Reliability of Magnetic Resonance Imaging Readings for Lumbar Disc Herniation in the Spine Patient Outcomes Research Trial (SPORT)

Jon D. Lurie, MD, MS[*], **Anna N. A. Tosteson, ScD**[†], **Tor D. Tosteson, ScD**[†], **Eugene Carragee, MD**[‡], **John Carrino, MD, MPH**[§], **Jay Kaiser, MD**[¶], **Roberto T. Blanco Sequeiros, MD, PhD**[‖], **Amy Rosen Lecomte, MD**[‖], **Margaret R. Grove, MS**[**], **Emily A. Blood, MS**[**], **Loretta H. Pearson, MPhil**[**], **Richard Herzog, MD**[††], and **James N. Weinstein, DO, MS**[**]

[*]Department of Medicine, Dartmouth Medical School, Lebanon, New Hampshire [†]Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, New Hampshire [‡]Department of Orthopaedic Surgery, Stanford School of Medicine, Palo Alto, California [§]Department of Radiology, John's Hopkins Medical School, Baltimore, Maryland [¶]National Orthopedic Imaging Associates, Greenbrae, California [‖]Department of Radiology, Harvard Medical School, Boston, Massachusetts [**]Department of Orthopaedics, Dartmouth Medical School, Lebanon, New Hampshire [††]Department of Radiology and Imaging, Hospital for Special Surgery, New York, New York

Abstract

Study Design—Assessment of the reliability of standardized magnetic resonance imaging (MRI) interpretations and measurements.

Objective—To determine the intra- and inter-reader reliability of MRI parameters relevant to patients with intervertebral disc herniation (IDH), including disc morphology classification, degree of thecal sac compromise, grading of nerve root impingement, and measurements of cross-sectional area of the spinal canal, thecal sac, and disc fragment.

Summary of Background Data—MRI is increasingly used to assess patients with sciatica and IDH, but the relationship between specific imaging characteristics and patient outcomes remains uncertain. Although other studies have evaluated the reliability of certain MRI characteristics, comprehensive evaluation of the reliability of readings of herniated disc features on MRI is lacking.

Methods—Sixty randomly selected MR images from patients with IDH enrolled in the Spine Patient Outcomes Research Trial were each rated according to defined criteria by 4 independent readers (3 radiologists and 1 orthopedic surgeon). Quantitative measurements were performed separately by 2 other radiologists. A sample of 20 MRIs was re-evaluated by each reader at least 1 month later. Agreement for rating data were assessed with kappa statistics using linear weights. Reliability of the quantitative measurements was assessed using intraclass correlation coefficients (ICCs) and summaries of measurement error.

Results—Inter-reader reliability was substantial for disc morphology [overall kappa 0.81 (95% confidence interval (CI): 0.78, 0.85)], moderate for thecal sac compression [overall kappa 0.54 (95% CI: 0.37, 0.68)], and moderate for grading nerve root impingement [overall kappa 0.47 (95% CI:

Address correspondence and reprint requests to Jon D. Lurie, MD, MS, Departments of Medicine and of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH 03756; jon.d.lurie@dartmouth.edu.

0.36, 0.56)]. Quantitative measures showed high ICCs of 0.87 to 0.96 for spinal canal and thecal sac cross-sectional areas. Measures of disc fragment area had moderate ICCs of 0.65 to 0.83. Mean absolute differences between measurements ranged from approximately 15% to 20%.

Conclusion—Classification of disc morphology showed substantial intra- and inter-reader agreement, whereas thecal sac and nerve root compression showed more moderate reader reliability. Quantitative measures of canal and thecal sac area showed good reliability, whereas measurement of disc fragment area showed more modest reliability.

Keywords

disc herniation; MRI; reliability study

Low back pain is one of the most prevalent and costly health problems in the industrial world. Magnetic resonance imaging (MRI) is increasingly used to assess patients with lumbar spine problems, particularly those with sciatica and intervertebral disc herniation (IDH). It is considered the diagnostic imaging procedure of choice for IDH,[1] as it can provide exquisite morphologic detail of the disc abnormality.[2,3] Unfortunately, the relationship between findings on MRI and clinical course remains controversial, with several studies showing a high prevalence of disc "herniations" in asymptomatic subjects.[4-7]

Efforts have been made to improve the specificity of MRI interpretation by developing more precise morphologic terminology than simply "herniation."[5,8] Although disc "extrusions" are much less common in asymptomatic subjects, the reliability of this determination has been variable.[9] Recently, a grading system for determining nerve root compression has been proposed that seems to have substantial reliability, but awaits confirmation in additional studies.[10] Another approach has been to look at quantitative measurement of disc fragment size and canal morphology; preliminary results of this approach appear promising.[11] However, a comprehensive evaluation of the reliability of different features of IDH is lacking.

In this study, we used baseline MRIs collected from patients enrolled in the Spine Patient Outcomes Research Trial (SPORT) with a diagnosis of lumbar IDH. We evaluated the reliability of MRI readings, *i.e.*, the variability in the interpretation and measurements of the same MR image by different readers. Interpretations and measurements were performed by readers using multiple predefined criteria.

Methods

Ratings

SPORT enrolled 1244 patients with IDH defined on the basis of 3 factors: radicular pain with a positive nerve root tension sign or neurologic deficit, a confirmatory imaging study demonstrating IDH corresponding to their symptoms, and presence of symptoms for at least 6 weeks. Baseline MRIs were available and archived for 763 patients. Of these, 92 were collected electronically, deidentified for patient confidentiality, and stored directly as DICOM files. Six hundred seventy-one were collected as printed films and then digitized using a high-definition scanner, deidentified, and stored in DICOM format. No standard imaging protocol was used; clinical films obtained at each participating site were used "as is." Images were provided to the readers on CDs using eFilm Lite software (Merge Technologies; Milwaukee, WI) as a viewer. Display monitors were not standardized across readers.

We randomly selected 60 complete MRI studies for use in this reliability study. Complete images were defined as those containing at least T1 and T2 sagittal series and a T2 axial series. The images were read by 4 clinical experts in spine MRI interpretation, including 3 musculoskeletal radiologists with subspecialty experience in spine imaging, and 1 orthopedic

spine surgeon. Image quality was assessed as good, fair, or inadequate for interpretation. Images deemed inadequate for interpretation were excluded from the study. Image interpretation was recorded using a standardized data collection form that prompted the reader to select from multiple choice lists of findings for imaging characteristics at each level. Images were prepared in monthly batches of approximately 12 studies, including some from patients with IDH and some from patients with spinal stenosis. To assess intrareader reliability, a random subsample of 20 MRIs was selected and reread by each reader at least 1 month after the initial reading.

Each reader received a handbook containing standardized definitions of imaging characteristics. Pictorial and diagrammatic examples were provided where appropriate, derived from the literature or by consensus when no relevant publication was available. Before beginning the study, the readers evaluated a sample set of images and then met in person to review each image and refine the standardized definitions.

The features assessed for IDH included disc morphology, using the published classification scheme of "normal," "bulge," "broad-based protrusion," "focal protrusion," "extrusion," and "sequestered."[12] For analytic purposes this scheme was collapsed into 3 categories: "normal/bulge," "protrusion," and "extrusion/sequestered." This was rated for all available lumbar disc levels. Additional features of thecal sac compression, nerve root impingement, apical location, and sagittal extent of the disc herniation and signal characteristics of the epidural mass were evaluated for all levels that were rated as protrusion or extrusion/sequestered. Thecal sac compression by the disc fragment was characterized as "none," "<1/3," "1/3 to 2/3," or ">2/3."[12] Nerve root impingement was evaluated using the grading system of Pfirrman *et al* and was characterized as "no impingement," "touching" (contact), "displaced" (deviation), or "compressed."[10]

Additional characteristics that were evaluated included the axial location (left extraforaminal/foraminal, left paracentral, central, right paracentral, right extraforaminal/foraminal) and sagittal extent of the herniation when present.[12] In addition, the T2-weighted signal intensity (bright, intermediate, dark) and the signal homogeneity (homogeneous, heterogenous) of the epidural material were rated.

Measurements

In addition to the readings described above, 2 other independent radiologists made quantitative measurements of selected imaging characteristics. For scanned images, scaling was taken from the printed centimeter scale when available; images without any scale were excluded. Measurements were made using ImageJ software's built-in measurement tools (Rasband, W.S., ImageJ, U.S. National Institutes of Health, Bethesda, MD, http://rsb.info.nih.gov/ij/, 1997-2006.) All area measurements were made using freehand areas (Appendix 1, available online through Article Plus). Bony and soft tissue canal area and thecal sac area were measured at all available disc and pedicle levels. Bony canal measurements used the osseous borders posterolaterally and the disc margin anteriorly. Soft tissue canal measurements used the ligamentous borders posterolaterally. Disc fragment area and the thecal sac area at the level of the largest disc fragment were measured only for those levels at which a disc herniation (protrusion, extrusion, or sequestered fragment) was identified.

A detailed handbook was provided to each reader with precise standardized definitions for each measured quantity. Each quantitative reader performed measurements on a training set of images, followed by a feedback session and refinement of the handbook before beginning the

Appendix available online through Article Plus.

study. Measurements were checked for consistency and anatomic plausibility and returned to the readers for remeasurement or rescaling when necessary.

Statistical Analysis

Initial analyses focused on the distribution of selected categories across readers for each imaging characteristic to look for systematic differences in the use of particular categories based on χ^2 tests. The means of the quantitative image measurements were compared between readers using paired t tests.

The kappa statistic[13] was used to summarize intrareader and inter-reader reliability of the rating data. Kappa statistics were calculated with linear weights to give less importance to disagreements closer together on an ordinal scale. Intrareader kappas were calculated for each reader individually and interreader kappas were calculated for each reader pair using the disc level as the unit of analysis. For the intrareader kappas, the bootstrap procedure was implemented using 1000 samples of size 20 from the individual image records included in the reliability study. A stratified estimate of the overall weighted intrareader kappa was formed at each bootstrap iteration. To accommodate the presence of multiple levels per image, overall inter-reader kappas and 95% confidence intervals (CIs) were calculated using the bootstrap technique with 1000 samples of size 58 taken with replacement from the individual image records. A weighted average of the pairwise kappas was taken using weights based on their estimated standard errors. The mean of the bootstrap distribution of the weighted averages was taken as the estimate of the inter-rater kappa. Interpretation of strength of agreement based on kappa values followed the schema of Landis and Koch[13]: <0 = Poor; 0 to 0.20 = Slight; 0.21 to 0.40 = Fair; 0.41 to 0.60 = Moderate; 0.61 to 0.80 = Substantial; 0.81 to 1.00 = Almost perfect.

The primary outcome measure for the quantitative measurements was the intraclass correlation coefficient (ICC) for both inter- and intrareader data. ICC and confidence interval values were calculated using analysis of variance methods as defined by Shrout and Fleiss.[14]

Sample Size

The sample size for this study was planned to allow for an intrareader kappa based on the 20 repeats of a feature with a prevalence of 0.44 to give an approximate predicted standard error for the estimated kappa of 0.11. For a kappa of 0.6 or higher, this would yield a coefficient of variation of less than 20%. For the interobserver kappa, approximately 55 readings by the 4 readers was chosen to give an approximate predicted standard error of 0.028, or a coefficient of less than 5% for kappas of 0.6 or higher.

Results

Of the 60 selected MRIs, 2 were found to be inadequate for interpretation, leaving a total sample size of 58 studies included in the ratings analysis. Of these, 8 did not have an appropriate scale, leaving 50 studies for inclusion in the analysis of quantitative measurements.

Characteristics of the study population are shown in Table 1. The average age was 42.3 years, about half were women, most were white and nonhispanic, most had nerve root tension signs and neurologic deficits, and average Oswestry Disability scores were 45 at baseline. These characteristics were generally similar to the overall IDH population in SPORT.[15,16]

Ratings

Major Characteristics—The distribution of ratings by the 4 expert readers for the main imaging characteristics are shown in Figure 1. Disc morphology had relatively similar

distributions across readers, although Reader C endorsed more protrusions and fewer extrusions than the other readers. Systematic differences in response patterns were more evident for degree of thecal sac compression and most striking for nerve root impingement.

Intrareader reliability for major characteristics is summarized in Table 2. Disc morphology showed almost perfect agreement, thecal sac compression showed substantial to almost perfect agreement, and nerve root impingement showed moderate to substantial agreement. The number of levels evaluated for thecal sac compression and nerve root impingement is substantially smaller than for disc morphology because only levels with a disc herniation present were assessed for these features.

Inter-reader reliability for major characteristics is summarized in Figure 2. Reliability across reader pairs was quite consistent for disc morphology, with an overall substantial to almost perfect agreement [summary kappa = 0.81 (95% CI: 0.78, 0.85)]. Thecal sac compression showed somewhat poorer overall agreement between reader pairs and more variability between pairs. The overall agreement was moderate, with a summary kappa of 0.54 (95% CI: 0.37, 0.68). Nerve root impingement was similar, with overall moderate agreement and a summary kappa of 0.47 (95% CI: 0.36, 0.56).

Additional Characteristics—The axial location showed substantial intrareader reliability [summary kappa 0.78 (95% CI: 0.61, 0.94)] and inter-reader reliability [average kappa 0.76 (95% CI: 0.66, 0.86)]. The sagittal extent of the herniation showed substantial reliability with an intrareader summary kappa of 0.67 (95% CI: 0.51, 0.79) and an inter-reader summary kappa of 0.63 (95% CI: 0.54, 0.70). Ratings of T2 signal characteristics had fair agreement for an intrareader summary kappa of 0.38 (95% CI: -0.02, 0.72) and a moderate inter-reader summary kappa of 0.43 (95% CI: 0.31, 0.54). The ratings of the signal homogeneity had moderate intrareader reliability [summary kappa of 0.58 (95% CI: 0.39, 0.75)] but only poor inter-reader reliability [summary kappa of 0.12 (95% CI: 0.05, 0.20)].

Quantitative Measurements

The results of the measurements by each of the 2 quantitative readers are summarized in Table 3. The mean soft tissue canal area measured at the disc level was 225 mm^2 and the mean thecal sac area was 141 mm^2, with no significant differences between the 2 readers. There were systematic differences between the readers in measures of the bony canal area, the area of the disc fragment, and most markedly in the anterior-posterior length of the disc fragment.

Intra- and inter-reader reliabilities for the quantitative measures are summarized in Table 4. There was excellent intra- and inter-reader reliability for all measures of the canal and thecal sac area. Disc fragment area was somewhat less reliable and the anterior-posterior length of the fragment showed the worst agreement.

Inter-reader agreement for thecal sac area at the disc level is shown graphically in Figure 3. The mean absolute difference between measurements by the 2 readers was 22 mm^2, approximately 15% of the mean 144 mm^2 size. Disc fragment area showed the largest discrepancy between readers. The absolute mean intrarater differences were approximately 19 mm^2 for each reader, with mean disc fragment sizes of 106 and 75 mm^2. The mean interrater difference in measurements was 39 mm^2.

Agreement of Ratings Versus *Measurements*

We also examined agreement between the rater assessments and the quantitative measurements for the 1 parameter that was directly comparable between the 2: degree of thecal sac compression caused by the disc herniation. We compared the subjective assessments of "<1/3,"

"1/3 to 2/3" and ">2/3" with the measured ratio of the thecal sac area at the level of the disc herniation to the thecal sac area at the level of the pedicle above the herniation (<33%, 33%-67%, >67%). The agreement for the subjective rating of thecal sac compression was moderate, with a kappa of 0.54 as previously described. The agreement of the corresponding measured ratio of thecal sac area to the area at the pedicle level was substantial, with an intrareader kappa of 0.63 and a moderate inter-reader kappa of 0.46. The agreement between the ratings of compression and the measured decrement in thecal sac area was fair, with an overall kappa of 0.22 (95% CI: 0.04, 0.41).

Discussion

We found excellent intra- and inter-reader reliability for most of the MRI features assessed in this study. Disc morphology, rated as "normal/bulge," "protrusion," or "extrusion/ sequestered," showed near-perfect intrareader agreement and substantial to near-perfect inter-reader agreement. The degree of thecal sac compression was also highly reliable, whereas the grading of nerve root impingement was only moderate. Our quantitative measurements generally had excellent intra- and inter-reader reliability by ICC, with modest absolute error sizes on remeasurement. Disc fragment area, however, was less reliably measured.

Our results for the reliability of the disc morphology classification compares favorably to prior studies. Brant-Zawadzki *et al* found substantial intrareader agreement (unweighted kappa 0.68) and moderate inter-reader agreement (unweighted kappa 0.59) using the terminology "normal," "bulge," "protrusion," or "extrusion."[8] Jarvick *et al* also found moderate to substantial interreader agreement for this classification with weighted kappas of 0.50 to 0.75 across reader pairs.[9] Similarly, Weishaupt *et al*[7] and Sorensen *et al*[17] found substantial agreement for classifying disc morphology, with interreader kappas of 0.79 and 0.68, respectively. Our reliability was slightly better than these prior studies and may be related to efforts to review criteria and build consensus before undertaking the readings. It may also represent increased familiarity and comfort with this classification system over time.

Despite these efforts, our reliability for grading nerve root impingement was only moderate (overall weighted kappa 0.47). This is lower than initial reliability estimates by Pfirrmann *et al*, who showed substantial interreader reliability for this grading system, with kappas of 0.62 to 0.67 across reader pairs.[10] Our finding of somewhat poorer reliability may be related to the lack of a consistent imaging protocol. Although Pfirrmann *et al* used standard imaging sequences on a single scanner, we used clinically available images with varying image acquisition protocols, field strength, slice orientation, *etc.* This may have contributed to poorer reliability on imaging characteristics that were more finely detailed than disc morphology. The somewhat variable appearance of nerve roots depending on the level and orientation of each slice may also contribute to decreased reliability of grading nerve root compression. It may, however, reflect the type of reliability that could be expected in clinical practice where there is substantial variability in image quality and characteristics.

Although the quantitative measurements showed good reliability in terms of intraclass correlations, the absolute measurement errors were larger than in previous studies. Carragee and Kim's finding of absolute measurement errors of <3% in repeated measures of disc herniations and canal size are much lower than the approximately 15% to 20% seen in the current study.[11] Carlisle *et al* also reported a 3% intraobserver overall measurement variability in a study looking at disc fragment and spinal canal areas.[18] The higher variability in our study may relate to heterogeneity among the images and the limited interaction before and during the study between the 2 radiologists who performed the measurements. The extent to which this degree of measurement error might impede the use of these imaging characteristics to predict clinical findings or outcomes is unknown. The measure of disc fragment area showed

somewhat lower reader reliability, perhaps related to the fact that these structures were irregular, sometimes small, and varied in terms of the level of maximal extent.

This study had a number of important limitations. As noted above, heterogeneity among the images used is a potential shortcoming, in terms of determining the ideal reliability, but may be more representative of actual clinical practice. In addition, there was no standardization across readers in terms of how the readings were done (*i.e.*, all in 1 sitting *vs.* a few at a time) or the monitor on which they were viewed. This could have created substantial differences across readers. In addition, the readers themselves were heterogenous, with 3 radiologists and an orthopedic spine surgeon. Differences in training and background may have affected the inter-reader reliability. Interestingly, however, when we assessed reliability across reader pairs, we did not see any systematic differences in inter-reader agreement based on reader specialty.

It is important to note our use of prestudy meetings, detailed handbooks of definitions, and standardized reporting forms with multiple choice categories for each parameter at each level. These features allowed the assessments to be structured far more than possible in general clinical practice. Thus, our results may overestimate the reliability that might be expected among readers doing routine clinical assessments. In addition, while the readers were not provided with specific clinical data on subjects except their age and sex, they were aware that all the images were from patients with either disc herniation or spinal stenosis severe enough to qualify them to be surgical candidates and enter the SPORT trial. How this knowledge may have affected the readers' interpretations is unknown.

Finally, we studied only the reliability of different readings of the same images. We did not assess the reliability of interpretations between different scans on the same patients, or different imaging protocols. These other factors may introduce entirely separate challenges and create additional possibilities for disagreement.

Disagreements between readers in our study were fairly modest overall. However, when they did occur, we had no gold standard by which to decide between differing interpretations. For example, it is unclear whether the measured thecal sac area or the subjective rating of thecal sac compression is the most "valid." The standard for preferring 1 assessment over the other should not be based on reliability, but rather on whether 1 assessment is able, or better able, to predict patient symptoms or outcome.

The assessment of reliability is merely the first step in this process. The imaging characteristics in this study generally had moderate to substantial intra- and inter-reader reliability. Carlisle *et al* showed that larger disc fragment size, smaller canal area, and larger proportion of canal compromise predicted the need for surgery using a clinical algorithm.[18] Caragee and Kim showed that larger disc fragment size predicted surgical outcomes, but not nonoperative outcomes.[11] Future studies should evaluate whether these features may have potential prognostic implications for the outcomes of surgery compared to nonoperative care in patients with IDH.

Key Points

- In this population of patients with clinical radiculopathy and intervertebral disc herniation, the classification of disc morphology as "normal/bulge," "protrusion," and "extrusion/sequestered" showed substantial agreement within and across readers.

- Classification of thecal sac compression by the disc fragment and grading of nerve root impingement showed moderate agreement.

> - Quantitative measures of canal and thecal sac area showed good reliability, whereas measurement of disc fragment area showed more modest reliability.

Acknowledgments

References

1. Practice parameters: magnetic resonance imaging in the evaluation of low back syndrome (summary statement). Report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology 1994;44:767–70. [PubMed: 8164844]

2. Herzog RJ, Guyer RD, Graham-Smith A, et al. Magnetic resonance imaging. Use in patients with low back or radicular pain. Spine 1995;20:1834–8. [PubMed: 7502142]

3. Modic MT, Ross JS. Magnetic resonance imaging in the evaluation of low back pain. Orthop Clin North Am 1991;22:283–301. [PubMed: 1826552]

4. Boden S, McCowin P, Davis D. Abnormal magnetic-resonance scans of the lumbar spine in asymptomatic subjects: a prospective investigation. J Bone Joint Surg 1990;72A:403–8. [PubMed: 2312537]

5. Jensen M, Brant-Zawadzki M, Obuchowski N. Magnetic resonance imaging of the lumbar spine in people without back pain. N Engl J Med 1994;331:69–73. [PubMed: 8208267]

6. Stadnik TW, Lee RR, Coen HL, et al. Annular tears and disk herniation: prevalence and contrast enhancement on MR images in the absence of low back pain or sciatica. Radiology 1998;206:49–55. [PubMed: 9423651]

7. Weishaupt D, Zanetti M, Hodler J, et al. MR imaging of the lumbar spine: prevalence of intervertebral disk extrusion and sequestration, nerve root compression, end plate abnormalities, and osteoarthritis of the facet joints in asymptomatic volunteers. Radiology 1998;209:661–6. [PubMed: 9844656]

8. Brant-Zawadzki MN, Jensen MC, Obuchowski N, et al. Interobserver and intraobserver variability in interpretation of lumbar disc abnormalities. A comparison of two nomenclatures. Spine 1995;20:1257–63. [PubMed: 7660234]discussion 64

9. Jarvik JG, Haynor DR, Koepsell TD, et al. Interreader reliability for a new classification of lumbar disk disease. Acad Radiol 1996;3:537–44. [PubMed: 8796714]

10. Pfirrmann CW, Dora C, Schmid MR, et al. MR image-based grading of lumbar nerve root compromise due to disk herniation: reliability study with surgical correlation. Radiology 2004;230:583–8. [PubMed: 14699183]

11. Carragee EJ, Kim DH. A prospective analysis of magnetic resonance imaging findings in patients with sciatica and lumbar disc herniation. Correlation of outcomes with disc fragment and canal morphology. Spine 1997;22:1650–60. [PubMed: 9253102]

12. Fardon DF, Milette PC. Nomenclature and classification of lumbar disc pathology. Recommendations of the Combined task Forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology. Spine 2001;26:E93–E113. [PubMed: 11242399]

13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74. [PubMed: 843571]

14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8. [PubMed: 18839484]

15. Weinstein JN, Lurie JD, Tosteson TD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT) observational cohort. JAMA 2006;296:2451–9. [PubMed: 17119141]

16. Weinstein JN, Tosteson TD, Lurie JD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. JAMA 2006;296:2441–50. [PubMed: 17119140]

17. Sorensen SJ, Kjaer P, Jensen ST, et al. Low-field magnetic resonance imaging of the lumbar spine: reliability of qualitative evaluation of disc and muscle parameters. Acta Radiol 2006;47:947–53. [PubMed: 17077047]

18. Carlisle E, Luna M, Tsou PM, et al. Percent spinal canal compromise on MRI utilized for predicting the need for surgical treatment in single-level lumbar intervertebral disc herniation. Spine J 2005;5:608–14. [PubMed: 16291099]
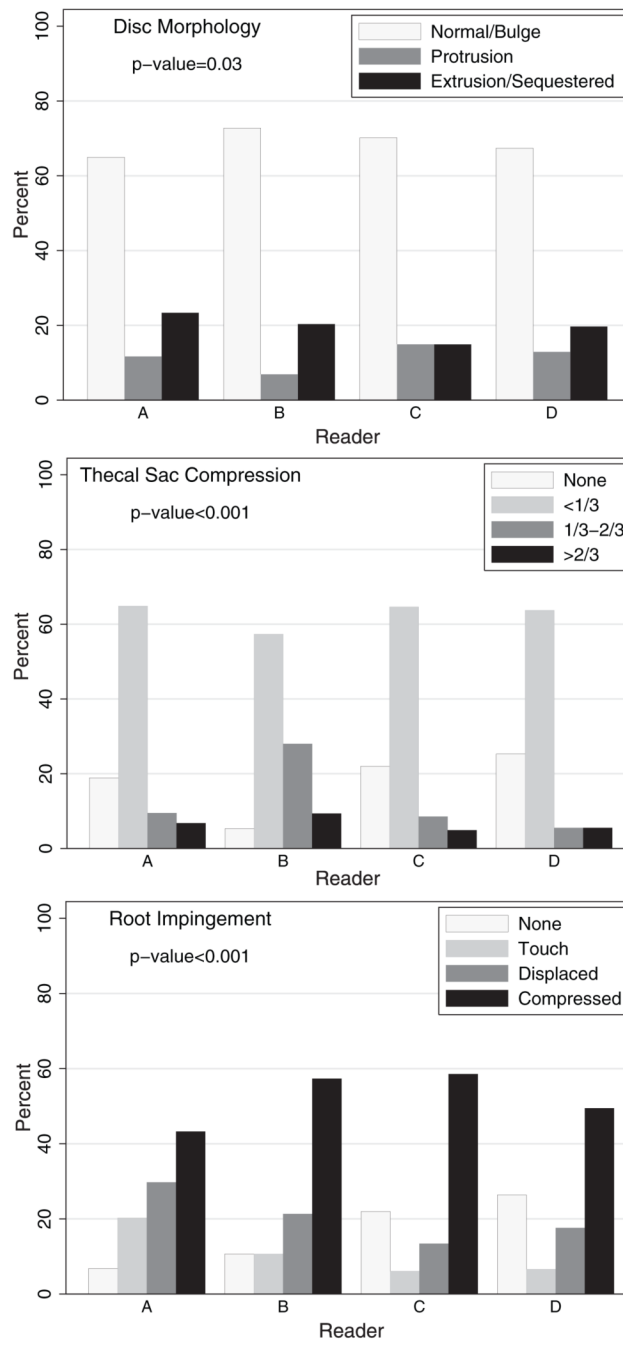
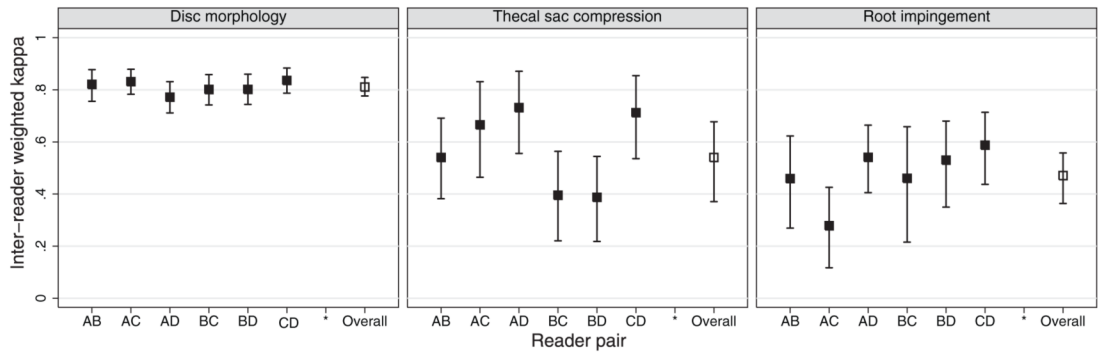**Figure 1.**
Frequency of selected characteristics by reader.

Figure 2.
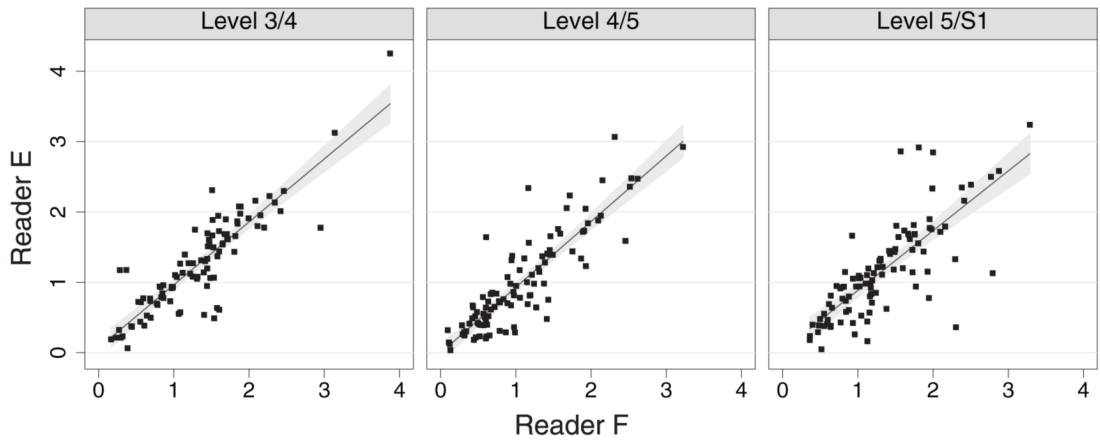Weighted kappas and 95% confidence intervals for inter-reader agreement by reader pairs and overall.

Figure 3.
Inter-reader comparison of thecal sac area at disc level.

Table 1

Baseline Characteristics (n = 58)

Characteristic	n (%)
Age, mean (SD)	42.3 (11.6)
Female	28 (48%)
Ethnicity: not Hispanic	56 (97%)
Race—white	51 (88%)
Mean body mass index (BMI), (stdev)	28.1 (5)
Time since recent episode <6 mo	48 (83%)
Straight Leg Raise Test—ipsilateral	37 (64%)
Straight Leg Raise Test—contralateral/both	7 (12%)
Any neurological deficit	41 (71%)
Herniation level	
L2-L3/L3-L4	6 (10%)
L4-L5	24 (41%)
L5-S1	28 (48%)
Herniation type	
Protruding	15 (26%)
Extruded	39 (67%)
Sequestered	4 (7%)
Posterolateral herniation	46 (79%)
SF-36 bodily pain (BP), score mean (SD)	29.3 (18.2)
Oswestry (ODI), mean (SD)	45 (22.6)

Table 2

Intrareader Reliability as Measured by Weighted Kappa Statistics

Reader	Disc Morphology			Thecal Sac Compression[*]			Root Impingement[*]		
	n1[†]/n2[‡]	kappa[§]	(95% CI)	n1[†]/n2[‡]	kappa[§]	(95% CI)	n1[†]/n2[‡]	kappa[§]	(95% CI)
A	20/83	0.84	(0.75, 0.92)	20/26	0.71	(0.46, 0.91)	20/26	0.46	(0.16, 0.76)
B	20/100	0.87	(0.76, 0.97)	20/23	0.86	(0.62, 1.00)	20/23	0.65	(0.43, 0.86)
C	20/100	0.93	(0.88, 0.97)	20/28	0.93	(0.74, 1.00)	20/28	0.44	(0.05, 0.72)
D	20/99	0.93	(0.88, 0.99)	20/30	0.80	(0.52, 1.00)	20/30	0.83	(0.63, 0.97)
Overall	20/382	0.90	(0.85, 0.94)	20/108	0.84	(0.71, 0.93)	20/108	0.63	(0.49, 0.76)

[*] Contingent on disc morphology.

[†] n1 is the no. of imaging studies.

[‡] n2 is the total no. of levels rated.

[§] Intrarater weighted kappas are based on 1000 bootstrapped samples.

Table 3

Quantitative Measurements by Reader[*]

Characteristic	Reader E Mean (SD)[†]	Reader F Mean (SD)[†]	P
Bony canal area at disc level	421 (142)	377 (129)	0.006
Soft tissue canal area at disc level	225 (87)	233 (85)	0.434
Thecal sac area at disc level	141 (66)	148 (62)	0.306
Thecal sac area at pedicle level above	186 (73)	185 (78)	0.958
Disk fragment area	106 (53)	76 (56)	0.007
Max AP disc fragment length	8.9 (3.0)	6.3 (2.7)	<0.001

[*] n = 50 images; 150 levels.

[†] Lengths were measured in mm, areas in mm^2.

Table 4

Intrareader and Inter-reader Intraclass Correlation Coefficients (ICC) for Quantitative Measurements

Measurement	Intrareader*				Inter-reader[†]	
	Reader E		Reader F			
	ICC	(95% CI)	ICC	(95% CI)	ICC	(95% CI)
Bony canal area at disc level	0.94	(0.89–0.96)	0.89	(0.82–0.93)	0.87	(0.82–0.90)
ST canal area disc level	0.92	(0.87–0.95)	0.95	(0.92–0.97)	0.88	(0.84–0.91)
Thecal sac area at disc level	0.95	(0.92–0.97)	0.96	(0.94–0.98)	0.89	(0.85–0.92)
Canal area at pedicle above	0.95	(0.90–0.97)	0.97	(0.94–0.98)	0.86	(0.80–0.90)
Thecal sac area at pedicle level above	0.94	(0.89–0.97)	0.96	(0.94–0.98)	0.90	(0.86–0.93)
Disk fragment area	0.77	(0.51–0.90)	0.83	(0.62–0.93)	0.65	(0.45–0.79)
Max AP disc fragment length	0.35	(−.09–0.67)	0.66	(0.31–0.85)	0.53	(0.29–0.71)

*
N = 20 images; 60 levels.

[†]
N = 50 images; 150 levels.