Quantitative Paraspinal Muscle Measurements: Inter-Software Reliability and Agreement Using OsiriX and ImageJ

Maryse Fortin and Michele C. Battié

PHYS THER. 2012; 92:853-864.
Originally published online March 8, 2012
doi: 10.2522/ptj.20110380

Quantitative Paraspinal Muscle Measurements: Inter-Software Reliability and Agreement Using OsiriX and ImageJ

Maryse Fortin, Michele C. Battié

Background. Variations in paraspinal muscle cross-sectional area (CSA) and composition, particularly of the multifidus muscle, have been of interest with respect to risk of, and recovery from, low back pain problems. Several investigators have reported on the reliability of such muscle measurements using various protocols and image analysis programs. However, there is no standard protocol for tissue segmentation, nor has there been an investigation of reliability or agreement of measurements using different software.

Objective. The purpose of this study was to provide a detailed muscle measurement protocol and determine the reliability and agreement of associated paraspinal muscle composition measurements obtained with 2 commonly used image analysis programs: OsiriX and ImageJ.

Design. This was a measurement reliability study.

Methods. Lumbar magnetic resonance images of 30 individuals were randomly selected from a cohort of patients with various low back conditions. Muscle CSA and composition measurements were acquired from axial T2-weighted magnetic resonance images of the multifidus muscle, the erector spinae muscle, and the 2 muscles combined at L4–L5 and S1 for each participant. All measurements were repeated twice using each software program, at least 5 days apart. The assessor was blinded to all earlier measurements.

Results. The intrarater reliability and standard error of measurement (SEM) were comparable for most measurements obtained using OsiriX or ImageJ, with reliability coefficients (intraclass correlation coefficients [ICCs]) varying between .77 and .99 for OsiriX and .78 and .99 for ImageJ. There was similarly excellent agreement between muscle composition measurements using the 2 software applications (inter-software ICCs=.81–.99).

Limitations. The high degree of inter-software measurement reliability may not generalize to protocols using other commercial or custom-made software.

Conclusion. The proposed method to investigate paraspinal muscle CSA, composition, and side-to-side asymmetry was highly reliable, with excellent agreement between the 2 software programs.

M. Fortin, BSc, CAT(C), PhD student, Faculty of Rehabilitation Medicine, University of Alberta, Edmonton, Alberta, Canada.

M.C. Battié, PhD, Department of Physical Therapy, Faculty of Rehabilitation Medicine, Department of Physical Therapy, University of Alberta, 2-50 Corbett Hall, Edmonton, Alberta, Canada T6G 2H4. Address all correspondence to Dr Battié at: mc.battie@ualberta.ca.

[Fortin M, Battié MC. Quantitative paraspinal muscle measurements: inter-software reliability and agreement using OsiriX and ImageJ. *Phys Ther.* 2012;92:853–864.]

Published Ahead of Print: March 8, 2012
Accepted: March 4, 2012
Submitted: November 5, 2011

Cross-sectional area (CSA) asymmetries of lumbar paraspinal muscles,[1–7] as well as fat infiltration,[8,9] have been associated with low back pain (LBP) and related pathologies using various imaging techniques. As a result, the measurement of paraspinal muscle asymmetry or composition has been emphasized in a number of studies related to the etiology and prognosis of LBP.[1–15] There are inconsistencies, however, in study findings of the association between painful spinal conditions and paraspinal muscle morphology. For example, Ploumis et al[6] used a manual segmenting technique to measure paraspinal muscle *functional CSA* (FCSA), defined as fat-free muscle mass, in a group of 40 patients with monosegmental disk disease and unilateral LBP, with or without radicular symptoms, and reported significant multifidus muscle atrophy on the symptomatic side. Yet, in another magnetic resonance imaging (MRI) study, Hyun et al[10] found no significant asymmetry between involved and uninvolved sides in a group of 39 patients with disk herniation, with or without radiculopathy. They also measured multifidus muscle FCSA, but used a technique to determine the proportion of muscle versus fat tissue based on a signal intensity threshold.

Similarly, 2 studies that quantitatively compared the degree of paraspinal muscle fatty infiltration present in patients with chronic LBP compared with a control group of individuals who were healthy showed conflicting results.[1,2] Different threshold techniques and measurement protocols were used to measure the proportion of muscle fatty infiltration, which may have contributed to the discrepant findings, but the effect of such differences on measurement is not known.

Variations in imaging modalities (MRI, computed tomography scan, and ultrasound), image analysis programs, and measurement protocols contribute to conflicting results. Currently, several methods are used to investigate paraspinal muscle morphology, and too little attention has been given to whether they lead to roughly equivalent measurements. Some investigators have focused on total CSA,[3,4,7,12–14] whereas others contend that FCSA is a better indicator of muscle atrophy and contractibility.[16] Functional CSA is calculated by using either a manual technique or a signal intensity threshold technique with the aid of computer software. Although the reliability of measurements of FCSA using the 2 different approaches has been investigated in several studies,[1,15–19] investigators interested in segmenting paraspinal muscles or fat tissues currently use a variety of computer software, including in-house custom-made software,[1,18] software that is part of an MRI scanner,[20] picture archiving and communications systems workstations,[17,19] commercial software,[10] computer-aided drafting (auto-CAD) software,[3,21] and freeware.[15,16,22] Moreover, the use of proprietary software and insufficient descriptions of measurement protocols hinder replication of results by others, and the comparability of measurements obtained using different software and measurement protocols has been neglected.

Although the measurement error related to the measurement methods used appears to be mostly associated with the observer,[23] the software used also might lead to measurement differences, and there is a need to determine whether direct comparisons can be made among different software packages (using comparable methods). There currently is no standard protocol, and we found no investigations of reliability or agreement among measurements obtained with different software or protocols.

To clarify the measurement error related to use of 2 widely available, free image analysis programs and associated measurement techniques, the purpose of the present study was to determine the reliability and agreement, as well as the standard error of measurement (SEM), of paraspinal muscle CSA and composition measurements obtained using 2 open source, readily available computer software programs: ImageJ and OsiriX. In addition, the associated image analysis protocol is proposed for standardized use to facilitate comparisons among studies.

Materials and Method

Measurement Study Design

Total CSA and FCSA measurements of the multifidus muscle, the erector spinae muscle, and the 2 muscles combined, bilaterally, were directly obtained for each participant using 2 open source software packages. ImageJ (version 1.43, National Institutes of Health, Bethesda, Maryland) is a free, downloadable, public domain image processing software program (http://rsbweb.nih.gov/ij/download.html) that was developed by the National Institutes of Health. The 32-bit OsiriX software (version 3.8.1, Pixmeo, Geneva, Switzerland) was downloaded from http://www.osirix-viewer.com/ and was previously assessed as a more user-friendly image analysis software package for the Apple Mac OS (Microsoft Corp, Redmond, Washington) than ImageJ.[24] One of the OsiriX program's main advantages is its integrated PAC system, which allows patient data to be stored automatically.[24] Both software packages have been utilized by clinicians and scientists in a wide variety of studies as functional tools for image analysis.[24–26]

To determine intrarater and intersoftware measurement reliability, each muscle measurement was acquired 4 times by the same rater, twice using each software program. In an effort to minimize bias from carryover or practice effects, the first complete set of measurements using each software program was obtained by alternating between programs after every block of 10 participants' images, randomly selected and ordered. After all magnetic resonance images were assessed once using either ImageJ or OsiriX, the images were reordered and blinded to be similarly assessed again a minimum of 5 days after the first measurements were completed.

Sample of Lumbar MRI

A sample of 30 patients (11 female and 19 male) were randomly selected from an ongoing study of patients attending spine specialty clinics and having commonly diagnosed lumbar pathologies, including disk herniation, spinal stenosis, spondylolisthesis, and nonspecific chronic LBP. Patients were excluded if they were below 18 or over 60 years of age, had a contrast agent allergy, had reduced renal function, were not able to undergo MRI acquisition, or had a tumor, infection, spinal fracture, or rheumatoid arthritis or were pregnant.

The MRI protocol included routine T2-weighted turbo spin echo sequences for both axial and sagittal images acquired with a Siemens Avanto 1.5T MRI system (Siemens AG, Erlangen, Germany) (axial T2 parameters included repetition time=4,000, echo time=113, and slice thickness=3 mm).

Muscle Measurements

All muscle measurements were acquired by one of the investigators (M.F.) who, in preparation for the measurements, received training in spine MRI assessments focusing on lumbar intervertebral disk and paraspinal muscle morphology. For practice purposes, a sample of about 15 images was analyzed with each software application prior to the beginning of the measurement study.

Quantitative measurements of the multifidus and erector spinae muscles individually and as a group (multifidus and erector spinae muscles together) were obtained from the T2-weighted axial images using ImageJ and OsiriX. ImageJ has already been used in previous studies to measure total CSA and FCSA using a threshold method, with previously reported intraclass correlation coefficients for intrarater reliability of both area measurements ranging from .89 to .99.[15,16] We are not aware of any reports of reliability of paraspinal muscle morphology measurements using OsiriX. The same MRI slices were used for the ImageJ and OsiriX muscle measurements. Because the reliability of FCSA and total CSA measurements has been shown to be relatively equivalent across spinal levels,[16] measurements for this study were taken only at mid-disk for L4–L5 and mid-S1 for every participant. The 2 levels were selected because most lumbar pathologies and muscle morphological changes occur between L4–L5 and L5–S1.[27]

The paraspinal muscle measurements of interest in this study for the multifidus and erector spinae muscles and the 2 muscles as a group included the following: total CSA, FCSA, ratio of FCSA to total CSA, side-to-side differences (muscle asymmetry) in total CSA and FCSA, and mean signal intensity of total CSA.

The FCSA measurement was obtained by selecting a threshold signal within the total muscle CSA to include only pixels within the lean muscle tissue range (Fig. 1A). The gray scale range for lean muscle tissue was established for every participant, on each scan slice. Four to 6 sample regions of interest (ROI) within the bilateral paraspinal muscle group (multifidus and erector spinae) were taken from areas of lean muscle tissue visible on each slice (Fig. 1B). If atrophied paraspinal muscle with significant fatty infiltration was encountered, care was taken to avoid the inclusion of any visible pixel of fat. The maximum value acquired from the sample ROIs was used as the highest threshold to distinguish muscle tissue from fat, in the same way the lower limit was determined by the minimum signal intensity value obtained from the sample ROIs. However, because we observed that the lower limit was typically 0 or 1, it might be best to standardize the lower limit at 0. This standardization could potentially decrease related measurement error and simplify the protocol. When timing a sample of measurements obtained with each software program, the average time taken to complete the measurements of the 3 muscle regions bilaterally at one spinal level was approximately 9 minutes with OsiriX and 5 minutes with ImageJ.

Data Analysis

The statistical analysis was performed using Statistical Package for the Social Sciences version 18.0 (SPSS Inc, Chicago, Illinois). Means and standard deviations for each variable were obtained. The ICC (2,1) was calculated to determine the intrarater reliability of measurements using OsiriX and ImageJ for each measurement variable and every muscle of interest using a 2-way random-effects model and absolute agreement. The ICC reflects both the degree of correlation and agreement between the ratings and was interpreted using the following criteria, as suggested by Portney and Watkins[28]: .00–.49=poor, .50–.74=

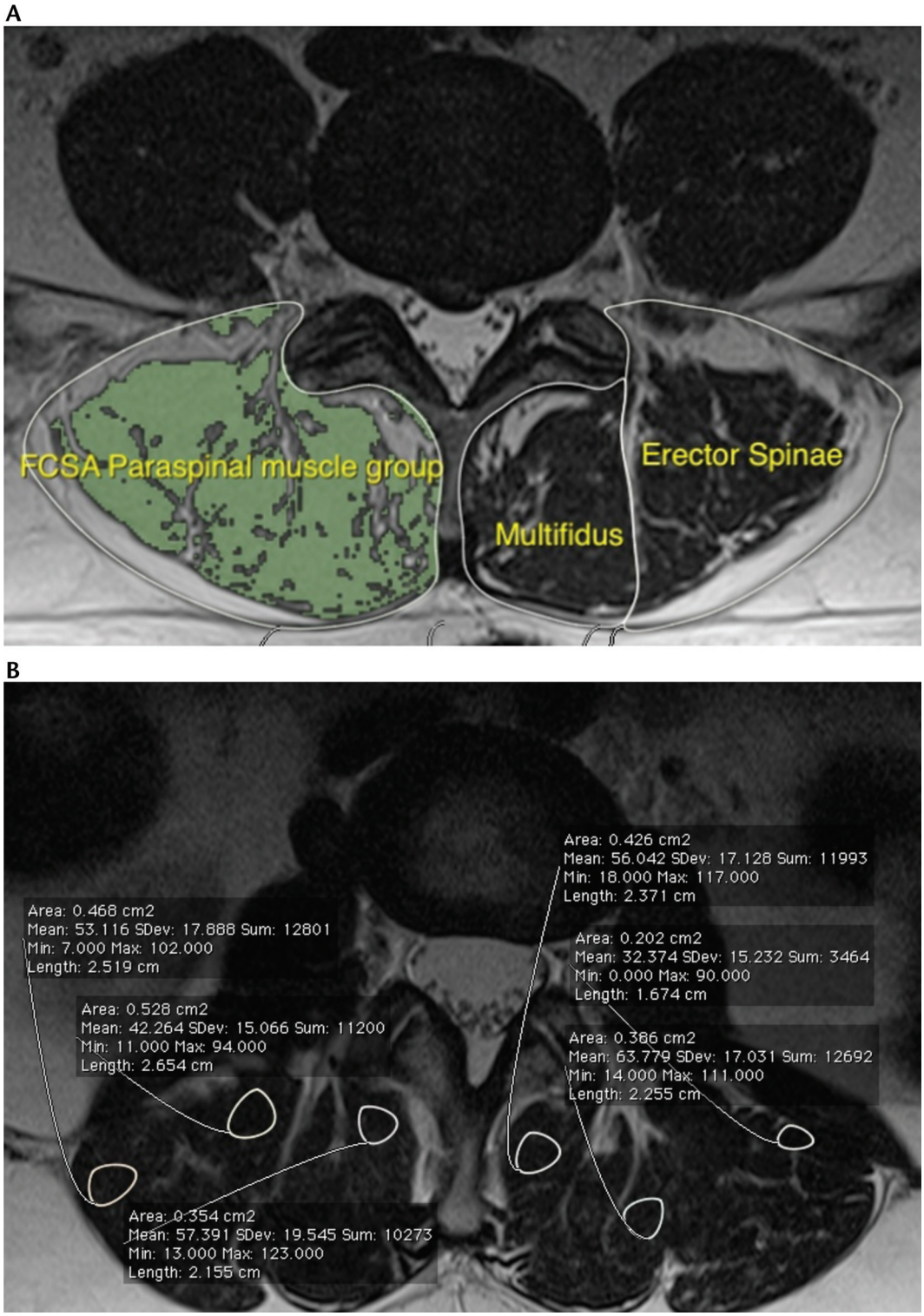

Figure 1.
(A) Measurement of total cross-sectional area of erector spinae and multifidus muscles (right) at L4–L5. Lean muscle functional cross-sectional area (FCSA) of the paraspinal muscle group using a threshold method is represented by the area highlighted in green (left). (B) Sample selection of regions of interest to define upper and lower signal intensity threshold limits.

moderate, and .75–1.00=excellent. The SEM was calculated to provide an estimate of the expected error related to a particular measurement.[28] The ICC defines the ability to discriminate among individuals, whereas the SEM defines the measurement error in the same units as the initial measurement.[29] Method agreement between the measurements acquired from the different software programs also was evaluated using the 95% limits of agreement as suggested by Bland and Altman.[30–32] Reliability results were analyzed and reported according to spinal level, muscle investigated, and muscle side.

Results

Inter-Software Reliability of Muscle Measurements Using OsiriX and ImageJ

The results for the inter-software reliability (ICC), SEM values, and descriptive statistics (mean±SD) for the left side are presented in Table 1 for the L4–L5 spinal measurements and in Table 2 for the S1 measurements. The results for the right side were virtually equivalent and are not presented. The inter-software reliability was analyzed by comparing the first set of measurements collected with each software program. The ICCs for all of the different muscle composition measurements, regardless of the muscle analyzed or spinal level, showed excellent agreement and varied between .81 and .99. However, the SEM associated with the side-to-side difference measurements was of greater magnitude in comparison with the rest of the other muscle measurements.

Table 1.
Inter-Software Reliability Indexes for Left Paraspinal Muscle Measurements at L4–L5[a]

Parameter	$\bar{X}$ (SD)	ICC (95% CI)	SEM
Multifidus muscle			
CSA (cm^2)	10.07 (1.47)	.96 (.92–.98)	0.29
SI	188.02 (40.89)	.99 (.99–1.00)	4.09
FCSA (cm^2)	5.92 (1.73)	.96 (.92–.98)	0.35
FCSA/CSA	0.58 (0.12)	.95 (.91–.98)	0.03
CSA diff (cm^2)	1.03 (0.77)	.81 (.63–.90)	0.33
FCSA diff (cm^2)	0.72 (0.58)	.87 (.75–.94)	0.21
Erector spinae muscle			
CSA (cm^2)	18.49 (3.95)	.99 (.98–1.00)	0.39
SI	226.07 (47.96)	.99 (.96–1.00)	4.80
FCSA (cm^2)	9.71 (3.37)	.97 (.95–.99)	0.58
FCSA/CSA	0.52 (0.13)	.94 (.88–.97)	0.03
CSA diff (cm^2)	1.31 (1.35)	.86 (.68–.94)	0.50
FCSA diff (cm^2)	1.22 (1.12)	.98 (.96–.99)	0.16
Paraspinal muscle group			
CSA (cm^2)	28.49 (4.52)	.99 (.99–1.00)	0.45
SI	212.28 (43.21)	.99 (.99–1.00)	4.32
FCSA (cm^2)	15.63 (4.47)	.97 (.94–.99)	0.77
FCSA/CSA	0.55 (0.12)	.95 (.91–.98)	0.03
CSA diff (cm^2)	1.27 (1.18)	.87 (.75–.94)	0.43
FCSA diff (cm^2)	1.23 (1.15)	.96 (.91–.99)	0.23

[a] ICC=intraclass correlation coefficient, CI=confidence interval, SEM=standard error of measurement, CSA=cross-sectional area, SI=signal intensity, FCSA=functional CSA, FCSA/CSA=ratio, CSA diff=side-to-side difference in CSA, FCSA diff=side-to-side difference in functional CSA.

Inter-Software Agreement

Figure 2 shows the combined Bland and Altman 95% limits of agreement plots for the different muscle composition measurements from the left multifidus muscle at L4–L5 using the first set of measurements collected with each software program. Two methods are considered to have good agreement when the measurement difference is small enough for both methods to be used interchangeably.[30] All of the plots show good agreement between OsiriX and ImageJ and no systematic bias; the distribution of the scores around the mean approximate zero and are spread evenly and randomly above and below the line.[28] As suggested by Bland and Altman, an initial histogram of the difference scores was performed for every measurement parameter, and all histograms followed a normal distribution. Because the error is normally distributed, we can observe that about 95% of the points are between the limits of agreement (noted by the dashed lines on the plots) for each measure. The width of the limits of agreement for the different measurements also was small (Fig. 2).

Intrarater Reliability of Muscle Measurements Using OsiriX and ImageJ

The intrarater reliability (ICC), SEM values, and descriptive statistics (mean±SD) related to OsiriX and ImageJ muscle measurements for the left side are presented in Table 3 for the L4–L5 level and in Table 4 for the S1 level. Again, the results for the right side were virtually equivalent and are not presented. The ICCs for intrarater reliability across both spinal levels for total CSA measurements of the paraspinal muscles, individually and as a group, ranged from .94 to .99 for ImageJ and from

Table 2.
Inter-Software Reliability Indexes for Left Paraspinal Muscle Measurements at S1[a]

Parameter	$\bar{X}$ (SD)	ICC (95% CI)	SEM
Multifidus muscle			
CSA (cm^2)	12.33 (1.74)	.97 (.93–.99)	0.30
SI	233.13 (49.64)	.99 (.99–1.00)	4.96
FCSA (cm^2)	6.91 (2.11)	.96 (.93–.98)	0.42
FCSA/CSA	0.56 (0.13)	.94 (.89–.97)	0.03
CSA diff (cm^2)	1.00 (0.81)	.88 (.77–.94)	0.28
FCSA diff (cm^2)	0.97 (1.03)	.97 (.94–.99)	0.18
Erector spinae muscle			
CSA (cm^2)	8.10 (4.10)	.99 (.98–1.00)	0.41
SI	304.52 (63.98)	.99 (.97–.99)	6.40
FCSA (cm^2)	2.59 (1.85)	.96 (.93–.98)	0.37
FCSA/CSA	0.31 (0.14)	.93 (.86–.97)	0.04
CSA diff (cm^2)	1.45 (1.24)	.87 (.75–.94)	0.45
FCSA diff (cm^2)	0.71 (0.65)	.86 (.73–.93)	0.24
Paraspinal muscle group			
CSA (cm^2)	20.34 (4.72)	.99 (.99–1.00)	0.47
SI	259.12 (51.19)	.99 (.99–1.00)	5.12
FCSA (cm^2)	9.47 (2.98)	.96 (.92–.98)	0.60
FCSA/CSA	0.47 (0.12)	.92 (.85–.96)	0.03
CSA diff (cm^2)	1.62 (1.19)	.89 (.79–.95)	0.40
FCSA diff (cm^2)	1.45 (1.16)	.96 (.93–.99)	0.23

[a] ICC=intraclass correlation coefficient, CI=confidence interval, SEM=standard error of measurement, CSA=cross-sectional area, SI=signal intensity, FCSA=functional CSA, FCSA/CSA=ratio, CSA diff=side-to-side difference in CSA, FCSA diff=side-to-side difference in functional CSA.

.97 to .99 for OsiriX. The FCSA ICCs across both spinal levels for all of the measured muscles tended to be slightly lower for ImageJ (ICC=.90–.96) compared with OsiriX (ICC=.97–.98), although all values were excellent.

The side-to-side difference measurements are of much smaller areas compared with the total CSA and FCSA measurements and had lower reliability values (ICC=.77–.97). The intrarater ICCs for the side-to-side difference in total CSA varied from .80 to .90 for OsiriX and from .78 to .91 for ImageJ, and the side-to-side difference in FCSA varied from .77 to .96 for OsiriX and from .85 to .97 for ImageJ. The reliability of the signal intensity of the total CSA and the ratio of FCSA/CSA also was measured because these data give a proportion estimate of a muscle fat content. The mean ICC for the signal intensity of the total CSA was .99 for measurements acquired with either software program, and the mean for the FCSA/CSA ratio was .96 for OsiriX and .91 for ImageJ (range=.88–.97). The SEM associated with each muscle composition measurement was generally comparable between the software programs, except for the FCSA measurement where the SEM tended to be higher for ImageJ.

Discussion

We have presented specific protocols for paraspinal muscle measurements using 2 readily available, free image analysis programs, OsiriX and ImageJ, in a level of detail to allow replication (Appendix). The reliability and agreement of related paraspinal muscle measurements were found to be reasonably comparable between software programs, with excellent reliability when applied to a clinically relevant population. These findings are supported by the Bland and Altman limits of agreement that indicate inter-software agreement is within an acceptable range to use either of the 2 methods. Furthermore, the similar intrarater and inter-software reliability coefficients and SEMs suggest that the software used contributes little to the measurement error.

A threshold technique was utilized to calculate FCSA based on differences in pixel intensities between muscle (low intensity) and fat tissues (high intensity) on T2-weighted axial images. The application used in OsiriX is based on a region-growing algorithm, whereas ImageJ uses a signal intensity threshold algorithm. With OsiriX, once the lean muscle signal intensity is defined, the region-growing image segmentation involves the selection of seed points, which determine whether neighboring pixels will be included in the selection. This method is more time-consuming compared with a straight threshold algorithm where the only step needed is to indicate the upper and lower bounds of the threshold limit for muscle tissue. However, as suggested by Dello et al,[24] our impression was that OsiriX is a more user-friendly software package than ImageJ. We are not aware of any other study that investigated the agreement of paraspinal muscle measurements between 2 different image analysis programs.

The results of this study related to intrarater reliability, however, are similar to those of other studies examining measurements of FCSA and total CSA that used a threshold

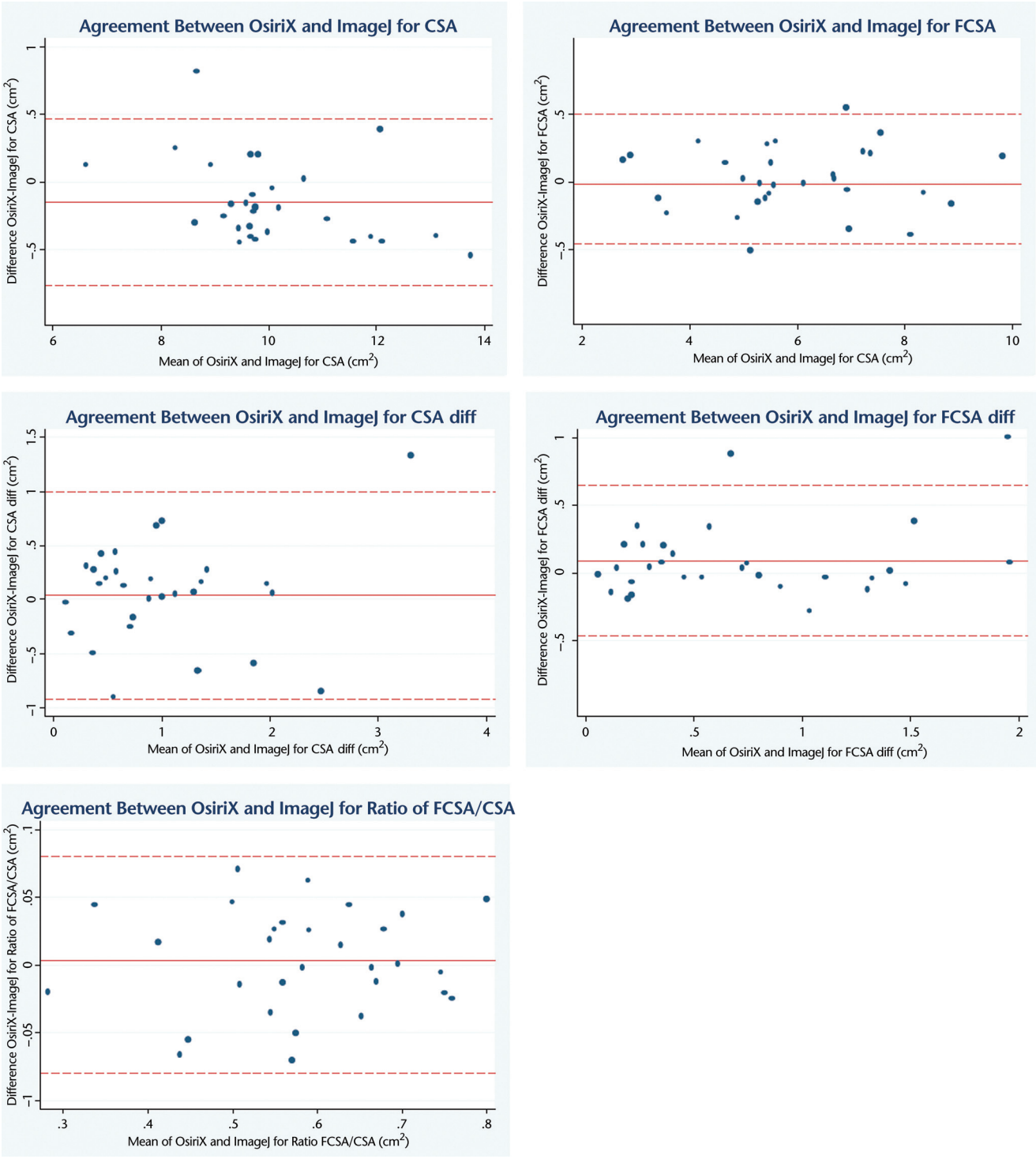

Figure 2.
Bland-Altman 95% limits of agreement plots for the different muscle composition measurements of the left multifidus muscle at L4–L5. CSA=cross-sectional area, FCSA=functional CSA, CSA diff=side-to-side difference in CSA, FCSA diff=side-to-side difference in functional CSA, FCSA/CSA=ratio.

Table 3.
Intrarater Reliability Indexes for OsiriX and ImageJ for Left Paraspinal Muscle Measurements at L4–L5[a]

Parameter	OsiriX			ImageJ		
	$\bar{X}$ (SD)	ICC (95% CI)	SEM	$\bar{X}$ (SD)	ICC (95% CI)	SEM
Multifidus muscle						
CSA (cm^2)	10.03 (1.47)	.97 (.93–.98)	0.26	10.14 (1.49)	.98 (.96–.99)	0.21
SI	188.49 (40.32)	.99 (.99–1.00)	4.03	187.30 (40.63)	.99 (.99–1.00)	4.06
FCSA (cm^2)	5.84 (1.71)	.97 (.93–.98)	0.30	5.81 (1.73)	.96 (.88–.99)	0.35
FCSA/CSA	0.58 (0.13)	.97 (.92–.99)	0.02	0.57 (0.12)	.93 (.70–.98)	0.03
CSA diff (cm^2)	1.01 (0.77)	.80 (.62–.90)	0.34	1.03 (0.74)	.87 (.75–.94)	0.27
FCSA diff (cm^2)	0.75 (0.59)	.90 (.78–.95)	0.19	0.66 (0.52)	.93 (.85–.96)	0.14
Erector spinae muscle						
CSA (cm^2)	18.45 (3.95)	.99 (.99–1.00)	0.39	18.45 (3.96)	.99 (.98–1.00)	0.40
SI	227.45 (47.69)	.99 (.99–1.00)	4.77	224.50 (48.42)	.99 (.99–1.00)	4.84
FCSA (cm^2)	9.48 (3.50)	.98 (.94–.99)	0.50	9.43 (3.19)	.96 (.71–.99)	0.64
FCSA/CSA	0.51 (0.13)	.97 (.88–.99)	0.02	0.51 (0.13)	.92 (.67–.97)	0.04
CSA diff (cm^2)	1.12 (1.16)	.86 (.72–.93)	0.42	1.34 (1.26)	.86 (.71–.94)	0.47
FCSA diff (cm^2)	1.17 (1.12)	.96 (.92–.98)	0.22	1.18 (1.09)	.97 (.92–.99)	0.19
Paraspinal muscle group						
CSA (cm^2)	28.42 (4.57)	.99 (.99–1.00)	0.46	28.60 (4.60)	.99 (.99–1.00)	0.46
SI	214.31 (43.34)	.99 (.99–1.00)	4.34	211.42 (43.00)	.99 (.99–1.00)	4.30
FCSA (cm^2)	15.30 (4.60)	.98 (.92–.99)	0.65	15.25 (4.35)	.95 (.76–.98)	0.97
FCSA/CSA	0.53 (0.12)	.96 (.83–.98)	0.02	0.53 (0.11)	.92 (.61–.97)	0.03
CSA diff (cm^2)	1.26 (1.14)	.87 (.74–.93)	0.41	1.27 (1.16)	.87 (.74–.93)	0.42
FCSA diff (cm^2)	1.20 (1.15)	.96 (.92–.98)	0.23	1.20 (1.16)	.97 (.94–.99)	0.20

[a] ICC=intraclass correlation coefficient, CI=confidence interval, SEM=standard error of measurement, CSA=cross-sectional area, SI=signal intensity, FCSA=functional CSA, FCSA/CSA=ratio, CSA diff=side-to-side difference in CSA, FCSA diff=side-to-side difference in functional CSA.

technique. Danneels et al[1] reported ICCs for intrarater reliability that varied between .81 and .92 for FCSA, whereas other authors reported ICCs for intrarater reliability that were slightly higher (.90–.99).[15,16,18] Studies using a tracing technique to measure FCSA by manually segmenting muscle from fat tissues have shown somewhat lower ICCs for intrarater reliability, varying between .81 and .96.[17,19] Other investigators measuring total CSA reported ICCs for intrarater reliability that varied between .89 and .99.[3,15,22,33,34] In the present study, however, intrarater reliability indexes were computed primarily in order to better interpret the contribution of inter-software reliability to measurement error. The fact that inter-software reliability is similarly high as intrarater reliability further suggests that using one software program as opposed to the other contributes little to measurement error.

One of the strengths of this study is the report of reliability indexes related to both individual muscle measurements and side-to-side differences. After several investigations of individuals with chronic LBP and those who were asymptomatic, Hides et al[4] suggested that total CSA side-to-side asymmetry of the multifidus muscle greater than 10% could potentially signify an abnormality. Other investigators are now referring to this guideline.[15]

However, to our knowledge, only 2 studies examined the reliability of side-to-side difference measurements, with ICCs varying between .77 and .97 for side-to-side difference measurements of total CSA and .82 to .94 for FCSA (Battié and colleagues, unpublished research).[15] The ICCs for both side-to-side difference measurements reported in our study are similar. Despite both single muscle measurements and side-to-side difference measurements having high reliability coefficients and similar SEMs, the error is relatively more important in the difference measurements, as they represent much smaller areas. For example, when using OsiriX, we found that the mean FCSA side-to-side difference of the multifidus muscle at L4–L5 was 0.75 cm^2 and the associated SEM was 0.19 cm^2, which is

Table 4.
Intrarater Reliability Indexes for OsiriX and ImageJ for Left Paraspinal Muscle Measurements at S1[a]

Parameter	OsiriX			ImageJ		
	$\bar{X}$ (SD)	ICC (95% CI)	SEM	$\bar{X}$ (SD)	ICC (95% CI)	SEM
Multifidus muscle						
CSA (cm^2)	12.25 (1.67)	.98 (.97–.99)	0.24	12.42 (1.75)	.99 (.97–.99)	0.18
SI	234.09 (50.66)	.99 (.99–1.00)	5.07	232.61 (48.43)	.99 (.99–1.00)	4.84
FCSA (cm^2)	6.86 (2.18)	.98 (.97–.99)	0.31	6.84 (2.05)	.94 (.88–.97)	0.50
FCSA/CSA	0.55 (0.14)	.97 (.94–.99)	0.02	0.55 (0.12)	.92 (.83–.96)	0.03
CSA diff (cm^2)	0.99 (0.78)	.88 (.76–.94)	0.27	1.05 (0.74)	.91 (.81–.95)	0.22
FCSA diff (cm^2)	0.95 (1.03)	.94 (.88–.97)	0.25	1.02 (1.04)	.95 (.91–.98)	0.23
Erector spinae muscle						
CSA (cm^2)	8.04 (4.19)	.99 (.99–1.00)	0.42	8.20 (4.12)	.99 (.98–1.00)	0.41
SI	305.10 (59.97)	.99 (.98–1.00)	6.00	305.00 (5.36)	.99 (.98–.99)	6.54
FCSA (cm^2)	2.54 (1.89)	.98 (.96–.99)	0.27	2.43 (1.60)	.92 (.80–.96)	0.45
FCSA/CSA	0.30 (0.14)	.95 (.90–.98)	0.03	0.29 (0.13)	.89 (.75–.95)	0.04
CSA diff (cm^2)	1.40 (1.24)	.80 (.62–.90)	0.55	1.46 (1.27)	.86 (.73–.93)	0.47
FCSA diff (cm^2)	0.66 (0.62)	.77 (.57–.88)	0.30	0.66 (0.58)	.85 (.72–.93)	0.22
Paraspinal muscle group						
CSA (cm^2)	20.33 (4.71)	.99 (.99–1.00)	0.47	20.43 (4.82)	.94 (.97–.99)	0.68
SI	260.20 (50.48)	.99 (.99–1.00)	5.05	258.30 (50.33)	.99 (.98–1.00)	5.03
FCSA (cm^2)	9.43 (3.12)	.98 (.96–.99)	0.44	9.25 (2.74)	.90 (.77–.95)	0.88
FCSA/CSA	0.47 (0.12)	.96 (.92–.98)	0.02	0.46 (0.11)	.88 (.75–.94)	0.04
CSA diff (cm^2)	1.55 (1.20)	.90 (.80–.95)	0.38	1.59 (1.22)	.78 (.58–.89)	0.56
FCSA diff (cm^2)	1.40 (1.16)	.96 (.91–.98)	0.23	1.43 (1.17)	.97 (.95–.98)	0.20

[a] ICC=intraclass correlation coefficient, CI=confidence interval, SEM=standard error of measurement, CSA=cross-sectional area, SI=signal intensity, FCSA=functional CSA, FCSA/CSA=ratio, CSA diff=side-to-side difference in CSA, FCSA diff=side-to-side difference in functional CSA.

small in absolute terms but still relatively large, as it represents approximately 25% of the mean measurement of multifidus asymmetry. The SEM of 0.30 cm^2 represents only approximately 5% of the mean multifidus muscle FCSA measurement of 5.84 cm^2. When changes over time are of interest, such as in preintervention and postintervention measurements, there may be a high probability that the differences observed are due to measurement error rather than true changes if they do not exceed 2 SEMs.[35] The greater measurement error related to side-to-side difference was confirmed by the Bland and Altman plots where the limits of agreement were relatively large in comparison with the other measurements.

Another strength of this study is that we studied patients with LBP conditions for whom the measurements are most likely to be of interest and who are expected to have more fatty infiltration[9,36] and muscle atrophy[1,4] compared with people who are healthy, increasing the difficulty of determining muscle boundaries during manual segmentation. Other authors reporting on the reliability of FCSA measurements primarily used samples of participants who were healthy.[15,16,18] Our results suggest that total muscle size, within the range studied, and spinal level (L4–L5, S1) do not influence intrarater reliability or inter-software agreement. Only the erector spinae muscle at S1 seems to have a proportionally higher SEM associated with the composition measurements with both software programs, in comparison with the other analyzed muscles. This finding could be explained by the high fatty infiltration and the smaller size of the erector spinae muscle at S1, which increased the difficulty in determining the muscle borders.

A limitation of this study is the restriction of the measurement analysis to only 2 software packages. Even though inter-software reliability and agreement between OsiriX and ImageJ were excellent, even when measurements were obtained by an individual with modest experience, this finding might not be the case for other custom-made and commercial software used for image

analysis. As determining inter-software reliability was the primary purpose of this study, replicate measurements were obtained from the same image to remove a potential extraneous source of measurement error. However, this represents a limitation when looking at intrarater reliability, where estimates might have been somewhat lower if the rater had repeated the entire procedure, including selecting the image from which to obtain the measurement.

In summary, a detailed protocol for paraspinal muscle CSA and composition measurements using 2 widely available, commonly used software programs was described, which yielded measurements with high inter-software and intrarater reliability. However, we found slightly lower reliability of side-to-side difference measurements compared with measurements of single muscles, which may be an important consideration in view of the current interest in muscle asymmetry. Future related studies would benefit from using a standard muscle measurement protocol to facilitate replication and comparisons among studies.

Both authors provided concept/idea/research design, writing, and data analysis. Ms Fortin provided data collection and project management. Dr Battié provided fund procurement and facilities/equipment. The authors thank Doug Gross and Luciana Macedo for their review of this work and helpful comments.

This study was approved by the Health Research Ethics Board of the University of Alberta.

Support was received from the Canada Research Chairs Program and the European Union Community's Seventh Framework Programme (FP7, 2007–2013; grant HEALTH F2–2008-201626; project GENODISC).

DOI: 10.2522/ptj.20110380

References

1 Danneels LA, Vanderstraeten GG, Cambier DC, et al. CT imaging of trunk muscles in chronic low back pain patients and healthy controls subjects. *Eur Spine J.* 2000;9:266–272.

2 Parkkola R, Rytökoski U, Kormano M. Magnetic resonance imaging of the discs and trunk muscles in patients with chronic low back pain and healthy controls subjects. *Spine (Phila Pa 1976).* 1993;18:830–836.

3 Barker KL, Shamley DR, Jackson D. Changes in the cross-sectional area of multifidus and psoas in patients with unilateral back pain: the relationship to pain and disability. *Spine (Phila Pa 1976).* 2004;29:E515–E519.

4 Hides JA, Gilmore C, Stanton W, et al. Multifidus size and symmetry among chronic LBP and healthy asymptomatic subjects. *Man Ther.* 2008;13:43–49.

5 Hodges P, Holm AK, Hansson T, et al. Rapid atrophy of the lumbar multifidus follows experimental disc or nerve root injury. *Spine (Phila Pa 1976).* 2006;31:2926–2933.

6 Ploumis A, Michailidis N, Christodoulou P, et al. Ipsilateral atrophy of paraspinal and psoas muscle in unilateral back pain patients with monosegmental degenerative disc disease. *Br J Radiol.* 2011;84:709–713.

7 Hides JA, Stokes MJ, Saide M, et al. 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.

8 Mengiardi B, Schmid MR, Boos N, et al. Fat content of lumbar paraspinal muscles in patients with chronic low back pain and in asymptomatic volunteers: quantification with MR spectroscopy. *Radiology.* 2006;240:786–792.

9 Kjaer P, Bendix T, Sorensen JS, et al. Are MRI-defined fat infiltrations in the multifidus muscles associated with low back pain? *BMC Med.* 2007;5:2.

10 Hyun JK, Lee JY, Lee SJ, et al. Asymmetric atrophy of multifidus muscle in patients with unilateral lumbosacral radiculopathy. *Spine (Phila Pa 1976).* 2007;32:E598–E602.

11 Kader DF, Wardlaw D, Smith FW. Correlation between MRI changes in the lumbar multifidus muscles and leg pain. *Clin Radiol.* 2000;55:145–149.

12 Stokes MJ, Cooper RG, Morris G. et al. Selective changes in multifidus dimensions in patients with chronic low back pain. *Eur Spine J.* 1992;1:38–42.

13 Cooper RG, St Clair Forbers W, Jayson MI. Radiographic demonstration of paraspinal muscle wasting in patients with chronic low back pain. *Br J Rheumatol.* 1992;31:389–394.

14 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.

15 Niemeläinen R, Briand MM, Battié MC. Substantial asymmetry in paraspinal muscle cross-sectional areas in healthy adults questions its value as a marker of LBP and pathology. *Spine (Phila Pa 1976).* 2011;36:2152–2157.

16 Ranson CA, Burnett AF, Kerslake R, et al. An investigation into the use of MR imaging to determine the functional cross sectional area of lumbar paraspinal muscles. *Eur Spine J.* 2006;15:764–773.

17 Fan S, Hu Z, Zhao F, et al. Multifidus muscle changes and clinical effects of one-level posterior lumbar interbody fusion: minimally invasive procedure versus conventional open approach. *Eur Spine J.* 2009;19:316–324.

18 Gille O, Jolivet E, Dousset V, et al. Erector spinae muscle changes on magnetic resonance imaging following lumbar surgery through a posterior approach. *Spine (Phila Pa 1976).* 2007;32:1236–1241.

19 Hu JZ, He J, Zhao FD, et al. An assessment of intra- and inter-reliability of the lumbar paraspinal muscle parameters using CT scan and magnetic resonance imaging. *Spine (Phila Pa 1976).* 2011;36:E868–E874.

20 Marras WS, Jorgensen MJ, Granata KP, et al. Female and male trunk geometry: size and prediction of the spine loading trunk muscles derived from MRI. *Clin Biomech (Bristol, Avon).* 2001;16:38–46.

21 Kang CH, Shin MJ, Kim SM, et al. MRI of paraspinal muscles in lumbar degenerative kyphosis patients and control patients with chronic low back pain. *Clin Radiol.* 2007;62:479–486.

22 Hides JA, Belavy DL, Stanton W, et al. Magnetic resonance imaging assessment of trunk muscles during prolonged bed rest. *Spine (Phila Pa 1976).* 2007;32:1687–1692.

23 Keller A, Gunderson R, Reikeras O, et al. Reliability of computed tomography measurements of paraspinal muscle cross-sectional area and density in patients with chronic low back pain. *Spine (Phila Pa 1976).* 2003;28:1455–1460.

24 Dello SA, Stoot JH, van Stipout RS, et al. Prospective volumetric assessment of the liver on a personal computer by nonradiologists prior to partial hepatectomy. *World J Surg.* 2011;35:386–392.

25 Yamauchi T, Yamazaki M, Okawa A, et al. Efficacy and reliability of highly functional open source DICOM software (OsiriX) in spine surgery. *J Clin Neurosci*. 2010;17:756–759.

26 Albert S, Cristofari JP, Cox A, et al. Reconstruction mandibulaire par lambeau microanstomosé de fibula: modélisation radiologique préopératoire par le logiciel OsiriX. *Ann Chir Plas Esthe.* 2011;56:494–503.

27 Takatalo J, Karppinen J, Niinimäki J, et al. Prevalence of degenerative imaging findings in lumbar magnetic resonance imaging among young adults. *Spine (Phila Pa 1976).* 2009;34:1716–1721.

28 Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall Inc; 2000.

29 Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther.* 1997;77:745–750.

30 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8:135–160.

31 Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol.* 2003;22:85–93.

32 Bland JM, Altman DG. Statistical method for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307–310.

33 Ropponen A, Videman T, Battié MC. The reliability of paraspinal muscles composition measurements using routine spine MRI and their association with back function. *Man Ther.* 2008;13:349–356.

34 Kim DY, Lee SH, Chung SK, et al. Comparison of multifidus muscle atrophy and trunk extension muscle strength: percutaneous versus open pedicle screw fixation. *Spine (Phila Pa 1976).* 2005;30:123–129.

35 Harvill LM. Standard error of measurement, module 9, summer 1991. NCME website. Available at: http://ncme.org/linkservid/6606715E-1320-5CAE-6E9DDC581EE47F88/showMeta/0/. Accessed September 1, 2011.

36 Lee JC, Cha JG, Kim Y, et al. Quantitative analysis of back muscle degeneration in the patients with the degenerative lumbar flat back using a digital image analysis: comparison with the normal controls. *Spine (Phila Pa 1976).* 2008;33:318–325.

Appendix.

Specific Protocols for Obtaining Muscle Cross-Sectional Area (CSA) and Functional CSA (FCSA) Signal Measurement

Muscle total CSA measurement protocol for both ImageJ and OsiriX

1. Begin defining each region of interest (ROI) at the inferior-medial corner of the muscle.
2. Include fat between multifidus muscle and lamina within the multifidus muscle ROI.
3. Include fat between erector spinae and multifidus muscles within the erector spinae muscle ROI.
4. Fat within the erector spinae muscle fascial boundary, lateral and posterior to iliocostalis lumborum, is included in the erector spinae muscle ROI for total CSA.
5. Fat within the erector spinae muscle fascial boundary posterior to the longissimus muscle component is included in the erector spinae muscle ROI for total CSA.
6. Isolated deposits of intramuscular fat are included in the total CSA ROI for the muscle.
7. When a clear boundary between fat and muscle is not evident (ie, when a region of gray pixels is encountered), the ROI is defined though the middle of this region and in a manner that allows a reasonable approximation of the muscle's anticipated boundary.

Defining the signal intensity range to measure muscle FCSA using OsiriX

1. Use the close polygon ROI tool (mouse button function) to select 4 to 6 ROIs of homogenous lean muscle tissue (excluding fat pixels) evenly and bilaterally (refer to Fig. 1B) within the paraspinal muscles (erector spinae and multifidus).
2. From the sample ROIs, use the lowest minimum value as the lower threshold bound and the highest maximum value as the upper bound.
3. Use the close polygon ROI tool to trace the contour of the muscle of interest.
4. Double click on the muscle ROI results box and name the ROI (eg, right multifidus muscle). Close the ROI information window.
5. Make sure that muscle ROI (eg, right multifidus muscle) is selected (results box should be highlighted in red). Open "ROI" pull-down menu in the main menu bar and choose *Set pixel values to . . .* Select the option *outside ROIs* and select the option *to this new value*. Change the new value to a negative number and click on "OK." This step will "delete" the image background to apply the region-growing threshold only to the specific selected muscle ROI.

(*Continued*)

Appendix.
Continued

6. Open ROI pull-down menu in the main menu bar and choose *Grow region (2D/3D segmentation).* In the parameters section of the window, select the algorithm *threshold (lower/upper bounds)*. Make sure that *brush ROI* option is selected in the results section of the window and leave the window open. No other parameters/options need to be changed.

7. In the appropriate space of the parameters window section, enter the upper and lower threshold values previously defined in step 2 and leave the window open.

8. Click inside the paraspinal muscle ROI in a homogenous lean muscle tissue area.

9. To calculate the new FCSA ROI, click on *compute* button of the segmentation parameters window.

10. If needed, repeat steps 8 and 9 until lean muscle tissue of the entire muscle ROI is highlighted.

11. To combine all the brush ROIs together, open "ROI" pull-down menu in the main menu bar and choose *Brush ROIs*, then select *Merge selected brush ROIs.* Close the segmentation parameters window.

12. When completed, close the image slice and reopen from the main patients local database. The same image slice will appear with the initial image background and the newly created region-growing ROI representing the muscle FCSA.

13. Repeat steps 3 to 12 to measure the FCSA of another muscle. Give a different name to every muscle ROI (step 4).

Defining signal the intensity range to measure muscle FCSA using ImageJ

1. Use the polygon selections ROI tool from the main menu bar to select 4 to 6 ROIs of sample homogenous lean muscle tissue (excluding fat pixels) evenly and bilaterally (refer to Fig. 1B) within the paraspinal muscles (erector spinae and multifidus). To obtain each ROI area, mean signal intensity, and minimum/maximum values open "Analyze" pull-down menu and select *Measure* (or click control + M).

2. From sample ROIs, use the lowest minimum value as the lower threshold bound and the highest maximum value as the upper bound.

3. Use the close polygon selections ROI tool from the main menu bar to trace the contour of the muscle of interest. To obtain muscle ROI area, mean signal intensity, and minimum/maximum values open "Analyze" pull-down menu and select *Measure* (or click control + M).

4. Open "Image" pull-down menu and select *Adjust*, then click on *Threshold.* Click on the *Set* button from the threshold window. Write the lower and upper threshold values previously determined in step 2 in the Set Threshold Levels window and click "OK." Leave the threshold window open.

5. The threshold color will be applied to the entire image. To calculate the FCSA of the selected ROI only, open the "Analyze" pull-down menu, then select *set measurement* and click on the option *limit to threshold.* This option modification needs to be done only once.

6. To obtain the FCSA of the selected muscle ROI, open "Analyze" pull-down menu and select *Measure* (or click control + M).

7. To reset the image to the initial background, click on the *Reset* button from the Threshold window.

8. Repeat steps 3 to 7 (excluding step 5) to measure the FCSA of another muscle.

Quantitative Paraspinal Muscle Measurements: Inter-Software Reliability and Agreement Using OsiriX and ImageJ

Maryse Fortin and Michele C. Battié

PHYS THER. 2012; 92:853-864.
Originally published online March 8, 2012
doi: 10.2522/ptj.20110380