coincided with those reported in studies involving symptomatic[23,25] and asymptomatic participants.[21,37] Although our reliability estimates of percent thickness change are higher than those reported in previous studies,[22,23] the reliability estimates of percent thickness change remained smaller than those of LM thickness measurements in the current study. This finding could be ascribed to the inherited measurement error from 2 imperfect measurements at rest and while contracted.[23]

The good agreement[3] of video RUSI and static RUSI in LM muscle measurements enables examiners to choose the appropriate RUSI method according to the demands of the situation. Video RUSI is suitable for assessing the LM thickness of individuals who cannot endure prolonged prone position. Examiners can adjust the preset recording duration to optimize the acquisition of RUSI video clips.

Improvement of Measurement Precision by Multiple Measures

Our results corroborated the use of an average of 2 LM muscle measures to optimize measurement precision based on the SEM reduction of muscle thickness measurements and percent thickness change (**TABLE 2**). This finding agrees with a study that calculated measurement precision improvement based on the SEMs of muscle thickness.[23] However, our results differ from those of another study that recommended using an average of 3 measures to optimize measurement precision of percent LM thickness change.[25] Because the discrepancy might be ascribed to the greater measurement error of percent thickness change,[25] we recalculated the improvements in measurement precision of averaging multiple measures using only the SEMs of percent thickness change. Overall, compared to the SEM using a single measure across different testing conditions, SEMs decreased by an average of 25.1% when using a mean of 2 measures, and by a further 8.3% (33.4%) when using a mean of 3 measures. Given the relatively small increase in measurement precision by averaging in a third measure, 2 measures appeared to be adequate for this specific methodology. However, future research should determine the optimal number of multiple measurements according to the precision requirement of individual studies.

The higher measurement precision of this study compared to previous research is reflected by the smaller relative MDC_{95} values. Given the differences in participants' characteristics (body mass index, age, sex), the absolute MDC_{95} values in the current study were not directly compared with those in previous research. However, the measurement precision between studies can be compared by relative MDC_{95} values (**TABLE 5**). A smaller relative MDC_{95} value represents higher measurement precision.[34] Our within-day relative MDC_{95} values represented 2.6% to 2.7% of the corresponding L4-5 LM thickness, whereas those reported in other studies[23,34] represented 4.2% to 16% of the L4-5 LM muscle thickness. Likewise, our relative MDC_{95} values of percent thickness change at the L4-5 level represented 14.3% to 46.6% of the corresponding percent LM muscle thickness change in within- and between-day comparisons, whereas the corresponding relative MDC_{95} values represented 98% to 112% of the mean percent thickness change in another study.[23] Our improved measurement precision may support the use of RUSI measurements for monitoring subtle change in LM muscle activity over time, which was deemed to be infeasible in the past.[23]

The improvement in measurement precision in the current study may be due to our experimental setup. Because there is no difference in resolution between the SonoSite M-Turbo used in this study and the SonoSite TITAN machine used in previous studies,[23,34] our improved measurement precision may be attributed to various procedural improvements (transparency sheets, horizontal feedback bar, and transducer support system).

Asymptomatic and Symptomatic Participants

Because RUSI is commonly used for LM measurements in individuals with and without LBP in clinics, an exploratory analysis was performed to investigate whether the reliability estimates of LM measurements using RUSI were different between the 2 participant groups. The within- and between-day reliability point estimates of both groups were high. Except for 1 measurement of percent thickness change, there was no significant difference in the reliability estimates between the groups, as their corresponding 95% CIs overlapped. The significantly lower within-day reliability estimate of day 1 percent thickness change at the L4-5 level using video RUSI ($ICC_{3,2}$ = 0.82) may be attributed to the effect of low variance (SD, 5.1%) and small range of thickness change (14.6%) in the asymptomatic group compared to the symptomatic group (SD, 14.2%; range of thickness change, 42.8%)[29] (**TABLE 3**). Similarly, the small asymptomatic sample variance (SD, 5.1%-5.5%) resulted in moderate between-day reliability estimates of percent thickness change at the L4-5 level ($ICC_{3,2}$ = 0.55-0.60), which were much higher prior to the group classification ($ICC_{3,2}$ = 0.80-0.89) (**TABLE 4**). The results of this exploratory analysis showed that static RUSI and video RUSI are reliable LM muscle measurement tools for the clinical population.

To investigate the localized effect of pain on the reliability of LM measurements, LM measurements at the L3-4 and L4-5 levels were chosen because they represent common asymptomatic and symptomatic levels, respectively. If LBP has a localized effect on the reliability of LM measurements, the reliability of LM measurements at the L3-4 level should be significantly different from that at the L4-5 level in symptomatic participants. Additionally, there should be no statistically significant difference in the reliability estimates of L3-4 LM measurements between the 2 participant groups. Our results showed that the reliability of LM measurements was not affected by localized pain, because the 95% CIs of reliability estimates at all lumbar levels overlapped one another in all conditions (**TABLES 3** and **4**).

Limitations

This study is limited by the inclusion of single-examiner data, which affects its generalizability to other examiners. However, our experimental setup (repositioning the ultrasound transducer between consecutive measurements) might have minimized many possible errors,[6] and it is very likely that trained examiners with comparable experience would produce similar results.

This study was planned to recruit representative samples of asymptomatic participants and patients with LBP. However, the enrolled symptomatic participants had low ODI scores. Only 5 of 14 LBP participants met the classification criteria of patients with LBP (NPRS of 2 or greater and ODI greater than or equal to 12%).[5,9] Although our participants might not represent typical samples of patients with acute or chronic LBP in clinics, they represented patients with back discomfort or recovering from LBP in clinical situations. Future research should investigate whether the reliability estimates may change in participants with higher intensity of LBP.

Finally, the transducer support system may not be suitable for participants with hypertrophied back muscles. Although this system held the ultrasound transducer in place during data collection in most of the participants, 2 participants with hypertrophied erector spinae caused the transducer to slide during CALTs. To minimize the risk of data loss, it is suggested that examiners use manual RUSI measurement in such cases.

Future Studies and Development

Given the high reliability of video RUSI in LM measurements and its ability to evaluate LM thickness changes over time, video RUSI may broaden the applications of B-mode ultrasound for trunk muscle assessments. Traditionally, B-mode static RUSI is solely used to evaluate muscle morphology, whereas M-mode ultrasound imaging is adopted to assess onset of muscle activity or relative timing of muscle thickness changes over time.[40] Video RUSI combines the advantages of B- and M-mode ultrasound imaging, allowing both the visualization of muscle morphology and the sequential measurement of muscle thickness over time. With these features, new software programs can be developed to determine a true maximum muscle contraction over the span of a RUSI video, or to calculate the average LM muscle thickness at rest and during contraction. Video RUSI may also be used to monitor muscle fatigue by evaluating the rate of muscle thickness change. Future research may incorporate video RUSI to investigate the onset of contraction within a muscle, or the relative timing and movement patterns of various target structures over time. These findings may deepen our understanding of back muscle activity in individuals with and without LBP.

CONCLUSION

There was no significant difference in the measured LM thickness and the intrarater reliability estimates of LM measurements between static RUSI and video RUSI. Both RUSI methods demonstrated high reliability of LM thickness measurements and adequate reliability of percent thickness change measurements in within- and between-day comparisons. While using an average of 2 measures optimizes measurement precision, the reliability estimates of all but 1 LM muscle measurement were not statistically different between individuals with and without LBP. Overall, video RUSI is a reliable surrogate for static RUSI for LM muscle measurements and offers the additional advantage of shorter data collection time.

KEY POINTS

FINDINGS: Both static RUSI and video RUSI methods showed moderate to high intrarater reliability estimates of LM measurements in individuals with and without LBP. Using the average of 2 measures optimizes intrarater measurement precision of LM muscle measurements.

IMPLICATIONS: The video RUSI method is a potential surrogate for the static RUSI method in measuring LM thickness or percent thickness change, given its measurement agreement with the static RUSI method and shorter duration of data collection.

CAUTION: The reliability estimates in this study were obtained based on a single rater's data; the results should be generalized to other raters with caution and should not be generalized to interrater comparisons.

ACKNOWLEDGEMENTS: *The authors thank Ms Martha Funabashi for assisting with the data collection.*

REFERENCES

1. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep*. 1966;19:3-11. http://dx.doi.org/10.2466/pr0.1966.19.1.3
2. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol*. 2003;22:85-93. http://dx.doi.org/10.1002/uog.122
3. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135-160.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Int J Nurs Stud*. 2010;47:931-936. http://dx.doi.org/10.1016/j.ijnurstu.2009.10.001
5. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine (Phila Pa 1976)*. 2005;30:1331-1334.
6. Costa LO, Maher CG, Latimer J, Smeets RJ. Reproducibility of rehabilitative ultrasound imaging for the measurement of abdominal muscle activity: a systematic review. *Phys Ther*. 2009;89:756-769. http://dx.doi.org/10.2522/ptj.20080331
7. Dickx N, Cagnie B, Parlevliet T, Lavens A, Danneels L. The effect of unilateral muscle pain on recruitment of the lumbar multifidus during automatic contraction. An experimental pain study. *Man Ther*. 2010;15:364-369. http://dx.doi.org/10.1016/j.math.2010.02.002
8. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Phys Ther*. 1994;74:777-788.
9. Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine (Phila Pa 1976)*. 2000;25:2940-2953. http://dx.doi.org/10.1097/00007632-200011150-00017
10. Farrar JT, Berlin JA, Strom BL. Clinically important changes in acute pain outcome measures: a validation study. *J Pain Symptom Manage*. 2003;25:406-411.
11. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. *Phys Ther*. 2001;81:776-788.
12. Hebert JJ, Koppenhaver SL, Parent EC, Fritz JM. A systematic review of the reliability of rehabilitative ultrasound imaging for the quantitative assessment of the abdominal and lumbar trunk muscles. *Spine (Phila Pa 1976)*. 2009;34:E848-E856. http://dx.doi.org/10.1097/BRS.0b013e3181ae625c
13. Herbert WJ, Heiss DG, Basso DM. Influence of feedback schedule in motor performance and learning of a lumbar multifidus muscle task using rehabilitative ultrasound imaging: a randomized clinical trial. *Phys Ther*. 2008;88:261-269. http://dx.doi.org/10.2522/ptj.20060308
14. Hides J, Gilmore C, Stanton W, Bohlscheid E. Multifidus size and symmetry among chronic LBP and healthy asymptomatic subjects. *Man Ther*. 2008;13:43-49. http://dx.doi.org/10.1016/j.math.2006.07.017
15. Hides JA, Richardson CA, Jull GA. Magnetic resonance imaging and ultrasonography of the lumbar multifidus muscle. Comparison of two different modalities. *Spine (Phila Pa 1976)*. 1995;20:54-58.
16. Hides JA, Richardson CA, Jull GA. Multifidus muscle recovery is not automatic after resolution of acute, first-episode low back pain. *Spine (Phila Pa 1976)*. 1996;21:2763-2769. http://dx.doi.org/10.1097/00007632-199612010-00011
17. Hides JA, Stokes MJ, Saide M, Jull GA, Cooper DH. Evidence of lumbar multifidus muscle wasting ipsilateral to symptoms in patients with acute/subacute low back pain. *Spine (Phila Pa 1976)*. 1994;19:165-172. http://dx.doi.org/10.1097/00007632-199401001-00009
18. Hu Y, Wong YL, Lu WW, Kawchuk GN. Creation of an asymmetrical gradient of back muscle activity and spinal stiffness during asymmetrical hip extension. *Clin Biomech (Bristol, Avon)*. 2009;24:799-806. http://dx.doi.org/10.1016/j.clinbiomech.2009.07.013
19. Jensen MP, Turner JA, Romano JM, Fisher LD. Comparative reliability and validity of chronic pain intensity measures. *Pain*. 1999;83:157-162. http://dx.doi.org/10.1016/S0304-3959(99)00101-3
20. Kiesel KB, Uhl T, Underwood FB, Nitz AJ. Rehabilitative ultrasound measurement of select trunk muscle activation during induced pain. *Man Ther*. 2008;13:132-138. http://dx.doi.org/10.1016/j.math.2006.10.003
21. Kiesel KB, Uhl TL, Underwood FB, Rodd DW, Nitz AJ. Measurement of lumbar multifidus muscle contraction with rehabilitative ultrasound imaging. *Man Ther*. 2007;12:161-166. http://dx.doi.org/10.1016/j.math.2006.06.011
22. Kiesel KB, Underwood FB, Mattacola CG, Nitz AJ, Malone TR. A comparison of select trunk muscle thickness change between subjects with

low back pain classified in the treatment-based classification system and asymptomatic controls. *J Orthop Sports Phys Ther*. 2007;37:596-607. http://dx.doi.org/10.2519/jospt.2007.2574

23. Koppenhaver SL, Hebert JJ, Fritz JM, Parent EC, Teyhen DS, Magel JS. Reliability of rehabilitative ultrasound imaging of the transversus abdominis and lumbar multifidus muscles. *Arch Phys Med Rehabil*. 2009;90:87-94. http://dx.doi.org/10.1016/j.apmr.2008.06.022
24. Koppenhaver SL, Hebert JJ, Parent EC, Fritz JM. Rehabilitative ultrasound imaging is a valid measure of trunk muscle size and activation during most isometric sub-maximal contractions: a systematic review. *Aust J Physiother*. 2009;55:153-169. http://dx.doi.org/10.1016/S0004-9514(09)70076-5
25. Koppenhaver SL, Parent EC, Teyhen DS, Hebert JJ, Fritz JM. The effect of averaging multiple trials on measurement error during ultrasound imaging of transversus abdominis and lumbar multifidus muscles in individuals with low back pain. *J Orthop Sports Phys Ther*. 2009;39:604-611. http://dx.doi.org/10.2519/jospt.2009.3088
26. Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther*. 1996;18:979-992. http://dx.doi.org/10.1016/S0149-2918(96)80054-3
27. MacDonald D, Moseley GL, Hodges PW. Why do some patients keep hurting their back? Evidence of ongoing back muscle dysfunction during remission from recurrent back pain. *Pain*. 2009;142:183-188. http://dx.doi.org/10.1016/j.pain.2008.12.002
28. Maughan EF, Lewis JS. Outcome measures in chronic low back pain. *Eur Spine J*. 2010;19:1484-1494. http://dx.doi.org/10.1007/s00586-010-1353-6
29. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, NJ: Pearson/Prentice Hall; 2008.
30. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil*. 1998;12:187-199. http://dx.doi.org/10.1191/026921598672178340
31. Stanton T, Kawchuk G. The effect of abdominal stabilization contractions on posteroanterior spinal stiffness. *Spine (Phila Pa 1976)*. 2008;33:694-701. http://dx.doi.org/10.1097/BRS.0b013e318166e034
32. Stokes M, Rankin G, Newham DJ. Ultrasound imaging of lumbar multifidus muscle: normal reference ranges for measurements and practical guidance on the technique. *Man Ther*. 2005;10:116-126. http://dx.doi.org/10.1016/j.math.2004.08.013
33. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th ed. New York, NY: Oxford University Press; 2008.
34. Teyhen DS, George SZ, Dugan JL, Williamson J, Neilson BD, Childs JD. Inter-rater reliability of ultrasound imaging of the trunk musculature among novice raters. *J Ultrasound Med*. 2011;30:347-356.
35. Van K, Hides JA, Richardson CA. The use of real-time ultrasound imaging for biofeedback of lumbar multifidus muscle contraction in healthy subjects. *J Orthop Sports Phys Ther*. 2006;36:920-925. http://dx.doi.org/10.2519/jospt.2006.2304
36. Vasseljen O, Dahl HH, Mork PJ, Torp HG. Muscle activity onset in the lumbar multifidus muscle recorded simultaneously by ultrasound imaging and intramuscular electromyography. *Clin Biomech (Bristol, Avon)*. 2006;21:905-913. http://dx.doi.org/10.1016/j.clinbiomech.2006.05.003
37. Wallwork TL, Hides JA, Stanton WR. Intrarater and interrater reliability of assessment of lumbar multifidus muscle thickness using rehabilitative ultrasound imaging. *J Orthop Sports Phys Ther*. 2007;37:608-612. http://dx.doi.org/10.2519/jospt.2007.2418
38. Wallwork TL, Stanton WR, Freke M, Hides JA. The effect of chronic low back pain on size and contraction of the lumbar multifidus muscle. *Man Ther*. 2009;14:496-500. http://dx.doi.org/10.1016/j.math.2008.09.006
39. Werneke M, Hart DL, Cook D. A descriptive study of the centralization phenomenon. A prospective analysis. *Spine (Phila Pa 1976)*. 1999;24:676-683.
40. Whittaker JL, Teyhen DS, Elliott JM, et al. Rehabilitative ultrasound imaging: understanding the technology and its applications. *J Orthop Sports Phys Ther*. 2007;37:434-449. http://dx.doi.org/10.2519/jospt.2007.2530
41. Whittaker JL, Warner MB, Stokes MJ. Induced transducer orientation during ultrasound imaging: effects on abdominal muscle thickness and bladder position. *Ultrasound Med Biol*. 2009;35:1803-1811. http://dx.doi.org/10.1016/j.ultrasmedbio.2009.05.018
42. Zhao WP, Kawaguchi Y, Matsui H, Kanamori M, Kimura T. Histochemistry and morphology of the multifidus muscle in lumbar disc herniation: comparative study between diseased and normal sides. *Spine (Phila Pa 1976)*. 2000;25:2191-2199.