




## RESEARCH ARTICLE

# Thresholding approaches for estimating paraspinal muscle fat infiltration using T1- and T2-weighted MRI: Comparative analysis using water-fat MRI

Jessica Ornowski<sup>1</sup>  | Lucas Dziesinski<sup>1</sup> | Madeline Hess<sup>2</sup>  | Roland Krug<sup>2</sup> |  
Maryse Fortin<sup>3</sup> | Abel Torres-Espin<sup>4,5,6</sup> | Sharmila Majumdar<sup>2</sup> |  
Valentina Pedaia<sup>2</sup> | Noah B. Bonnheim<sup>1</sup>  | Jeannie F. Bailey<sup>1</sup>

<sup>1</sup>Department of Orthopaedic Surgery, University of California, San Francisco, California, USA

<sup>2</sup>Department of Radiology and Biomedical Imaging, University of California, San Francisco, California, USA

<sup>3</sup>Department of Health, Kinesiology, and Applied Physiology, Concordia University, Montreal, Québec, Canada

<sup>4</sup>School of Public Health Sciences, Faculty of Health, University of Waterloo, Waterloo, Ontario, Canada

<sup>5</sup>Department of Physical Therapy, University of Alberta, Edmonton, Alberta, Canada

<sup>6</sup>Department of Neurological Surgery, University of California, San Francisco, California, USA

## Correspondence

Jeannie F. Bailey, Department of Orthopaedic Surgery, University of California, San Francisco, 95 Kirkham St., San Francisco, CA 94122-0514, USA.  
Email: [jeannie.bailey@ucsf.edu](mailto:jeannie.bailey@ucsf.edu)

## Funding information

National Institute of Arthritis and Musculoskeletal and Skin Diseases, Grant/Award Number: AR076737

## Abstract

**Background:** Paraspinal muscle fat infiltration is associated with spinal degeneration and low back pain, however, quantifying muscle fat using clinical magnetic resonance imaging (MRI) techniques continues to be a challenge. Advanced MRI techniques, including chemical-shift encoding (CSE) based water-fat MRI, enable accurate measurement of muscle fat, but such techniques are not widely available in routine clinical practice.

**Methods:** To facilitate assessment of paraspinal muscle fat using clinical imaging, we compared four thresholding approaches for estimating muscle fat fraction (FF) using T1- and T2-weighted images, with measurements from water-fat MRI as the ground truth: Gaussian thresholding, Otsu's method, K-mean clustering, and quadratic discriminant analysis. Pearson's correlation coefficients ( $r$ ), mean absolute errors, and mean bias errors were calculated for FF estimates from T1- and T2-weighted MRI with water-fat MRI for the lumbar multifidus (MF), erector spinae (ES), quadratus lumborum (QL), and psoas (PS), and for all muscles combined.

**Results:** We found that for all muscles combined, FF measurements from T1- and T2-weighted images were strongly positively correlated with measurements from the water-fat images for all thresholding techniques ( $r = 0.70-0.86$ ,  $p < 0.0001$ ) and that variations in inter-muscle correlation strength were much greater than variations in inter-method correlation strength.

**Conclusion:** We conclude that muscle FF can be quantified using thresholded T1- and T2-weighted MRI images with relatively low bias and absolute error in relation to water-fat MRI, particularly in the MF and ES, and the choice of thresholding technique should depend on the muscle and clinical MRI sequence of interest.

## KEYWORDS

fat infiltration, low back pain, MRI, muscle quality, paraspinal muscles, thresholding, water-fat MRI

## 1 | INTRODUCTION

Chronic low back pain (cLBP) is a leading cause of disability in the world. To better understand potential causes, symptoms, and pathologies of this ailment, there is growing interest in quantifying paraspinal muscle quality (e.g., composition of fat infiltration and lean muscle). Currently, water-fat MRI sequences are considered the contemporary standard for quantifying the fat fraction (FF) within muscles.<sup>1,2</sup> Unfortunately, because of time constraints and cost, these advanced sequences are not included in routine clinical MRI procedures and are difficult to segment due to a poor signal-to-noise ratio. Because of these challenges with water-fat sequences, we tested the validity of estimated FF on T1- and T2-weighted images—which are more suitable for segmentation and are routinely performed on patients during clinical scans—against values from water-fat sequences.

Multiple studies have used T1- and T2-weighted images to estimate FF based on voxel thresholding, but their estimates for FF are difficult to compare due to inconsistency among methodologies.<sup>1</sup> Many of those that do include validation rely on a qualitative assessment of fat (the Goutallier grading system) and MR spectroscopy.<sup>3-5</sup> Furthermore, many studies rely on the selection of single disc levels, unilateral muscle segmentation, and/or a summary statistic of muscle quality that fails to differentiate between the varying pathology along the spine.<sup>1</sup> In a study similar to this one, a research group calculated fat on a T1-weighted sequence and used a fat-water sequence for validation, but they focused their study on the shoulder and used a freely available fuzzy C-means segmentation software, thus leaving room for the analysis of the spine and for the development of more robust thresholding methods.<sup>6</sup> Recent work compared the calculation of FF using T2-weighted images with fat-water sequence for paraspinal muscles in the lumbar spine. Their results provided more evidence that thresholding was a viable way to analyze muscle quality in clinical sequences.<sup>7</sup> However, this group used manual segmentation and manual thresholding methods, focused solely on the L4-L5 and L5-S1 disc levels, and only estimated FF using the T2-weighted sequence. While there are many studies attempting to quantify fat using various imaging techniques, there are gaps in the literature when it comes to widely available, user-friendly, and time-efficient methodologies to better analyze the underlying paraspinal muscle pathology in cLBP patients.

To address this, we applied and compared four different automated thresholding approaches: Gaussian thresholding, Otsu's method, *k*-means clustering, and quadratic discriminant analysis (QDA). Using these methods, we predicted and applied thresholding values to T1- and T2-weighted MR images to quantify fat within the lumbar paraspinal muscles (multifidus (MF), erector spinae (ES), quadratus lumborum (QL), and psoas (PS)) and validated our results with chemical shift encoding-based (CSE) based water-fat MRI which enables accurate measurement of muscle fat.<sup>8</sup> In this study we sought to understand variation in accuracy among different thresholding techniques on the lumbar spine muscles to determine which methods are reliable while using clinical MRI sequences. The results of this study will support the use of more efficient and accurate estimations of fat infiltration in the paraspinal muscles.

## 2 | METHODS

### 2.1 | Subjects

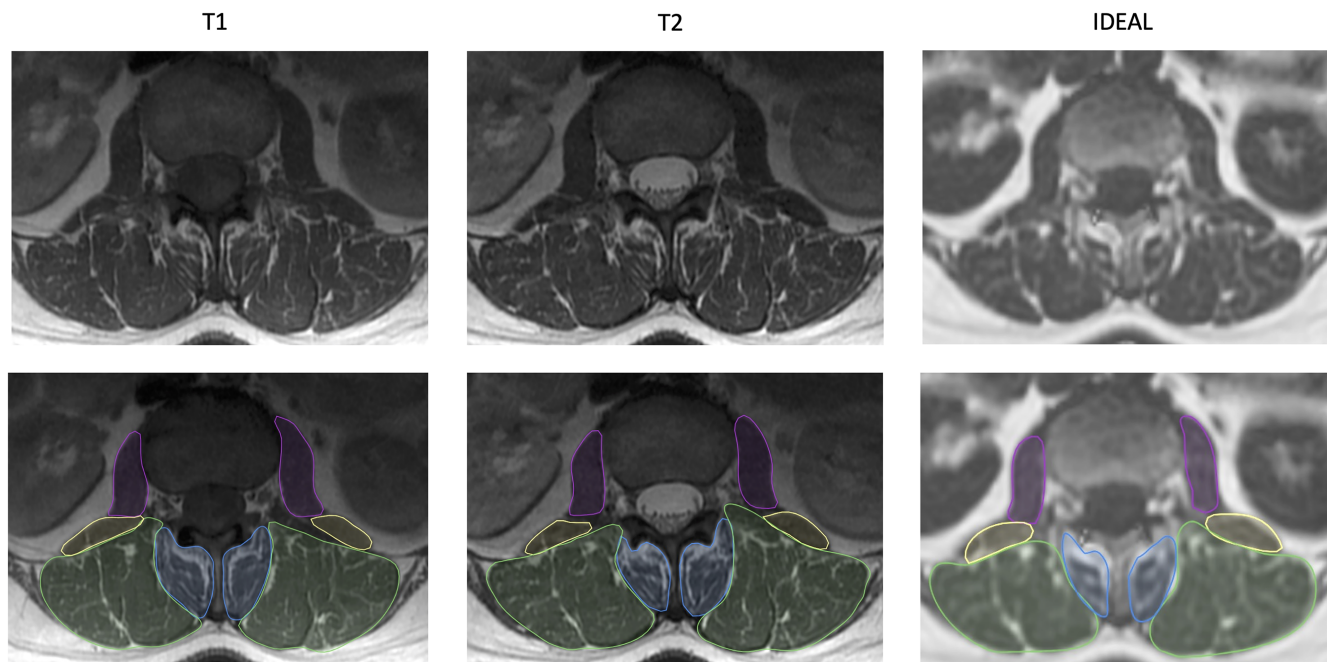
Following IRB (# 20-29928) approval and informed consent, lumbar MRI scans were acquired from 11 patients with cLBP. Patients were recruited from the spine clinic at our institution and were included if they met the criteria for cLBP established by the National Institutes of Health Research Taskforce: low back pain for at least 3 months or on at least half of the days in the past 6 months.<sup>9</sup> Subjects were excluded if they had prior spine surgeries. Subjects were aged 31 to 79 ( $55.0 \pm 14.4$ ), with a height ranging from 150.9 to 184.5 cm ( $167.4 \pm 8.4$ ), a weight ranging from 53.5 to 108.5 kg ( $53.4 \pm 14.3$ ), and BMI ranging from 19.2 to 28.8 kg/m<sup>2</sup> ( $24.3 \pm 2.9$ ).

### 2.2 | Imaging

All subjects were imaged using the same 3T MRI scanner (Discovery MR750; GE Medical Systems, Chicago IL) with an 8-channel phased-array spine coil. The acquisitions of the lumbar spine (L1 through S1) included standard clinical fast spin echo (FSE) sequences with T1- and T2-weighting, and a six echo CSE sequence to acquire water-fat images (Figure 1). The FSE images were acquired using the following parameters for T1- and T2-weighting, respectively: TE = 13, 56 ms; TR = 594, 8414 ms, field-of-view (FOV) = 18 cm; slice thickness = 4 mm; slice gap = 4 mm; and in-plane resolution = 0.35 mm. The CSE acquisition included a six echo 3D spoiled gradient-recalled echo (SPGR) sequence with iterative decomposition of water and fat with echo asymmetry and least-squares estimation (IDEAL) reconstruction with the following parameters: TE = 2, 3, 4, 5, 6, 7 ms; TR = 5.95 ms; FOV = 28 cm; flip angle = 3°; slice thickness = 4 mm; slice gap = 4 mm; in-plane resolution = 1.09 mm; receiver bandwidth = 125 kHz. All images were acquired in the axial plane. All sequences were prospectively applied to the same region of the lumbar spine (L1-S1) and were acquired contemporaneously, enabling accurate anatomic co-location across sequences.<sup>10</sup>

### 2.3 | Muscle segmentation

The bi-lateral MF, PS, ES, and QL muscles were automatically segmented on T1- and T2-weighted images using a previously developed neural network, followed by manual adjustments as needed (Figure 1).<sup>11,12</sup> Bordering epimuscular fat was excluded from muscle segmentations to facilitate assessment of fatty infiltration.<sup>1</sup> Segmentation masks were generated for two axial slices centered at each lumbar disc level (L1-L2, L2-L3, L3-L4, L4-L5, and L5-S1), thus yielding 10 annotated slices per patient. Eight muscle segmentations (one for each bi-lateral muscle) nominally comprised each annotated slice (Figure 1); however, due to variations in muscle morphology, some muscles (particularly the QL and ES) were not able to be segmented at some axial locations. Specifically, the QL was not segmentable



**FIGURE 1** Representative MRI images (upper) and muscle segmentations (lower) for an axial slice centered at the L2-L3 level from T1-, T2-, and IDEAL water-fat images. MF, ES, QL, and PS are highlighted as blue, green, yellow, and purple, respectively.

at L5-S1 for any of the 11 patients, or at L4-L5 for 8 of 11 (73%) patients. Additionally, the ES was not segmentable at L5-S1 level for 6 of 11 (55%) patients. Thus,  $n = 782$  individual muscle segmentations were included in the final analysis.

In order to analyze corresponding anatomic regions between T1-weighted, T2-weighted, and water-fat images (which have different spatial resolutions), the muscle segmentations were transformed from T1- and T2-weighted space to CSE space using an affine transformation utilizing the spatial information embedded in the DICOM metadata.<sup>12</sup> This technique enabled co-localization of the muscle segmentations between the different MRI sequences.

Clinical and water-fat axial slices were paired for segmentation by extracting the minimum difference in patient z-position from image metadata. Although the patient image position between the clinical and water-fat sequences do not maintain perfect alignment, the mean z-position difference per patient for annotated slices was less than 2 mm for both T1- and T2-weighted sequences ( $1.41 \pm 1.62$  mm,  $1.44 \pm 1.54$  mm).

## 2.4 | Image enhancement

The T1- and T2-weighted images were enhanced via contrast limited adaptive histogram equalization to improve tissue contrast (Figure 2). In this method, the DICOM image array is divided into non-overlapping tiles and a histogram of voxel intensities is created for each tile. An intensity clip limit is then set for each image, and the regional histograms are redistributed within that limit. Then, bilinear interpolation was used to re-sample the image.<sup>13</sup> The goal of this

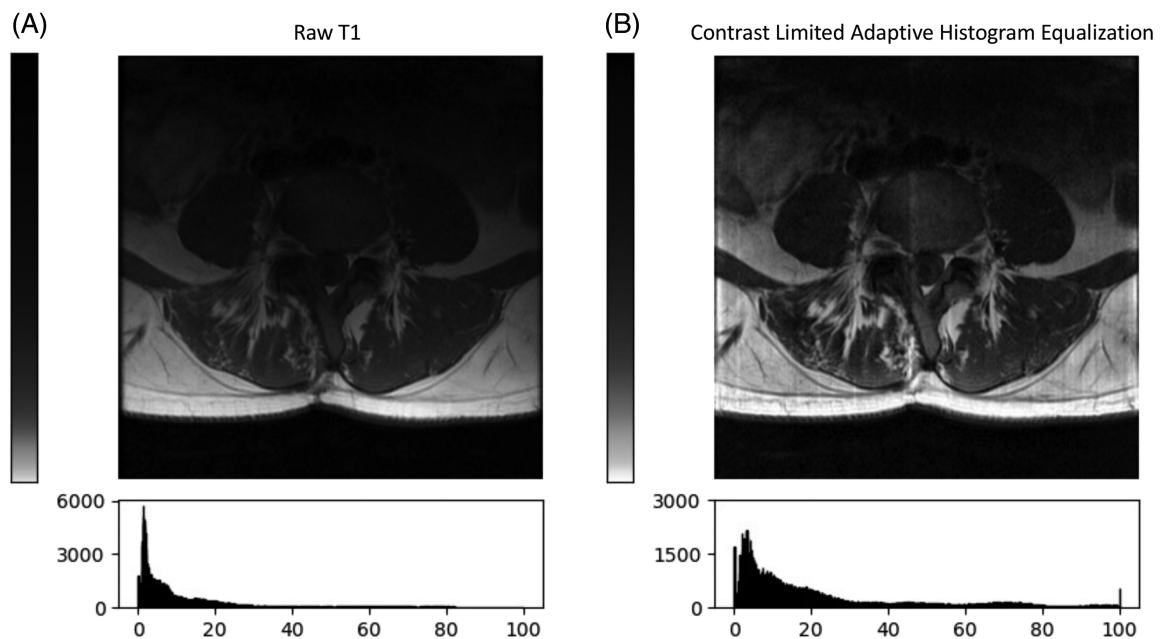
approach was to improve the contrast of the image and to account for any imaging discrepancies to help ensure that differences in voxel intensities are related to pathology. All image enhancement was implemented in Python (v 3.9.12) using scikit-image (v 0.19.2), scikit-learn (v 1.0.2), numpy (v 1.22.3), and pandas (v 1.4.4).

## 2.5 | Image thresholding

The segmented muscles from T1- and T2-weighted images were thresholded to differentiate muscle tissue from intramuscular fat. Each segmented muscle region of interest (ROI) was extracted from the full DICOM array into unique matrices for statistical analysis where each voxel in the matrix was scaled to a grayscale value (0 = black, 255 = white). Four thresholding approaches were then tested: Gaussian curve fitting, Otsu's method, *k*-means clustering, and QDA. Thresholding was implemented in Python (v 3.9.12) using scikit-image (v 0.19.2) and scikit-learn (v 1.0.2).<sup>14,15</sup>

### 2.5.1 | Gaussian method

For each image, two Gaussian curves were fit to the signal intensity histogram of each muscle, thereby assuming a bimodal histogram distribution based on the differential signal from muscle tissue and intramuscular fat. The fat threshold was calculated as the intersection of the Gaussian curves: voxels with signal intensity values below the threshold were classified as muscle and voxels above the threshold were classified as fat.<sup>16</sup>



**FIGURE 2** Representative MRI images depicting un-enhanced (left) and enhanced (right) T1-weighted images and associated histograms. After enhancement the voxels signal intensities have a lower maximum and are more evenly distributed within the histogram.

### 2.5.2 | Otsu's method

Otsu's method is a nonparametric and unsupervised method for image thresholding, which minimizes the intra-class variance in signal intensity (thereby maximizing inter-class variance).<sup>17</sup> Unlike the Gaussian method—which assumed a bi-modal histogram distribution—Otsu's method does not require a priori assumptions regarding histogram shape; rather, Otsu's method requires enumeration of discrete classes. Here, we attempted to differentiate between three tissue classes: fat, muscle, and underlying pathologies. We assumed that the Gaussian method was overpredicting FF because it was mis-classifying the tissue of underlying pathologies as fat, and that using a third class in Otsu's method would capture this erroneous group. To tease out the subtle differences in voxel pixel intensity, the fat-muscle threshold was calculated as the average signal intensity of the two Otsu-determined thresholds.

### 2.5.3 | k-means method

k-means clustering is an unsupervised machine learning approach that separates data into  $k$  clusters. To differentiate groups, the centroid of each cluster is iteratively tested until the intra-cluster sum of squares is minimized between each voxel and the cluster's centroid.<sup>18</sup> Here, we specified  $k = 3$  classes, consistent with the approach used for Otsu's method.

### 2.5.4 | Quadratic discriminant analysis (QDA)

QDA is a supervised learning approach that partitions classes through the optimization of the quadratic discriminant function, which

incorporates several parameters pertaining to voxel signal intensity.<sup>19</sup> Specifically, the model inputs were the signal intensity of each voxel, the mean signal intensity in a  $15 \times 15$  region surrounding the voxel of interest (i.e., the regionally-blurred signal intensity), and mean signal intensity of the entire muscle. For this supervised approach, the ground-truth dataset was developed by first finding the FF for the ROI in the fat-water sequence. The water-fat FF was then used to determine the number of T1- or T2-weighted voxels that needed to be classified as fat so that the clinical sequence FF matched the water-fat sequence FF ( $N_{\text{fat}} = N_{\text{total}}[1 - \text{FF}_{\text{water-fat}}]$  and  $N_{\text{fat}} = \sum n$  where  $SI_n < X_{\text{thresh}}$ ;  $N_{\text{fat}}$  and  $N_{\text{total}}$  are the number of fat voxels and total voxels in each ROI,  $SI_n$  is the signal intensity of the voxel of interest, and  $n$  is the binary value assigned to  $SI_n$ ). Using these formulas, each voxel was assigned a ground-truth binary value (1 for fat, 0 for not fat). With 782 segmented muscles (ranging from 50 to 2500 voxels each), we had a dataset of  $n = 1\,290\,378$  voxels divided into an 80/20 test/train split. The accuracy score for the model's prediction of each voxel as either fat or not-fat was consistently around 87%.

## 2.6 | Outcomes and statistical methods

The primary outcomes of this analysis were the muscle FF values measured for each thresholding approach using the T1- and T2-weighted images, and the ground-truth FF value measured from the water-fat images. For each muscle at each axial slice, the FF value was computed by dividing the total number of voxels classified as fat by the total number of voxels comprising each segmented muscle. The ground-truth muscle FF was quantified as the mean voxel signal intensity within each muscle from the corresponding water-fat image.<sup>8</sup> For each thresholding technique, Pearson correlation coefficients ( $r$ ), mean absolute errors ( $\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$ ), and mean bias

errors ( $MBE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$ ) were calculated for the muscle FF estimations for the thresholded T1- and T2-weighted images relative to the water-fat FF (see Tables 2–4). MBE was used to indicate whether each method over- or under-estimated the true FF, while MAE indicated the average magnitude of the errors regardless of direction. A Fisher's *r*-to-*z* transformation was calculated to obtain 95% confidence intervals (CIs) for *r*. Z-scores were calculated for all permutations of method comparisons (see Supporting Information Table B). Statistical analyses were conducted in Python (v 3.9.12) using numpy (v 1.22.3), pandas (v 1.4.4), scikit-learn (v 1.0.2), and scipy (v 1.7.3).

### 3 | RESULTS

Consistent with demographic and clinical heterogeneity, there was wide variation in mean muscle FF values measured with water-fat MRI (2–74% fat, depending on patient, muscle, and lumbar level; Table 1, Supporting Information Table A). Muscle FF values tended to be highest in the MF (mean  $\pm$  SD = 27.2  $\pm$  8.7%) and lowest in the PS (8.4  $\pm$  4.1%).

Results from pooled muscle analysis show that FF measurements from the T1- and T2-weighted images were strongly correlated with

measurements from the water-fat images across all thresholding techniques ( $r = 0.70$ – $0.86$  depending on thresholding technique and MRI sequence [T1- or T2-weighted],  $p < 0.0001$ , Table 2). Correlation strength tended to be slightly higher for T2-weighted images than T1-weighted images for all thresholding techniques except QDA ( $r = 0.84, 0.83$  for QDA thresholding for T1-, T2-weighted images, respectively). Of the approaches tested, QDA demonstrated the strongest correlation ( $r = 0.84, p < 0.0001$ ), the lowest absolute error (MAE = 5.9%), and the smallest bias (MBE =  $-1.46\%$ ) on T1-weighted images, whereas Otsu's method demonstrated strongest correlation ( $r = 0.86, p < 0.0001$ ), the lowest error (MAE = 5.3%), and the smallest bias (MBE = 1.31%) on T2-weighted images (Tables 2–4, Figure 3).

Variations in inter-muscle correlation strength were much greater than variations in inter-method correlation strength. Specifically, FF estimates from T1- and T2-weighted images were strongly correlated with water-fat measurements for the MF and ES regardless of thresholding technique ( $r = 0.66$ – $0.89$  across thresholding techniques,  $p < 0.0001$ , Table 2). Conversely, T1- and T2-weighted measurements were only moderately or weakly correlated with water-fat measurements in the PS and QL for all thresholding techniques ( $r = 0.16$ – $0.48$  depending on technique,  $p < 0.0001$ ). Thus, the accuracy of FF estimates from T1- and T2-weighted MRI depend more on which muscle is analyzed than which thresholding approach is used.

To get a better visual understanding of the discrepancies in correlation between muscles, we plotted differences in the estimated fat from each thresholding method and the water-fat sequence. Figure 4B (voxels are highlighted under the condition where the clinical sequence voxel was classified as fat *and* the corresponding water-fat sequence voxel was not) allows us to see that the inaccuracies in our estimations follow a similar pattern across methods and sequences. The thresholding methods incorrectly estimate fat along the medial and superior borders of the ES and the medial border of MF, while the mis-estimation of PS is random. Furthermore, the

**TABLE 1** Mean  $\pm$  standard deviation (range) fat fraction measured from IDEAL water-fat MRI.

Muscle	Fat fraction (%)
Multifidus ( $n = 220$ )	27.2 $\pm$ 8.7 (8.2–58.2)
Psoas ( $n = 216$ )	8.4 $\pm$ 4.1 (3.0–23.4)
Erector spinae ( $n = 196$ )	23.5 $\pm$ 13.1 (6.5–73.8)
Quadratus lumborum ( $n = 150$ )	14.3 $\pm$ 7.8 (2.1–45.4)
All ( $n = 782$ )	18.9 $\pm$ 12.1 (2.1–73.8)

**TABLE 2** Pearson's correlation coefficient (*r*) [95% CI] for fat fraction estimates from T1- and T2-weighted MRI relative to IDEAL water-fat MRI for the four thresholding methods tested (Gaussian, Otsu, *k*-means, QDA).

		Gaussian		Otsu		<i>k</i> -means		QDA	
		<i>r</i>	95% CI	<i>r</i>	95% CI	<i>r</i>	95% CI	<i>r</i>	95% CI
Multifidus ( $n = 220$ )	T1	0.72	[0.65, 0.78]	0.71	[0.64, 0.77]	0.66	[0.58, 0.73]	0.78	[0.73, 0.83]
	T2	0.80	[0.75, 0.84]	0.79	[0.74, 0.84]	0.73	[0.66, 0.79]	0.72	[0.65, 0.78]
Psoas ( $n = 216$ )	T1	0.31	[0.18, 0.43]	0.23	[0.01, 0.35]	0.16	[0.03, 0.29]	0.31	[0.19, 0.42]
	T2	0.44	[0.33, 0.54]	0.39	[0.27, 0.50]	0.30	[0.17, 0.42]	0.29	[0.17, 0.41]
Erector spinae ( $n = 196$ )	T1	0.72	[0.65, 0.78]	0.84	[0.79, 0.88]	0.83	[0.78, 0.87]	0.80	[0.75, 0.84]
	T2	0.84	[0.79, 0.88]	0.89	[0.86, 0.92]	0.87	[0.83, 0.90]	0.80	[0.75, 0.85]
Quadratus lumborum ( $n = 150$ )	T1	0.43	[0.29, 0.55]	0.36	[0.21, 0.49]	0.29	[0.14, 0.43]	0.44	[0.31, 0.56]
	T2	0.35	[0.20, 0.48]	0.48	[0.35, 0.60]	0.32	[0.17, 0.46]	0.44	[0.30, 0.56]
All ( $n = 782$ )	T1	0.73	[0.70, 0.76]	0.82	[0.80, 0.84]	0.70	[0.66, 0.73]	0.84	[0.82, 0.86]
	T2	0.8	[0.77, 0.82]	0.86	[0.84, 0.88]	0.76	[0.73, 0.79]	0.83	[0.81, 0.85]

Note: All correlation *p*-values were below 0.0001.

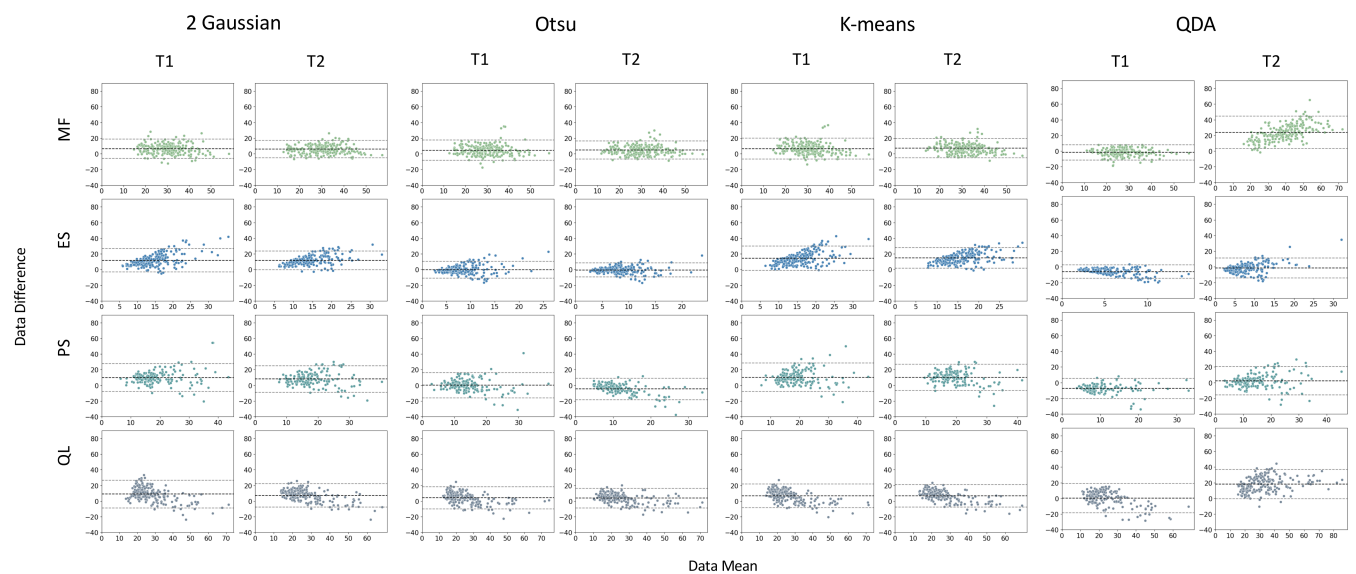
**TABLE 3** Mean absolute error (MAE, %) [95% CI] for fat fraction estimates from T1- and T2-weighted MRI relative to IDEAL water-fat MRI for the four thresholding methods tested (Gaussian, Otsu, *k*-means, QDA).

		Gaussian		Otsu		<i>k</i> -means		QDA	
		MAE	95% CI	MAE	95% CI	MAE	95% CI	MAE	95% CI
Multifidus (n = 220)	T1	7.35	[6.7, 8.0]	5.96	[5.29, 6.63]	7.41	[6.66, 8.16]	5.24	[4.76, 5.73]
	T2	6.68	[6.06, 7.3]	5.93	[5.29, 6.57]	7.78	[7.05, 8.51]	24.0	[22.48, 25.26]
Psoas (n = 216)	T1	1.19	[10.98, 12.87]	3.91	[3.41, 4.4]	14.4	[13.37, 15.4]	5.35	[4.81, 5.9]
	T2	1.17	[10.94, 12.51]	3.25	[2.82, 3.68]	14.9	[14.01, 15.77]	4.88	[4.28, 5.48]
Erector spinae (n = 196)	T1	1.08	[9.83, 11.66]	6.82	[6.2, 7.44]	8.72	[8.0, 9.43]	6.24	[5.62, 6.86]
	T2	9.03	[8.28, 9.78]	6.28	[5.71, 6.85]	8.55	[7.86, 9.24]	18.4	[17.05, 19.65]
Quadratus lumborum (n = 150)	T1	10.9	[9.74, 12.12]	5.82	[4.92, 6.72]	11.2	[9.93, 12.48]	6.95	[5.99, 7.91]
	T2	9.71	[8.63, 10.79]	6.02	[5.08, 6.96]	11.3	[10.2, 12.41]	6.88	[5.82, 7.93]
All (n = 782)	T1	10.2	[9.68, 10.62]	5.59	[5.25, 5.92]	10.4	[9.89, 10.89]	5.85	[5.53, 6.17]
	T2	9.25	[8.83, 9.66]	5.29	[4.97, 5.62]	10.6	[10.15, 11.08]	14.0	[13.18, 14.79]

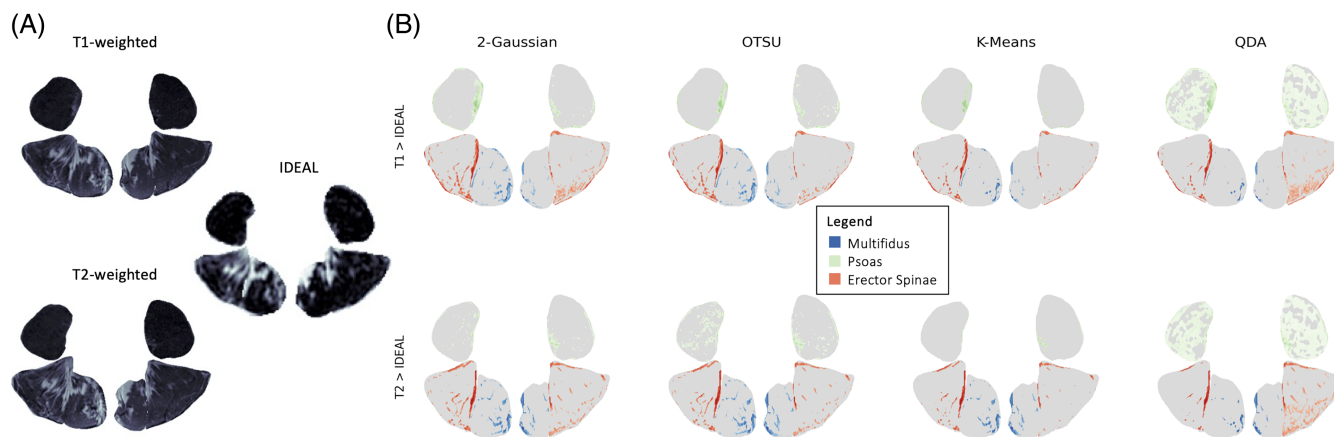
		Gaussian	Otsu	<i>k</i> -means	QDA
Multifidus (n = 220)	T1	6.34	4.48	6.57	2.94
	T2	5.91	4.86	7.15	23.85
Psoas (n = 216)	T1	11.79	-0.23	14.38	-4.34
	T2	11.65	-0.48	14.88	-1.58
Erector spinae (n = 196)	T1	8.91	4.12	6.77	-0.13
	T2	7.26	3.85	6.71	18.18
Quadratus lumborum (n = 150)	T1	9.87	0.08	10.34	-5.54
	T2	7.99	-4.62	10.15	2.26
All (n = 782)	T1	9.18	2.25	9.49	-1.46
	T2	8.23	1.31	9.75	11.26

**TABLE 4** Mean bias error (MBE, %) for fat fraction estimates from T1- and T2-weighted MRI relative to IDEAL water-fat MRI for the four thresholding methods tested (Gaussian, Otsu, *k* means, QDA).

Note: Positive values represent over-estimation, and negative values represent under-estimation.



**FIGURE 3** Bland-Altman between clinical and water-fat sequence of each method, separated by sequence and muscle.



**FIGURE 4** (A) Raw T1-weighted, T2-weighted, and IDEAL water-fat muscle segmentations. (B) Visual representation of the variability in fat classifications between each thresholding method using the T1- and T2-weighted MRI sequence on the L4-L5 disc level of a single patient scan. The colored pixels in the top row represent where each T2 thresholding algorithm classified the pixel as fat, but the IDEAL water-fat sequence did not. The bottom row shows where each T1-weighted thresholding algorithm classified the pixel as fat, but the IDEAL water-fat sequence did not. QL was omitted from this visualization.

thresholding methods did not recognize fat that was classified as such by the water-fat sequence in regions of muscle neighboring the bone.

Z-score calculations and Bland-Altman (BA) plots were used to compare performance between methods. After performing Z-score calculations on all permutations of method comparisons, it can be confirmed that the Gaussian, Otsu, and k-means methods perform better on T2-weighted sequences than on T1-weighted sequences (see Supporting Information Table B). BA plots showed consistency in error shape within muscles across methods, but differences across muscles within methods suggesting that no method was wholly superior.

## 4 | DISCUSSION

This study presents several automated thresholding methods for the quantification of fat infiltration within the paraspinal muscles using clinical MRI sequences. All methods yielded an  $r$  of 0.70 or greater for the overall set of paraspinal muscles, as well as for ES and MF when assessed individually. The T2-Otsu method showed the highest correlation coefficient over all muscles with an  $r$  of 0.86 ( $p < 0.0001$ ) and a 95% CI of [0.841, 0.877], and all methods had the highest correlation coefficients for ES with  $r$ 's ranging from 0.72 for T1-Gaussian to 0.89 for T2-Otsu.

Consistent with a recent study done by Masi et al., the muscles that saw the lowest  $r$  were consistently PS and QL, which also contain the least amount of fatty tissue.<sup>7</sup> Muscles with higher amounts of fat were associated with better estimation results which is likely because the spread of voxel SI values within the muscular region of interest is larger (see Table 1, Table 2, and Figure S2). The clustering algorithms, namely k-means and Otsu, are optimized based on the variance within the set of voxel SI values. Therefore, a set of voxel SI values with a small variance, such as PS and QL, are likely to result in less accurate

clusters. Despite the large differences in performance between muscles, the differences in performance within each muscle varies only slightly between methods. Because of this it is impossible to recommend a "perfect" thresholding option for FF estimation for all muscles of the lumbar spine. However, this analysis does provide us with evidence to suggest that the best thresholding option is dependent on the clinical sequence as well as the muscle of interest.

Assessing the thresholding methods based on their algorithms in addition to their interaction with the data can provide insight into why the correlation varies so much between muscles as well as between methods. Differences between the k-means and Otsu methods are interesting because both techniques use an unsupervised clustering method. The clustering of voxels, however, is where their similarities end. The Otsu method optimizes variance within and between groups, establishing complementary relationships between the clusters, while the k-means method is more single-cluster-focused with the goal of minimizing each point's distance from its cluster's center. Looking at this information in tandem with ground-truth FF of the muscles, it becomes clear why the Otsu method outperformed the k-means method. Muscles with a lower FF have a smaller voxel SI variance which making the separation of voxels more arbitrary for the Otsu method, while the nature of the k-means method allows for a better separation of voxels due to the need to maximize inter-group variance. This small difference is less noticeable in muscles that have higher FF, such as MF and ES because the Otsu's method of minimizing the in-cluster sum of squares is a sufficient way to distinguish the groups.

With the goal to achieve a higher Pearson correlation than the three unsupervised methods, a fourth, supervised learning method was applied: the QDA method. By adding more information about the image to the thresholding process, we predicted that this method would either outperform our previous attempts or would illuminate a ceiling to the Pearson correlation value. Ultimately, this method did

the latter. With an overall  $r$  of 0.82, this method was comparable to the unsupervised methods and requires training data that makes it difficult to reproduce unless fat-water sequences are available. Furthermore, as with any supervised learning method, there is a risk of overfitting the data, especially with a small sample population. This potential challenge was exemplified by the results of the T2-weighted estimations (MBE = 11.26), but interestingly was not a problem for the T1-weighted estimations (MBE = -1.46).

Once it was determined that our summary statistics were likely hitting a ceiling, we considered confounding factors that could be affecting the thresholding performance for estimating FF. The first potential issue is the individual muscle segmentations. Despite the consistency provided by the automated nature of the segmentations, there remains the potential for error. To account for this, we applied secondary manual adjustments to maximize accuracy of the segmentations. Additionally, because the size ratios of the ROI between the clinical sequences and the water-fat sequences are not 1:1, and the difference in patient z-position patient between clinical and water-fat sequences is greater than zero, each segmentation is not 100% reproducible thus leading to skewed FF calculations. Furthermore, within the clinical sequences, each image slice had an intensity gradient leaving it darker at the top than at the bottom. Although image adjustment was applied to each slice to normalize the pixel values, a perfectly balanced image was likely not achieved, leading to slightly skewed clustering.

While there are a variety of potential problems regarding image quality and segmentations, another potential factor influencing performance is the diverse composition of the muscle itself. Our results show that most of the methods overestimated FF. Originally, we suspected that this was a methodological issue, but because of the consistency of the error across all methods, we found that the error likely comes from the incorrect classification of tissue representing various pathologies as fat. When using the  $k$ -means and Otsu methods, we chose a cluster number of 3 to attempt to tease out the intensity differences between fat and other tissues. While the Otsu method did improve upon the Gaussian method, it continued to overestimate FF.

To better understand the meaning behind the inaccuracies in our estimations we created a visualization that highlights the differences between the T2-weighted estimations and the ground-truth FF (see Figure 4B). The location of the classification differences is notable as the T1- and T2-weighted thresholding algorithms appear to classify tissue as fat on the borders of the muscle (where the water-fat calculations do not). Highlighting of the perimeter by the T1- and T2-weighted estimations is likely due to the higher resolution of the clinical images while the highlighting of the deeper muscle fat by the water-fat calculations is likely due to voxel intensity discrepancy.

Finally, we found the BA plots to be insightful regarding our conclusions as to which method is best suited for a given muscle or muscle quality (Figure 3). For example, the MF plots suggest that the errors are consistently random across the range of mean FF (indicating low bias) for all methods except for the QDA method on the T2-weighted sequence. Looking at the QDA method applied to the T2-weighted sequence, higher mean FF values yield higher

positive errors, suggesting that this approach may be sub-optimal for degenerate muscle (despite the low mean bias). For ES, many of the plots exhibit a fan-shaped appearance, indicating heteroscedasticity such that the error increases (both positive and negative) as mean FF increases.

Despite the limitations in this study, all four thresholding methods provide viable options for estimating FF in the lumbar spine. The best method, however, is dependent on the sequence type of the image, muscle of interest, and the muscle quality. The Gaussian, Otsu, and  $k$ -means methods perform better on T2-weighted sequences than on T1-weighted sequences (see Supporting Information Table B). While the differences in performance within each muscle varies only slightly between methods, it is clear that thresholding as a form of FF estimation is much more viable for muscles with higher amounts of fatty infiltration such as MF and ES. While these methods were tested solely on axial scans of the lumbar spine, we are hopeful that their straightforward nature will prove to be easily reproducible for estimating intra-muscular FF in any area of the body.

## AUTHOR CONTRIBUTIONS

**Jessica Ornowski:** Conceived study design; designed research protocol; collected and analyzed the data; wrote the manuscript; and conceived and designed visual figures and tables; **Lucas Dziesinski:** Assisted in study design; assisted in analyzing the data; assisted in and conceived of visual figure design; and critically reviewed manuscript; **Madeline Hess:** Conceived of segmentation protocol and assisted with data analysis; **Noah B. Bonnheim:** Critically reviewed manuscript; **Roland Krug:** Helped conceive study; collected data; and reviewed manuscript; **Maryse Fortin:** Helped conceive study; collected data; and reviewed manuscript; **Abel Torres-Espin:** Provided expertise in data analysis; **Sharmila Majumdar:** Helped conceive study; collected data; and reviewed manuscript; **Valentina Pedita:** Helped conceive study; collected data; and reviewed manuscript; **Jeannie F. Bailey:** Helped conceive study design; assisted in design of research protocol; collected data; and critically reviewed the manuscript.

## CONFLICT OF INTEREST STATEMENT

The authors have declared that there are no conflicts of interest.

## ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute Of Arthritis And Musculoskeletal And Skin Diseases of the National Institutes of Health under Award Number U19AR076737. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also partially funded by an NSF Industry/University Cooperative Research Program called the Center for Disruptive Musculoskeletal Innovations (IIP-1916629).

## ORCID

Jessica Ornowski  <https://orcid.org/0000-0002-4259-6506>

Madeline Hess  <https://orcid.org/0000-0002-3313-932X>

Noah B. Bonnheim  <https://orcid.org/0000-0003-2191-180X>



## REFERENCES

1. Hodges PW, Bailey JF, Fortin M, Battié MC. Paraspinal muscle imaging measurements for common spinal disorders: review and consensus-based recommendations from the ISSLS degenerative spinal phenotypes group. *Eur Spine J*. 2021;30(12):3428-3441. doi:10.1007/s00586-021-06990-2
2. Sollmann N, Bonnheim NB, Joseph GB, et al. Paraspinal muscle in chronic low back pain: comparison between standard parameters and chemical shift encoding-based water-fat MRI. *J Magn Reson Imaging*. 2022;56(5):1600-1608. doi:10.1002/jmri.28145
3. Lee D, Hong KT, Lee W, et al. Threshold-based quantification of fatty degeneration in the supraspinatus muscle on MRI as an alternative method to Goutallier classification and single-voxel MR spectroscopy. *BMC Musculoskelet Disord*. 2020;21(1):362. doi:10.1186/s12891-020-03400-4
4. Ro K, Kim JY, Park H, et al. Deep-learning framework and computer assisted fatty infiltration analysis for the supraspinatus muscle in MRI. *Sci Rep*. 2021;11(1):15065. doi:10.1038/s41598-021-93026-w
5. Shen W, Gong X, Weiss J, Jin Y. Comparison among T1-weighted magnetic resonance imaging, modified Dixon method, and magnetic resonance spectroscopy in measuring bone marrow fat. *J Obes*. 2013; 2013:e298675. doi:10.1155/2013/298675
6. Davis DL, Kesler T, Gilotra MN, et al. Quantification of shoulder muscle intramuscular fatty infiltration on T1-weighted MRI: a viable alternative to the Goutallier classification system. *Skeletal Radiol*. 2019; 48(4):535-541. doi:10.1007/s00256-018-3057-7
7. Masi S, Rye M, Roussac A, et al. Comparison of paraspinal muscle composition measurements using IDEAL fat-water and T2-weighted MR images. *BMC Med Imaging*. 2023;23(1):48. doi:10.1186/s12880-023-00992-w
8. Hu HH, Li Y, Nagy TR, Goran MI, Nayak KS. Quantification of absolute fat mass by magnetic resonance imaging: a validation study against chemical analysis. *Int J Body Compos Res*. 2011;9(3):111-122.
9. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH task force on research standards for chronic low back pain. *J Pain*. 2014; 15(6):569-585. doi:10.1016/j.jpain.2014.03.005
10. Bonnheim NB, Wang L, Lazar AA, et al. Deep-learning-based biomarker of spinal cartilage endplate health using ultra-short echo time magnetic resonance imaging. *Quant Imaging Med Surg*. 2023;13(5): 2807821. doi:10.21037/qims-22-729
11. Bailey JF, Fields AJ, Ballatori A, et al. The relationship between endplate pathology and patient-reported symptoms for chronic low back pain depends on lumbar paraspinal muscle quality. *Spine*. 2019; 44(14):1010-1017. doi:10.1097/BRS.0000000000003035
12. Hess M, Allaire B, Gao KT, et al. Deep learning for multi-tissue segmentation and fully automatic personalized biomechanical models from BACPAC clinical lumbar spine MRI. *Pain Med*. 2023;24(Suppl 1): S139-S148. doi:10.1093/pm/pnac142
13. Anifah L, Purnama IKE, Hariadi M, Purnomo MH. Osteoarthritis classification using self organizing map based on Gabor kernel and contrast-limited adaptive histogram equalization. *Open Biomed Eng J*. 2013;7:18-28. doi:10.2174/1874120701307010018
14. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. Scikit-image: image processing in Python. *PeerJ*. 2014;2:e453. doi:10.7717/peerj.453
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.
16. Shahidi B, Parra CL, Berry DB, et al. Contribution of lumbar spine pathology and age to paraspinal muscle size and fatty infiltration. *Spine*. 2017;42(8):616-623. doi:10.1097/BRS.0000000000001848
17. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1):62-66. doi:10.1109/TSMC.1979.4310076
18. Lloyd S. Least squares quantization in PCM. *IEEE Trans Inf Theory*. 1982;28(2):129-137. doi:10.1109/TIT.1982.1056489
19. Ghoghj B, Crowley M. Linear and quadratic discriminant analysis: tutorial. <http://arxiv.org/abs/1906.02590>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Ornowski, J., Dzieszinski, L., Hess, M., Krug, R., Fortin, M., Torres-Espin, A., Majumdar, S., Padoia, V., Bonnheim, N. B., & Bailey, J. F. (2024). Thresholding approaches for estimating paraspinal muscle fat infiltration using T1- and T2-weighted MRI: Comparative analysis using water-fat MRI. *JOR Spine*, 7(1), e1301. <https://doi.org/10.1002/jsp2.1301>