Research Report

Reproducibility of Rehabilitative Ultrasound Imaging for the Measurement of Abdominal Muscle Activity: A Systematic Review

Leonardo Oliveira Pena Costa, Chris G. Maher, Jane Latimer, Rob J.E.M. Smeets

L.O.P. Costa, PT, is a PhD candidate, The George Institute for International Health, The University of Sydney, PO Box M201, Missenden Road, Sydney, New South Wales, 2050 Australia. Address all correspondence to Mr Costa at: lcos3060@gmail.com.

C.G. Maher, PT, PhD, is Director, Musculoskeletal Division, The George Institute for International Health, The University of Sydney.

J. Latimer, PT, PhD, is Associate Professor, The George Institute for International Health, The University of Sydney.

R.J.E.M. Smeets, MD, is Rehabilitation Physician, The George Institute for International Health, The University of Sydney; Rehabilitation Foundation Limburg, Hoensbroek, the Netherlands; and Department of Rehabilitation Medicine, Care and Public Health Research Institute, Maastricht University, Maastricht, the Netherlands.

[Costa LOP, Maher CG, Latimer J, Smeets RJEM. Reproducibility of rehabilitative ultrasound imaging for the measurement of abdominal muscle activity: a systematic review. *Phys Ther.* 2009;89:756–769.]

Background. Rehabilitative ultrasound imaging (RUSI) measures of abdominal wall muscles are used to indirectly measure muscle activity. These measures are used to identify suitable patients and to monitor progress of motor control exercise treatment of people with low back pain.

Purpose. The purpose of this study was to systematically review reproducibility studies of RUSI for measuring thickness of abdominal wall muscles.

Data Sources. Eligible studies were identified via searches of MEDLINE, EMBASE, and CINAHL. The authors also searched personal files and tracked references of the retrieved studies via the Web of Science Index.

Study Selection. Studies involving any type of reliability and or agreement of any type of ultrasound measurements (B or M mode) for any of the abdominal wall muscles were selected.

Data Extraction. Two independent reviewers extracted data and assessed methodological quality.

Data Synthesis. Due to heterogeneity of the studies' designs, pooling the data for a meta-analysis was not possible. Twenty-one studies were included, and these studies were typically of low quality and studied subjects who were healthy rather than people seeking care for low back pain. The studies reported good to excellent reliability for single measures of thickness and poor to good reliability for measures of thickness change (reflecting the muscle activity). Interestingly, no studies checked reliability of measures of the difference in thickness changes over time (representing improvement or deterioration in muscle activity).

Conclusions. The current evidence of the reproducibility of RUSI for measuring abdominal muscle activity is based mainly on studies with suboptimal designs and the study of people who were healthy. The critical question of whether RUSI provides reliable measures of improvement in abdominal muscle activity remains to be evaluated.

The use of motor control exercise in the management of nonspecific low back pain (LBP) has become increasingly popular in clinical practice. The rationale for the use of motor control exercises arises from the view that the activity of the deep abdominal muscles is critical for the dynamic control of the lumbar spine, with poor control resulting in lumbar spine symptoms.[1,2] Most of the studies[3,4] to date that have measured the activity of the deep spinal muscles were based on fine-wire electromyographic (EMG) examination, which is costly and uncomfortable and has associated risks such as infection, making its use in clinical practice difficult. An alternate approach is to indirectly measure the recruitment of the abdominal muscles by assessing morphologic changes (thickness changes) using rehabilitative ultrasound imaging (RUSI)[5,6] There is some evidence that RUSI measurements of thickness change are correlated to EMG measurements of muscle activity at low levels of contraction force (up to approximately 30% of maximal force).[7–9] This has resulted in the increasing use of ultrasound machines by health care professionals to assess motor control deficits and to provide feedback for patients receiving treatment for LBP.

It is widely accepted that clinical measurements need to be reproducible. Without an acceptable level of reproducibility, the clinical utility of assessment tools becomes substantially compromised.[10] It is important to state that *reproducibility* (ie, degree to which repeated measurements in people in stable health provide similar answers[11]) should be understood as an umbrella term for reliability and agreement,[12] where *reliability* could be defined as "the extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error)"[11(p36)] and *agreement* could be defined as "the extent to which the scores on repeated measures are close to each other (absolute measurement error)."[11(p36)] In an ideal scenario, a reproducibility study should be designed with particular attention to 5 points.[11] First, a reproducibility study should be performed in patients with the condition for which the test will be used (eg, ultrasound tests for the abdominal wall in patients with LBP). Second, the evaluation of reproducibility should be performed in a manner as similar as possible to the conditions used in clinical practice. Third, the study must be controlled for the order of the tests and for memory bias (which can easily be performed with blinding, with an appropriate time interval, and with randomization or counterbalancing procedures for ordering of the tests). Fourth, the study must be sufficiently statistically powered. Finally, the study must be analyzed in a way that the results can be reasonably generalized to a certain population of clinicians (ie, appropriate description of the tester and appropriate statistical analysis).[13]

In the last decade, a large number of studies evaluating the reproducibility of RUSI measures of abdominal muscle activity have been published,[14–17] and some of these studies have been reviewed in a nonsystematic design.[18] To date, there is no comprehensive systematic review that has attempted to investigate the reproducibility of RUSI measurements of the activity of the abdominal wall muscles. The objective of this study was to systematically review all reproducibility studies of RUSI for abdominal wall muscles.

Method

Data Sources and Searches

Studies were identified through searches of MEDLINE (1950 to 2008), EMBASE (1974 to 2008), and CINAHL (1982 to 2008). The results of the searches were combined in an Endnote X software file.* Additionally, hand searches of journals, references lists, and textbooks related to ultrasound imaging were performed. We also searched personal files and tracked references of the retrieved studies via the Web of Science Index.

There were no language restrictions. A record was kept of the number of articles retrieved and the number of articles included. The search terms are displayed in Appendix 1.

Study Selection

To be included in the systematic review, a study had to meet 2 criteria: (1) the study had to involve any type of reliability or agreement of any type of ultrasound measurements (B or M mode) for any of the abdominal wall muscles, and (2) the characteristics of the participants had to be described (eg, individuals who were healthy, patients with LBP). Relevant studies were identified by one of the authors (L.O.P.C.) and admitted to the study with agreement from a second author (R.J.E.M.S.).

Data Extraction

Data from eligible studies were extracted by 2 independent reviewers (L.O.P.C. and R.J.E.M.S.). Appendix 2 presents all items that were extracted from the studies.

* Thomson Reuters, 3 Times Square, New York, NY 10036.

Table 1.
Item 6 (Reproducibility) of the *Quality Criteria for Measurement Properties*[11,a]

Property	Definition	Quality Criteria
Reliability	The extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error)	+ ICC or kappa ≥.70[b] ? Doubtful design or method[c] (eg, time interval not mentioned, inadequate description of the ICC tests) − ICC or kappa <.70, despite adequate design and method 0 No information found on reliability
Agreement[d]	The extent to which the scores on repeated measures are close to each other (absolute measurement error)	+ MIC > SDC *or* MIC outside the LOA *or* convincing arguments that agreement is acceptable ? Doubtful design or method[c] *or* MIC not defined and no convincing arguments that agreement is acceptable − MIC ≤ SDC or MIC inside LOA, despite adequate design and method 0 No information found on agreement

[a] +=positive rating, ?=indeterminate rating, −=negative rating, 0=no information available. ICC=intraclass correlation coefficient, MIC=minimal important change, SDC=smallest detectable change, LOA=limits of agreement.
[b] In case of multiple reliability tests, the study will be rated + only if 75% or more of the tests achieved the benchmark of 0.70.
[c] Doubtful design or method = lacking a clear description of the design or methods of the study, sample size smaller than 50 subjects (should be at least 50 in every subgroup analysis),[36] or any important methodological weakness in the design or execution of the study.
[d] MIC was not considered in the ratings of agreement because it is related more to self-report patient outcome measures than to physiological measures such as rehabilitative ultrasound imaging.

The reliability and agreement indexes were extracted for 3 different measures: (1) thickness (ie, static measures of muscle thickness at rest or contracted), (2) thickness changes (ie, measuring muscle activity by determining the degree of change in thickness between the resting and contracted states), and (3) differences in thickness changes over time (ie, measuring improvements or deterioration of muscle activity described above). We considered thickness changes and differences in thickness changes to be the most important measures because they reflect the measures used in current clinical practice.

The studies also were divided according to the study design into 3 categories: (1) the study reported the reproducibility of taking repeated measurements of the same set of images, (2) the study reported the reproducibility of repeating the total measurement procedure (ie, positioning the participant, positioning the ultrasound transducer, acquiring the images, and measuring the images), and (3) the study reported the reproducibility of a portion of the whole procedure (eg, keeping the patient in the same position, but repositioning the transducer and acquiring new images for analysis).

Quality Assessment

The quality of the studies was rated, using item number 6 (reproducibility) from the *Quality Criteria for Measurement Properties*[11] (Tab. 1), by 2 independent reviewers (L.O.P.C. and R.J.E.M.S.). This item evaluates the design of the study, as well as the reproducibility values, by analyzing 2 dimensions: reliability and agreement. These criteria form a checklist that considers both the methodological quality of the reproducibility testing and the results from the testing and, therefore, is somewhat different from scales used to measure the methodological quality of clinical trials.[19]

Results

From the search strategy, 315 potentially relevant studies were found. From these, only 21 studies were considered eligible for data analysis (Figure), being 17 full manuscripts from peer-reviewed journals and 4 abstracts from conference presentations.[20–23] Twelve studies calculated the reproducibility of the whole process of measuring (ie, re-positioning the participant and ultrasound transducer, obtaining the images and measuring them),[8,14–17,20,22,24–28] with intraclass correlation coefficients (ICCs) ranging from .81 to .92 for static images and from .26 to .85 for thickness changes. Of the studies that did not evaluate the whole protocol, 6 studies calculated the reproducibility of measures from the same images only,[22,23,27–30] with ICCs ranging from .62 to .99 for static images and from .48 to .78 for thickness changes; 2 studies calculated partially the process of positioning the participant and the transducer, but fully repeated the process of obtaining the images and measuring them,[31,32] with ICCs ranging from .81 to .92 (static images only, no values for thickness changes were found); and 4 studies could not be classified due to unclear reporting.[21,33–35] The ICC values arose from different ICC models and, therefore, caution should be taken in interpreting these ranges. Given the heterogeneity of the studies in terms of

Table 2.
Description of the Eligible Studies[a]

Study	Are Patients Seeking Care for LBP?	Description of the Sample	Description of the Assessor	Ultrasound Mode	Muscle Task	Muscles Investigated	Interval	Order of Tests	Blinding
Ainscough-Potts et al, 2006[33]	No	30 subjects who were healthy (physical therapist students and staff with no history of LBP in the past 6 months), but only 10 subjects were required for the reliability testing	No information	B	Automatic (1) Supine (2) Sitting on a chair (3) Sitting on a ball (4) Sitting on a ball while lifting the left foot by approximately 10 cm Images were taken at the end of inspiration and expiration	TrA, OI (right side)	No information	Random order using a random square Latin table	No information
Beazell et al, 2006[23]	Mixed	19 subjects who were healthy and 20 patients with LBP	No information	No information	Voluntary (1) Abdominal drawing-in maneuver (2) "First four components of the Abdominal Muscle Strength Test (AMST)"	TrA, OI, OE (right side for subjects who were healthy and affected side for patients with LBP)	"Over three treatment sessions"	No information	No information
Bunce et al, 2002[24]	No	22 adults with no history of LBP in the past 6 months	No information	M	Automatic (1) Supine (2) Standing (3) Walking on a 3-kph treadmill	TrA	"Three separate days"	No information	No information
Critchley and Coutts, 2002[14]	No	"10 different subjects"	No information	B	Rest images only (10 times/patient) in 4-point kneeling	TrA, OI, OE	Immediately after	No information	No information
Ferreira et al, 2003[22]	Yes	20 patients with LBP	"One of the testers was trained for 3 months prior to data collection, while the other had no previous training."	B	No information	TrA	No information	No information	No information
Hides et al, 2007[27]	No	19 subjects with no previous history of LBP	"A physiotherapist underwent training and performed all of subsequent measurements."	B	Voluntary "Subjects were positioned in a supine hook-lying position with their hips in 45 degrees of flexion." (1) Rest (2) Abdominal drawing-in maneuver (the procedure was repeated 6 times/patient)	TrA, OI (both sides)	(1) Measured the same image 3 times (ie, no interval) (2) Comparison of 3 images from the same task, same day (ie, interval = immediately after) (3) 4–7 days interval	Randomized	No information

(Continued)

Table 2.
Continued

Study	Are Patients Seeking Care for LBP?	Description of the Sample	Description of the Assessor	Ultrasound Mode	Muscle Task	Muscles Investigated	Interval	Order of Tests	Blinding
John and Beith, 2007[31]	No	24 subjects who were healthy	No information	B and M	Automatic Rest in crook-lying position	OE	Immediately after (3 images)	No information	No information
Kidd et al, 2002[15]	No	11 subjects without a history of LBP	No information	M	Abdominal drawing-in maneuver (sitting and standing, 4 times each task)	TrA	"2 separate occasions"	No information	No information
Kiesel et al, 2008[29]	No	8 subjects who were asymptomatic	No information	B	Voluntary Abdominal wall drawing-in maneuver in prone position (rest and activated)	TrA	No information	No	No information
Kiesel et al, 2007[25]	56 patients were seeking care + 20 controls	15 subjects from the main study (unclear whether they were patients or controls)	No information	B	Abdominal drawing-in maneuver (rest and activated) in supine hook-lying position	TrA	Same day	No information	Yes
Mannion et al, 2008[16]	14 patients who were seeking care + 14 controls	14 patients with chronic LBP + 14 controls	No information	M	Voluntary Abdominal hollowing in supine hook-lying position (hips in 30° of flexion) (rest and activated)	TrA, OI, OE	7±2 days	No information	No information
McMeeken et al, 2004[8]	No	13 subjects who were healthy	No information	B and M	Voluntary Abdominal hollowing in supine position and knees bent to 20 degrees of flexion (8 times)	TrA	7 days	Randomized	No information
Misuri et al, 1997[34]	No	6 male subjects who were healthy	No information	B	Voluntary Seated in a high-backed armchair at 90 degrees (1) Functional residual capacity (2) Breath holding at residual volume (3) Total lung capacity	TrA, OI , OE, RA	No information	No information	No information
Norasteh et al, 2007[17]	No	27 subjects who were healthy + 12 patients with acute LBP	No information	B	Automatic (1) Supine (2) Standing (3) Sitting	TrA, OI, OE, RA	7 days	No information	No information

(Continued)

Table 2.
Continued

Study	Are Patients Seeking Care for LBP?	Description of the Sample	Description of the Assessor	Ultrasound Mode	Muscle Task	Muscles Investigated	Interval	Order of Tests	Blinding
Pietrek et al, 2000[21]	No	12 subjects who were healthy	No information	No information	Voluntary Isometric trunk flexion, extension, and rotation tasks with 4 levels of exertion	TrA, OI, OE	No information	Randomized	No information
Rankin et al, 2006[26]	No	10 subjects	No information	B	Rest	TrA, OE, OI	7 days	No information	No information
Roddey et al, 2007[32]	No	70 subjects who were healthy	No information	B	Voluntary Lying supine with knees flexed Rest and contracting TrA by "performing pelvic floor contraction or bracing contraction"	TrA	Immediately after	No information	No information
Springer et al, 2006[35]	No	32 Department of Defense beneficiaries who were healthy (no history of LBP in the last 3 years)	"The examiner consisted of a senior physical therapist with 18 years of clinical experience and a student, who were provided a training session."	B	Voluntary Abdominal drawing-in maneuver and rest in supine hook-lying posture	TrA and total abdominal muscle thickness	Immediately after	Randomized	Not for the measurements (assessors were blinded to the treatment allocation)
Teyhen et al, 2005[28]	Yes	30 subjects who had been seeking care for LBP within the previous 3 months	"Physical therapy students who were provided a 3-hour training session in the measurement procedures."	B	Rest in supine hook-lying position Reliability was calculated with TrA images in rest + of total abdominal muscle thickness	TrA, and total lateral abdominal muscle thickness (TrA+OI+OE)	Immediately after	No information	No blinding for images
Teyhen et al, 2008[30]	No	10 subjects who were healthy	No information	B	Abdominal drawing-in maneuver, abdominal crunch, abdominal sit-back, quadruped opposite upper and lower extremity lift	TrA and OI	Same day	No information	No information
Toma et al, 2006[20]	No	16 subjects	No information	M	Rest and abdominal hollowing in lying position (5 times); total muscle thickness at maximal TrA contraction	TrA, OI and OE	1-day interval	No Information	No information

[a] TrA=transversus abdominis muscle, OI=internal oblique muscle, OE=external oblique muscle, B=bright mode, M=motion mode, LBP=low back pain.

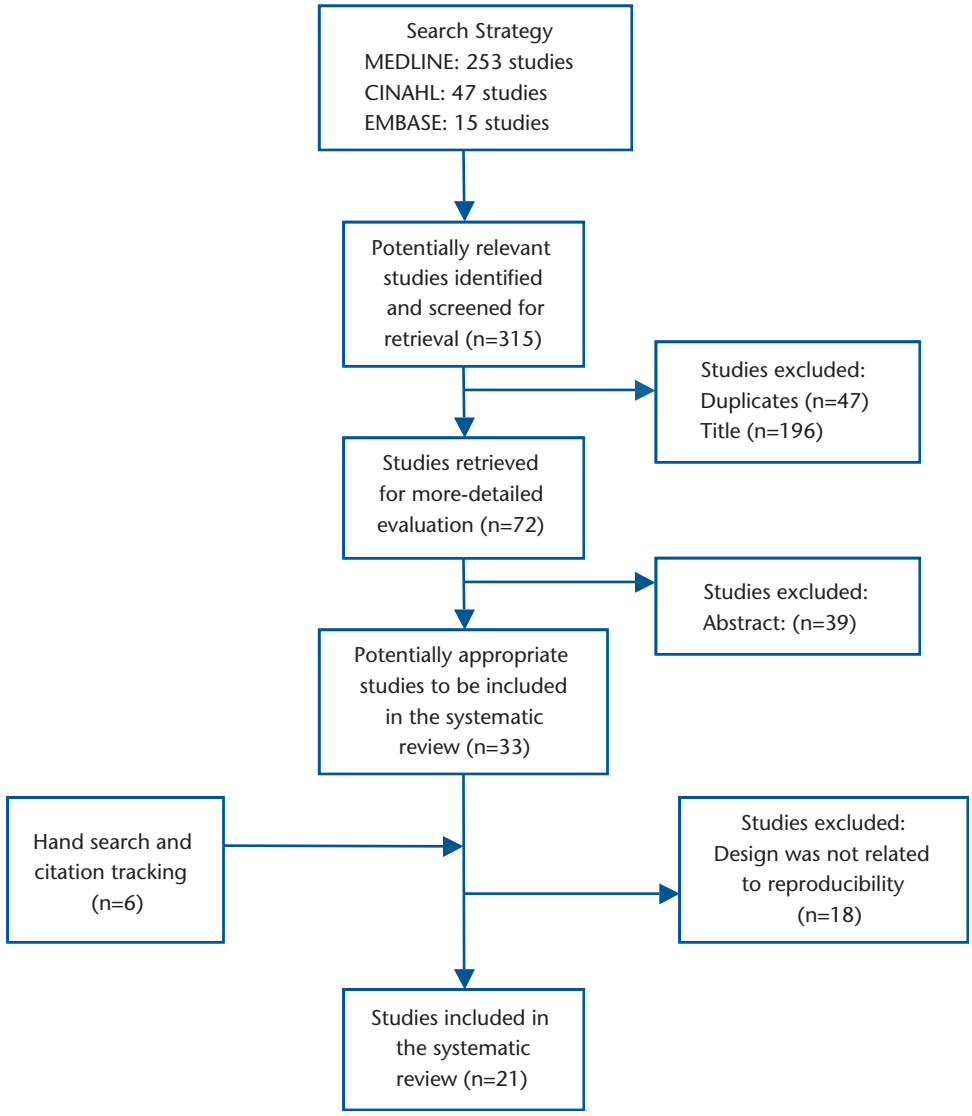

Figure.
QUORUM (Quality of Reporting of Meta-analyses) flow diagram of the literature search.

design and statistical analysis, pooling of the data for a meta-analysis approach was not possible; therefore, the presentation of our results is descriptive.

Table 2 describes the characteristics of the eligible studies. There were large differences in the sample sizes used in the reproducibility tests, ranging from 8 to 70 participants. Remarkably, only 2 studies recruited patients seeking care for their LBP, 3 studies recruited a mixed sample of patients with LBP and volunteers who were asymptomatic, and 16 studies recruited only subjects who were healthy. Information about the assessors, blinding, and how the order of the tests was controlled for bias was presented in only 4, 1, and 4 studies, respectively. Sixteen studies used B mode ultrasound for collecting the images, and the transversus abdominis was the most commonly investigated muscle (20 of 21 studies). Finally, the time interval between tests ranged from "immediately after" to 7 days, and 13 studies investigated reproducibility while performing a voluntary task.

Table 3 describes the results of the eligible studies. Although all studies investigated the reliability of measurements of abdominal muscle thickness, only 6 studies investigated the reliability of thickness changes (reflecting muscle activity), and none tested the reproducibility of the differences in thickness changes over time (reflecting improvement/deterioration in muscle activity). The ICC was used to assess reliability in

Table 3.
Reproducibility Values for Each Study[a]

Study	Reliability for Single Measures	Reliability for Changes in Thickness	Notes
Ainscough-Potts et al, 2006[33]	Intrarater reliability (ICC) Inspiration: .97 Expiration: .99	No information	The study did not specify to which muscles the ICC values pertain ICC type not specified
Beazell et al, 2006[23]	Intrarater reliability (ICC [3,1]) .94 to .99	No information	The study did not specify to which muscles the ICC values pertain
Bunce et al, 2002[24]	Intrarater reliability, TrA (ICC [1,1]; SEM [mm]) Supine: .94; .35 Standing: .88; .66 Walking: .88; .56	TrA ICC (1,1) Between supine and standing: .78 Between supine and walking: .48	
Critchley and Coutts, 2002[14]	Intrarater reliability (ICC [1,1]; SEM [mm]) OE: .95; .66 OI: .98; .80 TrA: .94; .60	No information	
Ferreira et al, 2003[22]	Interrater reliability, same image (ICC [2,1]) .70 to .98	Intrarater reliability (ICC [2,1]) Experienced rater: .85 Inexperienced rater: .28 Interrater reliability: .26	The study did not specify to which muscles the ICC values pertain
Hides et al, 2007[27]	Intrarater reliability, same image (ICC [3,1]; 95% CI; SEM [mm]) IO rest: .99; .97 to 1.00; .028 IO contracted: .98; .95 to 1.00; .033 TrA rest: .98; .95 to 1.00; 0.01 TrA contracted: .97; .98 to 1.00; .020 "Across 3 images" (ICC [3,4]; 95% CI; SEM [mm]) IO rest: .82; .55 to .95; .157 IO contracted: .66; .23 to .92; .385 TrA rest: .62; .32 to .85; .247 TrA contracted: .80; .56 to .93; .224 2-day interval (ICC [3,6]; 95% CI; SEM [mm]) IO rest: .69; .30 to .92; .295 IO contracted: .63; .21 to .94; .432 TrA rest: .85; .42 to .98; .089 TrA contracted: .84; .52 to .96; .176	No information	
John and Beith, 2007[31]	OE intrarater reliability (ICC): .92	No information	ICC type not specified
Kidd et al, 2002[15]	TrA, same day (ICC [1,1]; SEM [mm]; SEM [%]) "Range from .79 to .96"; 0.29 to 0.57; 3.7 to 8.9 TrA, between days (ICC [1,1]; SEM [mm]; SEM [%]) Sitting: .96; 0.18; 1.2 Standing: .88; 0.33; 3.6	No information	
Kiesel et al, 2008[29]	TrA ICC [3,1]: .95	No information	

(Continued)

Table 3.
Continued

Study	Reliability for Single Measures		Reliability for Changes in Thickness		Notes
Kiesel et al, 2007[25]	TrA (ICC [3,3]; 95% CI; SEM [cm]; MDC [cm]) Rest: .98; .91 to .99; 0.01; 0.03 Contracted: .97; .91 to .98; 0.02; 0.06		TrA % change (ICC [3,3]; 95% CI; SEM [%]; MDC [%]) .96; .91 to .99; 6.26; 17.34		
Mannion et al, 2008[16]	Controls (ICC [3,1]; SEM [mm]; CV) TrA rest Left: .86; 0.40; 10.7 Right: .83; 0.40; 10.3 TrA Max Left: .75; 0.65; 12.0 Right: .78; 0.58; 10.7 OI rest Left: .94; 0.72; 9.8 Right: .92; 0.58; 8.8 OE rest Left: .59; 1.03; 19.6 Right: .26; 0.84; 17.0	Patients with LBP (ICC [3,1]; SEM [mm]; CV) TrA rest Left: .63; 0.46; 11.5 Right: .89; 0.27; 7.2 TrA maximum Left: .41; 0.78; 14.3 Right: .88; 0.41; 7.7 OI rest Left: .85; 0.68; 9.8 Right: .73; 0.82; 10.8 OE rest Left: .51; 0.84; 14.6 Right: .42; 1.20; 19.1	Controls (ICC [3,1]; SEM [mm]; CV) TrA contraction ratio Left: .50; 0.16; 10.9 Right: .52; 0.16; 11.4 OE contraction ratio Left: .60; 0.06; 5.8 Right: .66; 0.04; 4.4 OE+OI contraction ratio Left: .61; 0.03; 3.2 Right: .72; 0.04; 3.9 TrA preferential activation ratio Left: .55; 0.02; 38.0 Right: .62; 0.02; 30.2	Patients with LBP (ICC [3,1]; SEM [mm]; CV) TrA contraction ratio Left: .28; 0.16; 11.6 Right: .80; 0.09; 6.0 OE contraction ratio Left: .57; 0.05; 5.6 Right: .43; 0.05; 5.4 OE+OI contraction ratio Left: .39; 0.05; 5.6 Right: .25; 0.05; 4.5 TrA preferential activation ratio Left: .32; 0.03; 49.5 Right: .48; 0.02; 27.4	
McMeeken et al, 2004[8]	TrA (ICC; 95% CI) B mode: .99; .96 to .99 M mode: .98; .94 to .99 Linear versus curvilinear transducers: .82, .87 to .96 Bland and Altman tests (mean difference [mm]; SD of difference [mm]; 95% limits of agreement [mm]; repeatability coefficient) B mode: 0.03; 0.03; −0.17 to 0.24; 0.023 M mode: 0.04; 0.04; −0.23 to 0.03; 0.038 Linear versus curvilinear transducers: −0.14; 0.14; −0.95 to 0.66; 0.045		No information		ICC type not specified
Misuri et al, 1997[34]	CV=0% to 15.7% "For each muscle, between-subject variability was highly significant (F ranging from 9.1 to 273, *P* values ranging from .003 to .00001), whereas intrasubject variability was not (F ranging from 0.54 to 3.2, *P* values ranging from .66 to .06)."		No information		The study presented CV for individual patients for each task
Norasteh et al, 2007[17]	Subjects who were asymptomatic (ICC [1,2]; ICC [1,3]; SEM) OE: .96; .72; 0.33 OI: .97; .91; 0.07 TrA: .81; .80; 0.45 RA: n/a; .85; 0.84 Patients with acute LBP (ICC [2,1]; SEM) OE: .87; 0.35 OI: .87; 0.31 TrA: .91; 0.30		No information		No unit of measurement was provided
Pietrek et al, 2000[21]	Pearson *r*: .55 to .97		No information		

(Continued)

Table 3.
Continued

Study	Reliability for Single Measures	Reliability for Changes in Thickness	Notes
Rankin et al, 2006[26]	Between scans, ICC [1,1]=.98 to .99; 95% CI=.91 to 1.00 Between days, ICC [1,2]=.96 to .99; 95% CI=.85 to 1.00 "Bland and Altman tests produced mean differences close to zero, and SD difference values were very low." 95% limits of agreement (cm) OI: 0.22 OE: 0.12 TrA: 0.09	No information	
Roddey et al, 2007[32]	Deferred assessment of TrA (ICC [2,1]; SEM [mm]) Right relaxed: .84; 0.04 Right contracted: .83; 0.07 Left relaxed: .90; 0.03 Left contracted: .91; 0.06 Immediate assessment of TrA (ICC [2,1]; SEM [mm]) Right relaxed: .83; 0.03 Right contracted: .81; 0.09 Left relaxed: .93; 0.02 Left contracted: .92; 0.04	No information	
Springer et al, 2006[35]	TrA single measure (ICC [2,1]; 95% CI; SEM [mm]) Rest: .93; .86 to .96; 0.32 Contracted: .96; .92 to .98; 0.45 TrA average of 3 measures (ICC [2,3]; 95% CI; SEM [mm]) Rest: .98; .92 to .99; 0.13 Contracted: .99; .98 to .99; 0.20 TrA+OE+OI single measure (ICC [2,1]; 95% CI; SEM [mm]) Rest: .98; .96 to .99; 0.80 Contracted: .99; .98 to 1.00; 0.71 TrA+OE+OI average of 3 measures (ICC [2,3]; 95% CI; SEM [mm]) Rest: 1.00; .99 to 1.00; 0.35 Contracted: 1.00; .99 to 1.00; 0.34	TrA/total single measure (ICC [2,1]; 95% CI; SEM [%]) Rest: 0.91; .82 to .95; 2 Contracted: .98; .96 to .99; 1.2 TrA/total average of 3 measures (ICC [2,3]; 95% CI; SEM [%]) Rest: .99; .97 to .99; 0.5 Contracted: .99; .99 to 1.00; 0.7	The ratios presented (TrA/total) were calculated from values of the same image
Teyhen et al, 2005[28]	TrA (ICC [3,1]; 95% CI; SEM [cm]; CV) Intra-image: .98; .96 to .99; 0.013; 5 Inter-image: .93; .77 to .99; 0.031; 11 TrA+OE+OI (ICC [3,1]; 95% CI; SEM [cm]; CV) Intra-image: .99; .99 to 1.0; 0.018; 5 Inter-image: .97; .77 to .99; 0.087; 14	No information	The values from TrA and total muscle thickness (TrA+OE+OI thickness) were obtained from the same image
Teyhen et al, 2008[30]	ICC [2,2] "greater or equal to .95" SEM (mm) TrA: 0.09 OI: 0.29		
Toma et al, 2006[20]	TrA (ICC; SEM [mm]; SEM [%]) .70 to .94; 0.44 to 0.74; 8 to 15	TrA ICC Left: .44 Right: .70	ICC type not specified

[a] TrA=transversus abdominis muscle, OI=internal oblique muscle, OE=external oblique muscle, B=bright mode, M=motion mode, LBP=low back pain, ICC=intraclass correlation coefficient, MDC=minimum detectable change, SD=standard deviation, SEM=standard error of the measurement, CI=confidence internal, CV=coefficient of variation, RA=rectus abdominis muscle, n/a=not applicable.

Table 4.
Quality of Reliability for Abdominal Muscle Thickness and Thickness Changes and Agreement Rated by the Quality Criteria[a]

Study	Reliability for Thickness	Reliability for Changes in Thickness	Agreement	Notes
Ainscough-Potts et al, 2006[33]	?	0	0	No time interval provided, did not provide ICC type, underpowered
Beazell et al, 2006[23]	?	0	0	Underpowered
Bunce et al, 2002[24]	?	?	?	Only 50% of the ICC values were above .70, underpowered
Critchley and Coutts, 2002[14]	?	0	?	Underpowered
Ferreira et al, 2003[22]	?	–	0	No time interval provided, only 33% of the ICC values were above .70, underpowered
Hides et al, 2007[27]	–	0	?	Across 3 images and 2 days: only 50% of the ICC values were above .70, underpowered
John and Beith, 2007[31]	?	0	0	Did not provide ICC type, underpowered
Kidd et al, 2002[15]	?	0	?	Underpowered
Kiesel et al, 2008[29]	?	0	0	No time interval provided, underpowered
Kiesel et al, 2007[25]	?	?	?	Underpowered
Mannion et al, 2008[16]	?	–	0	ICC values for reliability of thickness changes were below .70, underpowered
McMeeken et al, 2004[8]	?	0	?	Did not provide ICC type, underpowered
Misuri et al, 1997[34]	?	0	0	Used CV only, underpowered
Norasteh et al, 2007[17]	?	0	?	Underpowered
Pietrek et al, 2000[21]	?	0	0	Used Pearson *r*, no time interval provided, underpowered
Rankin et al, 2006[26]	?	0	?	Underpowered
Roddey et al, 2007[32]	?	0	?	The time interval (immediately after) could inflate the reliability
Springer et al, 2006[35]	?	?	?	The time interval (immediately after) could inflate the reliability, underpowered
Teyhen et al, 2005[28]	?	0	?	The time interval (immediately after) could inflate the reliability, underpowered
Teyhen et al, 2008[30]	?	0	?	Underpowered
Toma et al, 2006[20]	?	?	?	Did not provide ICC type, underpowered

[a] ICC=intraclass correlation coefficient, CV=coefficient of variation, ?=indeterminate rating, –=negative rating, 0=no information available.

18 studies; however, a range of forms of this statistic were used. Five studies used the ICC (1,k), 2 studies used the ICC (2,k), 6 studies used the ICC (3,k), and 4 studies did not specify which type of ICC was chosen. Unfortunately, only 5 studies provided confidence intervals for the ICC. Some studies used Pearson *r* or coefficient of variation (CV) as a measure of reliability. Agreement was calculated in 12 studies (12 for abdominal muscle thickness and 3 for thickness changes). Most of the studies used the standard error of the measurement (SEM) as the agreement parameter, 2 studies used Bland and Altman plots, and 1 study calculated the minimum detectable change (MDC).

In terms of reliability, although more than 80% of the ICC values for measuring abdominal muscle thickness ranged from .80 to 1.00, most ICC values for measuring changes in thickness were less then .70. Interestingly, the ICC values tend to be slightly lower in the 5 studies that used participants with LBP compared with the 16 studies that recruited only participants who were asymptomatic (Tab. 3).

Table 4 presents the quality assessment of the 21 studies, summarizing each criterion as positive, doubtful, negative, or no information. None of the studies demonstrated positive ratings for both reliability and agreement. For the reliability of measurements of abdominal muscle thick-

ness, no studies were rated as positive, 20 were rated as doubtful, and 1 was rated as negative. For the reliability of measurements of thickness change, no study out of 6 was rated as positive, 2 were rated as negative, and 4 were rated as doubtful. Twelve studies provided some information in regard to agreement (SEM, Bland and Altman plots, or smallest detectable change), with all of them rated as doubtful. The major reasons for doubtful and negative ratings were the lack of precise information about the time interval between the measures, lack of precise information about the ICC type, lack of statistical power (ie, fewer than 50 subjects for the analysis),[36] and ICC values below .70.

Discussion and Conclusions

This review highlights the limitations of existing research evaluating the reproducibility of RUSI measures of abdominal wall muscles. Few studies analyzed the reproducibility for the measurement of thickness change, and no studies evaluated the reproducibility of the difference in thickness change over time. The available studies were frequently of low quality, recruited subjects who were healthy, and evaluated only a portion of the RUSI measurement protocol. The existing data are of limited value in estimating the reproducibility of RUSI measures undertaken in a clinical setting to guide a motor control exercise program for people with LBP.

The whole process of performing ultrasound measurements has multiple sources of error (eg, accuracy of measurements of distance, identification of landmarks, ability to perform the tasks properly, and position of patient and transducer). Additionally, it has to be acknowledged that trial-to-trial variation in performance of the activation tasks is expected. It would be useful to consider whether modifications in the test protocol may enhance the reproducibility of ultrasound measures (especially for thickness changes). One approach that has been shown to enhance reliability of other low back assessments is to further standardize the protocol.[37–39]

The main methodological weaknesses found in the studies can be summarized into 4 main issues: (1) generalizability of findings (due to sampling issues and to the description of the assessor), (2) inadequate statistics, (3) bias (due to absence of blinding or to not controlling the order of the tests), and (4) the lack of studies that investigated the reliability and especially the agreement of thickness changes and differences in thickness changes over time.

The generalizability of the findings from the individual studies selected from this systematic review is substantially limited given the fact that from 21 studies, only 2 recruited patients with LBP[22,28] and 3 recruited a mixed sample of patients with LBP and individuals who were healthy.[16,23,25] Additionally, only 4 studies provided some description or source of the assessors.[22,27,28,35] Moreover, from the studies cited above, only 2 provided a description of the assessor and recruited patients seeking care for LBP.[22,28] We believe that clinically relevant studies must recruit participants seeking care for the condition in which the test will be used and include a description of the assessor to enable better understanding of the assessor's qualifications, skills, and length of training for future clinical comparisons.

The most widely used statistical test for the calculation of reliability was the ICC (18 of 21 occasions), which is a recommended option for testing reliability for continuous scales[40] (which is the case in studies of muscle thickness). There are multiple types of ICC, and the choice of the correct ICC model depends on 3 considerations[40,41]: (1) the wish to generalize, or not, the findings to other assessors, (2) whether the same set of assessors rate each subject, and (3) whether the authors are interested in the reliability of the ratings of an individual assessor or the reliability of the mean rating of a group of assessors.

Often, in the case of clinical research, only one judge is used, and it is important to generalize the results. Therefore, ICC type 2 (ie, ICC [2,1]) or type 1 (ie, ICC [1,1]) is preferred, and ICC (3,1) or ICC (3,k) (k=number of assessors) should be used only if the authors do not want to generalize their results.[42] We found 4 studies that provided no information on the ICC type,[8,20,31,33] and 6 studies used ICC type 3 only as their reliability index.[16,23,25,27–29] Pearson *r* was used as a measure of reliability in one study,[21] and the CV was used as a measure of reliability in one study.[34] However the use of Pearson *r* and CV are likely to provide overly optimistic estimates of reliability, as they do not consider the "between-judges variance."[40,41] We considered that investigators should used ICC (2,1) or ICC (1,1) as a measure of reliability for thickness or thickness changes, as they should provide the most relevant estimation of reliability. Agreement was analyzed by 12 studies (10 studies calculated the SEM, and 2 studies used Bland and Altman plots), and the evidence for all studies was classified as doubtful due to small sample sizes and small time interval between the measures.

An issue that needs to be borne in mind is that many of the studies reported the reproducibility of the mean of replicate measures. Although this is an accepted method of enhancing the reliability of a measure, it does make the measurement protocol more time-consuming. In a

similar fashion, it needs to be remembered that some studies used highly trained raters, intricate equipment to control the position of the participant, and load cells to standardize the activation of trunk muscles. Although each of these 3 elements makes sense, they again limit the generalizability of the results.

Surprisingly, we found that most of the studies calculated the reproducibility of measurements of abdominal muscle thickness only. The problem with this approach is that in a clinical setting, the most important measure would be either thickness changes (comparing one image in rest state with another image during muscle activity), which was calculated in 6 studies,[16,20,22,24,25,35] or differences in thickness changes over time (comparing thickness changes at 2 different time points for quantifying improvement or deterioration), which was not performed in any study. We suggest strongly that more studies investigating the reproducibility of thickness changes and differences in thickness changes over time should be undertaken.

We believe that our study provides important information for clinicians and researchers about the use of RUSI for abdominal wall muscles. It is important for clinicians to understand the limited evidence for reproducibility of the measurements made over time when used to document success of a motor control treatment program. Additionally, researchers have to acknowledge that the most important clinical questions about reproducibility of RUSI for abdominal wall muscles have been not answered, and further studies are urgently needed.

All authors provided concept/idea/research design, writing, and data analysis. Dr Costa and Dr Smeets provided data collection. Dr Latimer provided consultation (including review of manuscript before submission).

Mr Costa is supported by CAPES, Ministério da Educação, Brazil, and Pontifícia Universidade Católica de Minas Gerais, Brazil. Dr Maher holds a research fellowship funded by the National Health and Medical Research Council of Australia.

This article was received October 19, 2008, and was accepted April 7, 2009.

DOI: 10.2522/ptj.20080331

References

1 O'Sullivan P. Diagnosis and classification of chronic low back pain disorders: maladaptive movement and motor control impairments as underlying mechanism. *Man Ther.* 2005;10:242–255.

2 Hodges PW, Richardson CA. Feedforward contraction of transversus abdominis is not influenced by the direction of arm movement. *Exp Brain Res.* 1997;114:362–370.

3 Hodges PW, Richardson CA. Altered trunk muscle recruitment in people with low back pain with upper limb movement at different speeds. *Arch Phys Med Rehabil.* 1999;80:1005–1012.

4 Tsao H, Hodges PW. Persistence of improvements in postural strategies following motor control training in people with recurrent low back pain. *J Electromyogr Kinesiol.* 2008;18:559–567.

5 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.

6 Teyhen DS. Rehabilitative Ultrasound Imaging Symposium. *J Orthop Sports Phys Ther.* 2006;36:A1–A17.

7 Ferreira PH, Ferreira ML, Hodges PW. Changes in recruitment of the abdominal muscles in people with low back pain: ultrasound measurement of muscle activity. *Spine.* 2004;29:2560–2566.

8 McMeeken JM, Beith ID, Newham DJ, et al. The relationship between EMG and change in thickness of transversus abdominis. *Clin Biomech.* 2004;19:337–342.

9 Hodges PW, Pengel LHM, Herbert RD, et al. Measurement of muscle contraction with ultrasound imaging. *Muscle Nerve.* 2003;27:682–692.

10 May S, Littlewood C, Bishop A. Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review. *Aust J Physiother.* 2006;52:91–102.

11 Terwee CB, Bot SDM, Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34–42.

12 de Vet HCW, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *J Clin Epidemiol.* 2006;59:1033–1039.

13 Krebs DE. Declare your ICC type [letter to the editor]. *Phys Ther.* 1986;66:1431.

14 Critchley DJ, Coutts FJ. Abdominal muscle function in chronic low back pain patients: measurement with real-time ultrasound scanning. *Physiotherapy.* 2002;88:322–332.

15 Kidd AW, Magee S, Richardson CA. Reliability of real-time ultrasound for the assessment of the transversus abdominis function. *J Grav Physiol.* 2002;9:131–132.

16 Mannion AF, Pulkovski N, Gubler D, et al. Muscle thickness changes during abdominal hollowing: an assessment of between-day measurement error in controls and patients with chronic low back pain. *Eur Spine J.* 2008;17:494–501.

17 Norasteh A, Ebrahimi E, Salavati M, et al. Reliability of B-mode ultrasonography for abdominal muscles in asymptomatic and patients with acute low back pain. *J Body Mov Ther.* 2007;11:17–20.

18 Teyhen DS, Gill NW, Whittaker JL, et al. Rehabilitative ultrasound imaging of the abdominal muscles. *J Orthop Sports Phys Ther.* 2007;37:450–466.

19 Maher CG, Sherrington C, Herbert RD, et al. Reliability of the PEDro scale for rating quality of randomized controlled trials. *Phys Ther.* 2003;83:713–721.

20 Toma V, Pulkovski N, Sprott H, et al. Reliability of measures of abdominal muscle thickness during abdominal hollowing as assessed with M-mode ultrasound. *Eur J Pain.* 2006;10:S109.

21 Pietrek M, Sheikhzadeh A, Hagins M, et al. Evaluation of abdominal muscles by ultrasound imaging: reliability, and comparison to electromyography. *Eur Spine J.* 2000;9:309.

22 Ferreira PH, Ferreira ML, Maher CG, et al. Clinical ultrasound test for transversus abdominis thickness: investigation of reliability. Presented at: 13th Biennial Conference—Musculoskeletal Physiotherapy Australia; November 27–30, 2003; Sydney, New South Wales, Australia.

23 Beazell JR, Grindstaff TL, Magrum EM, et al. Comparison of clinical test and real time ultrasound evaluation of muscle contraction in normals and patients with low back pain. *J Man Manip Ther.* 2006;14:168–169.

24 Bunce SM, Moore AP, Hough AD. M-mode ultrasound: a reliable measure of transversus abdominis thickness? *Clin Biomech.* 2002;17:315–317.

25 Kiesel KB, Underwood FB, Mattacola CG, et al. A comparison of select trunk muscle thickness change between subjects with low back pain classified in treatment-based classification system and asymptomatic controls. *J Orthop Sports Phys Ther.* 2007;37:596–607.

26 Rankin G, Stokes M, Newham DJ. Abdominal muscle size and symmetry in normal subjects. *Muscle Nerve.* 2006;34:320–326.

27 Hides JA, Miokovic T, Belavy DL, et al. Ultrasound imaging assessment of abdominal muscle function during drawing-in of the abdominal wall: an intrarater reliability study. *J Orthop Sports Phys Ther.* 2007;37:480–486.

28 Teyhen DS, Miltenberger CE, Deiters HM, et al. The use of ultrasound imaging of the abdominal drawing-in maneuver in subjects with low back pain. *J Orthop Sports Phys Ther.* 2005;35:346–355.

29 Kiesel KB, Uhl T, Underwood FB, et al. Rehabilitative ultrasound measurement of select trunk muscle activation during induced pain. *Man Ther.* 2008;13:132–138.

30 Teyhen DS, Rieger JL, Westrick RB, et al. Changes in deep abdominal muscle thickness during common trunk-strengthening exercises using ultrasound imaging. *J Orthop Sports Phys Ther.* 2008;38:596–605.

31 John EK, Beith ID. Can activity within the external abdominal oblique be measured using real-time ultrasound imaging? *Clin Biomech.* 2007;22:972–979.

32 Roddey TS, Brizzolara KJ, Cook KF. A comparison of two methods of assessing transverse abdominal muscle thickness in participants using real-time ultrasound in a clinical setting. *Orthop Phys Ther Pract.* 2007;19:198–201.

33 Ainscough-Potts AM, Morrisey MC, Critchley DJ. The response of the transverse abdominis and internal oblique to different postures. *Man Ther.* 2006;11:54–60.

34 Misuri G, Colagrande S, Gorini M, et al. In vivo ultrasound assessment of respiratory function of abdominal muscles in normal subjects. *Eur Respir J.* 1997;10:2861–2867.

35 Springer BA, Mielcarek BJ, Nesfield TK, et al. Relationships among lateral abdominal muscles, gender, body mass index, and hand dominance. *J Orthop Sports Phys Ther.* 2006;36:289–297.

36 Altman DG. *Practical Statistics for Medical Research*. London, United Kingdom: Chapman and Hall, 1991.

37 Chiradejnant A, Maher CG, Latimer J. Objective manual assessment of lumbar PA stiffness is now possible. *J Manip Physiol Ther.* 2003;26:34–39.

38 Maher CG, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Phys Ther.* 1994;74:801–809.

39 Maher CG, Latimer J, Adams R. An investigation of the reliability and validity of posteroanterior spinal stiffness judgments made using a reference-based protocol. *Phys Ther.* 1998;78:829–837.

40 Fleiss J. *The Design and Analysis of Clinical Experiments*. New York, NY: John Wiley & Sons Inc; 1986.

41 Armstrong GD. The intraclass correlation as a measure of interrater reliability of subjective judgments. *Nurs Res* 1981;30:314–315.

42 Laschinger HKS. Intraclass correlations as estimates of interrater reliability in nursing research. *West J Nurs Res* 1992;14:246–249.

Appendix 1.
Search Strategy

1. ultrasound OR ultra-sound OR ultra sound OR scanning OR imaging
2. reliability OR repeatability OR test-retest OR assessment OR evaluation OR examination OR thickness OR activ$ OR function OR change$ OR investigation OR ICC OR limits of agreement OR critical difference
3. Transversus abdominis OR internal oblique OR external oblique OR abdominal muscle$
4. 1 and 2
5. 3 and 4

Appendix 2.
Data Extraction

Description of the sample
Sample size
Ultrasound mode (B or M mode)
Task performed by the participants (eg, abdominal hollowing, rest)
Muscles investigated (eg, transversus abdominis, internal oblique, external oblique)
Length of the interval between the rehabilitative ultrasound imaging assessments
Blinding
Ordering of the tests (eg, alternation, randomization)
Description of the assessor
Description of the type of reliability (ie, intrarater/interrater, intra-image/inter-image)
Reliability and agreement values (eg, intraclass correlation coefficient,[a] kappa, standard error of the measurement, coefficient of variation, Bland and Altman plots)

[a] Ten authors from studies that did not specify the type of intraclass correlation coefficient were contacted by e-mail, and we received responses from 6 of them.[14,15,22,23,26,32]