Reliability of Readings of Magnetic Resonance Imaging Features of Lumbar Spinal Stenosis

Jon D. Lurie, MD, MS[*,†], **Anna N. 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**[**], **James N. Weinstein, DO, MS**[**], and **Richard Herzog, MD**[††]

[*]Department of Medicine, Dartmouth Medical School, Lebanon, NH

[†]Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH

[**]Department of Orthopaedics, Dartmouth Medical School, Lebanon, NH

[‡]Department of Orthopaedic Surgery, Stanford School of Medicine, Palo Alto, CA

[§]Department of Radiology, John's Hopkins Medical School, Baltimore, MD

[¶]National Orthopedic Imaging Associates, Greenbrae, CA

[‖]Department of Radiology, Harvard Medical School, Boston, MA

[††]Department of Radiology and Imaging, Hospital for Special Surgery, New York, NY

Abstract

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

Objective—To determine the intra- and inter-reader reliability of MRI features of lumbar spinal stenosis (SPS), including severity of central, subarticular, and foraminal stenoses, grading of nerve root impingement, and measurements of cross-sectional area of the spinal canal and thecal sac.

Summary of Background Data—MRI is commonly used to assess patients with spinal stenosis. Although a number of studies have evaluated the reliability of certain MRI characteristics, comprehensive evaluation of the reliability of MRI readings in spinal stenosis is lacking.

Methods—Fifty-eight randomly selected MR images from patients with SPS enrolled in the Spine Patient Outcomes Research Trial were evaluated. Qualitative ratings of imaging features were performed according to defined criteria by 4 independent readers (3 radiologists and 1 orthopedic surgeon). A sample of 20 MRIs was reevaluated by each reader at least 1 month later. Weighted κ statistics were used to characterize intra- and inter-reader reliability for qualitative rating data. Separate quantitative measurements were performed by 2 other radiologists. Intraclass correlation coefficients and summaries of measurement error were used to characterize reliability for quantitative measurements.

Results—Intra-reader reliability was higher than interreader reliability for all features. Inter-reader reliability in assessing central stenosis was substantial, with an overall κ of 0.73 (95% CI 0.69-0.77). Foraminal stenosis and nerve root impingement showed moderate to substantial agreement with overall κ of 0.58 (95% CI 0.53-0.63) and 0.51 (95% CI 0.42-0.59), respectively. Subarticular zone

Address correspondence and reprint request 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.

stenosis yielded the poorest agreement (overall κ 0.49; 95% CI 0.42-0.55) and showed marked variability in agreement between reader pairs. Quantitative measures showed inter-reader intraclass correlation coefficients ranging from 0.58 to 0.90. The mean absolute difference between readers in measured thecal sac area was 128 mm^2 (13%).

Conclusion—The imaging characteristics of spinal stenosis assessed in this study showed moderate to substantial reliability; future studies should assess whether these findings have prognostic significance in SPS patients.

Keywords

spinal stenosis; MRI; reliability

With the aging US population, lumbar spinal stenosis (SPS) is becoming more commonly diagnosed. Magnetic resonance imaging (MRI) is frequently used to assess patients with SPS. Unfortunately, the relationship between findings on MRI and clinical course remains controversial, with several studies showing a high prevalence of anatomic SPS in asymptomatic subjects.[1-4]

A prior study showed only fair inter-reader reliability in the grading of SPS severity on MRI. [5] This result may be due, in part, to the lack of consensus-based criteria for grading SPS. Speciale *et al* also looked at measured crosssectional areas of the spinal canal and found poor agreement between the measured area and the rated severity of stenosis.

In this study, we used baseline MRIs collected from patients enrolled in the Spine Patient Outcomes Research Trial (SPORT) with a diagnosis of SPS with or without degenerative spondylolisthesis. Our objectives were to characterize intra-reader and inter-reader reliability of both qualitative ratings of imaging features and quantitative measurements and to compare them to each other.

Methods

Overview

SPORT enrolled 1261 patients with SPS. This diagnosis was defined on the basis of 3 factors: neurogenic claudication or radicular leg symptoms, a confirmatory imaging study demonstrating SPS, and presence of symptoms for at least 12 weeks. Baseline MRIs were available and archived for 703 patients. Of these, 90 were collected electronically, deidentified for patient confidentiality, and stored directly as DICOM files. 613 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." We randomly selected 60 MRI studies of which 58 were complete and were used in this reliability study. Complete images were defined as those containing at least T1-weighted and T2-weighted sagittal images and T2-weighted axial images.

Qualitative Ratings of Imaging Features

Baseline MRI studies from SPORT participants were rated according to defined criteria by 4 independent clinical experts in spine MRI interpretation, including 3 musculoskeletal radiologists with subspecialty experience in spine imaging, and 1 orthopedic spine surgeon. Quantitative measurements were performed by 2 additional independent radiologists. Images were provided to the readers on CDs using eFilm Lite software as a viewer (Merge Technologies; Milwaukee, WI). Display monitors were not standardized across readers.

Image quality was assessed as good, fair, or inadequate for interpretation. Images deemed inadequate for interpretation by at least 3 of the 4 readers were excluded from the study. Image interpretation was recorded using a standardized data collection form prompting 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 intervertebral disc herniation (another SPORT cohort) and some from patients with SPS. To assess intra-reader 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 SPS included severity of central, subarticular zone, and foraminal stensoses, rated as "none," "mild," "moderate," or "severe." Subarticular zone and foraminal stenosis were rated separately on each side. The subarticular zone, or lateral recess, was defined, as per Fardon and Millette, as extending from the medial edge of the articular facet to the edge of the neuroforamen.[6] A general guideline for severity rating was that mild stenosis represented a compromise of the area in question of ≤1/3 of its normal size, moderate was a compromise between 1/3 and 2/3 of normal size and severe was a compromise >2/3 of normal size. Central and subarticular zone stenoses were rated on the axial T2-weighted images, and foraminal stenosis on the sagittal T1-weighted images. The degree of nerve root impingement by the foramen was rated as "none," "touching," "displacing," or "compressing."

Quantitative Measurements

In addition to the qualitative ratings described above, 2 additional radiologists made quantitative measurements of selected imaging characteristics. For scanned images, scaling was taken from the printed centimeter scale when available. Images without an appropriate scale were excluded. The readers used ImageJ software's built-in measurement tools (Rasband, WS, ImageJ, US National Institutes of Health, Bethesda, MD, http://rsb.info.nih.gov/ij/, 1997-2006.) All area measurements were made using freehand areas. Measurements included cross-sectional areas of the osseous spinal canal, the soft tissue spinal canal, and thecal sac area. Spinal canal and thecal sac area were measured both at the level of the disc and, when possible, at the pedicle level above. This allowed for calculation of the stenosis ratio obtained by dividing the thecal sac area at the disc level by the area at the pedicle level.[7]

A detailed handbook was provided to the 2 quantitative readers, with precise standardized definitions for each measured quantity (see Appendix, available online through Article Plus). Before beginning the study, each reader first performed measurements on a set of training images, followed by a feedback session and refinement of the handbook. As the study progressed, 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. χ^2 tests were used to detect systematic differences among readers in the use of particular categories. The characteristics of the group of patients with image data

Appendix available online through Article Plus.

and the group of SPORT patients without image data were compared using t tests and χ^2 tests, to check the representativeness of patients with image data.

The κ statistic[8] was used to summarize intra-reader and inter-reader reliability of the ratings. κ statistics were calculated with linear weights to give less importance to disagreements closer together on an ordinal scale. Intra-reader κ were estimated for each reader individually and inter-reader κ were estimated for each reader pair using the disc level as the unit of analysis. As there were 4 readers, this resulted in 6 unique reader pairs—readers A, B, C, D result in reader pairs AB, AC, AD, BC, BD, and CD. To accommodate the existence of multiple spinal levels with SPS for any individual, overall and pairwise inter-reader κ and 95% confidence intervals were calculated using bootstrap techniques, with 1000 samples of size 58 taken with replacement from the individual image records including all levels. The mean of the bootstrap distribution was used as the reliability estimate for the pair wise inter-rater κ. A weighted average of the pairwise κ was calculated using weights based on the estimated standard errors to obtain an estimate of the overall κ. The mean of the bootstrap distribution of the weighted averages was used as the reliability estimate for the overall inter-rater κ. For the intra-reader κ, the bootstrap procedure was implemented using 1000 samples of size 20 from the individual image records used in the reliability study. An estimate of the overall weighted intra-reader κ was made at each bootstrap iteration.

The schema of Landis and Koch was used to interpret the strength of agreement based on κ values: <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.[8]

The means of the quantitative image measurements were compared between readers using paired t tests. The primary outcome measure for the quantitative measurements was the intraclass correlation coefficient as applied to intra-reader and inter-reader measurements. Confidence intervals were formed using analysis of variance methods for estimating intraclass correlations.[9]

Results

Of the 60 selected MRIs, 2 were found to be inadequate for interpretation, leaving a total sample size of 58 studies for the qualitative ratings analysis. Because 6 of these did not have an appropriate scale, only 52 studies were included in the analysis of quantitative measurements.

Characteristics of the study population are shown in Table 1. The average age was 64.8 years; about 60% were women; most were white (79%) and non-Hispanic (95%); almost all had neurogenic claudication (93%); most had neurologic deficits (59%); and the average Oswestry Disability score was 46.4 at baseline. These characteristics were generally similar to the SPS/ DS population in SPORT as a whole. Compared with the rest of the SPORT population, patients whose images were included in this reliability study were slightly more likely to have neurogenic claudication (93% *vs*. 83%; $P = 0.048$); slightly less likely to have central stenosis recorded by the enrolling MD (79% *vs*. 89%; $P = 0.042$); and had a trend toward slightly greater disability on the Oswestry disability index (46.4 *vs*. 41.9; $P = 0.061$).

Qualitative Readings

Intra-reader reliability for major characteristics is summarized in Table 2. Overall agreement for central stenosis was excellent with a κ of 0.82 (95% CI 0.78 - 0.87). Overall agreement for subarticular zone stenosis, foraminal stenosis, and nerve root impingement were substantial with κ ranging from 0.75 to 0.77.

Inter-reader reliability is summarized in Figure 1. Agreement on central stenosis was substantial with an overall κ of 0.73 (95% CI 0.69 - 0.77) and very consistent across reader pairs. Foraminal stenosis showed moderate to substantial agreement with an overall κ of 0.58 (95% CI 0.53-0.63) and was also very consistent across reader pairs. Nerve root impingement showed moderate agreement with an overall κ of 0.51 (95% CI 0.42-0.59). Subarticular zone stenosis was the most problematic finding, with moderate agreement (overall κ 0.49; 95% CI 0.42-0.55) and marked variability in agreement between reader pairs.

Quantitative Measurements

The results of the measurements made by each of the 2 quantitative readers are summarized in Table 3. The mean soft tissue canal area measured at the disc level was 166 mm^2 and the mean thecal sac area was 95 mm^2, with no statistically significant difference between the 2 readers. The absolute mean differences in measurements between readers were modest: 7.4 mm^2 (4.5%) for soft tissue canal area and 12.8 mm^2 (13%) for thecal sac area. There were systematic differences between the readers in measures of the osseous canal area.

Intra- and inter-reader reliability for the quantitative measures are summarized in Table 4. There was good intra-reader reliability for soft tissue canal area and thecal sac area, with intraclass correlation coefficients (ICCs) ranging from 0.77 to 0.94, and somewhat lower reliability for the osseous canal area. The inter-reader ICCs were slightly lower than the intra-reader ICCs for all measures, but displayed the same general pattern.

Inter-reader agreement for thecal sac area is shown graphically in Figure 2. Interestingly, the plot shows more scatter at L5-S1, which may be related to variability among studies in the angle of slices through this level.

Agreement of Qualitative Ratings Versus Quantitative Measurements

We also examined agreement between the rater assessments and the quantitative measurements for thecal sac compression with central SPS. We compared the subjective assessments of mild (<1/3), moderate (1/3-2/3), and severe (>2/3) stenosis with the measured ratio of the thecal sac area at the level of the disc to the thecal sac area at the level of the pedicle above (>66%, 66%-33%, <33%). The agreement of the tricotomized measurements was moderate, with an inter-reader κ of 0.45 (95% CI 0.31-0.58). This was lower than the agreement for the subjective severity of central stenosis, which had an overall inter-reader κ of 0.73 (95% CI 0.69 - 0.77).

Discussion

We found substantial reliability for many of the qualitative and quantitative MRI features of SPS assessed in this study. Agreement on the severity of central canal stenosis and foraminal stenosis was good, whereas subarticular zone stenosis showed markedly variable agreement between reader pairs. The measurements of soft tissue canal area and thecal sac area were reasonably reliable, though the tricotomized stenosis ratio showed less reliability than the ratings of central stenosis as mild, moderate, or severe. These findings are important because they suggest that some MRI features may be measured reliably enough to be examined as correlates of prognosis.

To judge the clinical applicability of the levels of agreement seen in this study, we can compare them to the reliability of physical examination features that have been studied in various spine populations. The substantial agreement for central stenosis (κ 0.73) was similar to the most reliable physical examination features studied, such as calfwasting with a κ of 0.80 and crossed straight leg raising in patients with disc herniation with a κ of 0.74.[10,11] The moderate agreement for foraminal stenosis (κ 0.58) is similar to the reliability of the assessment of pain

with bending (κ 0.56) or pain with resisted external hip rotation (0.63).[11] The lowest agreement in our study (κ 0.45) was similar to the agreement seen for reproducibility of bony tenderness (κ 0.40) or Achilles reflex deficit (κ 0.39-0.50).[10]

Our results compare favorably with prior studies of imaging interpretation in SPS. Speciale *et al* reported an overall interobserver κ of 0.26 for ratings of stenosis severity.[5] This much poorer agreement may stem from the lack of discussion or definition of what constituted mild, moderate, or severe in that study. We attempted to define in advance all ratings used in our study, convening in-person meetings to review cases in order to reach consensus on an approach. In addition, the agreement reported by Speciale *et al* seems to include foraminal and lateral recess stenosis along with central stenosis. We found wide variability in the agreement between reader pairs for subarticular (lateral recess) stenosis ratings.

The quantitative measurements showed reasonably good reliability in terms of intraclass correlations. The differences between measurements ranged from 4.8% to 13%. The reliability of the thecal sac area has been previously studied. Haminishi reported a correlation coefficient for the dural sac area of 0.92 and Weiner a correlation of 0.91.[12,13] However, these values were for Pearson's correlation coefficient rather than the intraclass correlation and values for absolute differences between measurers were not reported. In our current study, the reliability of the measured thecal sac stenosis ratio did not seem to be more reliable than the subjective rating of central canal severity.

This study had a number of important limitations. Despite our efforts to define terms and reach consensus on rating procedures, we relied on clinically available images with varying image acquisition protocols, field strength, slice orientation, *etc*. This may have contributed to poorer reliability for some imaging characteristics. However, it is likely to reflect the level of reliability that could be expected in clinical practice where there is substantial variability in image quality.[14] In addition, there was no standardization across readers in terms of the setting or equipment on which the readings were done. This could have contributed to the differences between readers. In addition, the readers themselves were heterogeneous (3 radiologists and an orthopedic spine surgeon); 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, although 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 SPS severe enough to qualify them as surgical candidates. How the lack of "normal" studies may have affected the readers' interpretations is unknown.

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 central stenosis severity is the most "valid." The standard for preferring 1 assessment over another should not be based on reliability alone, 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. Future studies should assess these ratings and measurements for their potential prognostic implications in predicting outcomes.

Key Points

- In this cohort of patients with spinal stenosis and neurogenic claudication with or without associated degenerative spondylolisthesis, ratings of central stenosis, foraminal stenosis, and thecal sac area showed moderate to substantial intra-reader and inter-reader reliability.

- Rating of subarticular zone stenosis and measures of osseous canal area were less reliable.

- Future studies should assess the prognostic significance of these findings.

Acknowledgments

References

1. 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]

2. 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]

3. 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]

4. 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]

5. Speciale AC, Pietrobon R, Urban CW, et al. Observer variability in assessing lumbar spinal stenosis severity on magnetic resonance imaging and its relation to cross-sectional spinal canal area. Spine 2002;27:1082–6. [PubMed: 12004176]

6. 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]

7. Laurencin CT, Lipson SJ, Senatus P, et al. The stenosis ratio: a new tool for the diagnosis of degenerative spinal stenosis. Int J Surg Investig 1999;1:127–31.

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

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

10. Deyo RA, Rainville J, Kent DL. What can the history and physical examination tell us about low back pain? JAMA 1992;268:760–5. [PubMed: 1386391]

11. McCombe PF, Fairbank JC, Cockersole BC, et al. 1989 Volvo Award in clinical sciences. Reproducibility of physical signs in low-back pain. Spine 1989;14:908–18. [PubMed: 2528822]

12. Hamanishi C, Matukura N, Fujita M, et al. Cross-sectional area of the stenotic lumbar dural tube measured from the transverse views of magnetic resonance imaging. J Spinal Disord 1994;7:388–93. [PubMed: 7819638]

13. Weiner BK, Patel NM, Walker MA. Outcomes of decompression for lumbar spinal canal stenosis based upon preoperative radiographic severity. J Orthop Surg 2007;2:3.

14. Jarvik JG, Robertson WD, Wessbecher F, et al. Variation in the quality of lumbar spine MR images in Washington State. Radiology 2000;215:483–90. [PubMed: 10796929]

Figure 1.
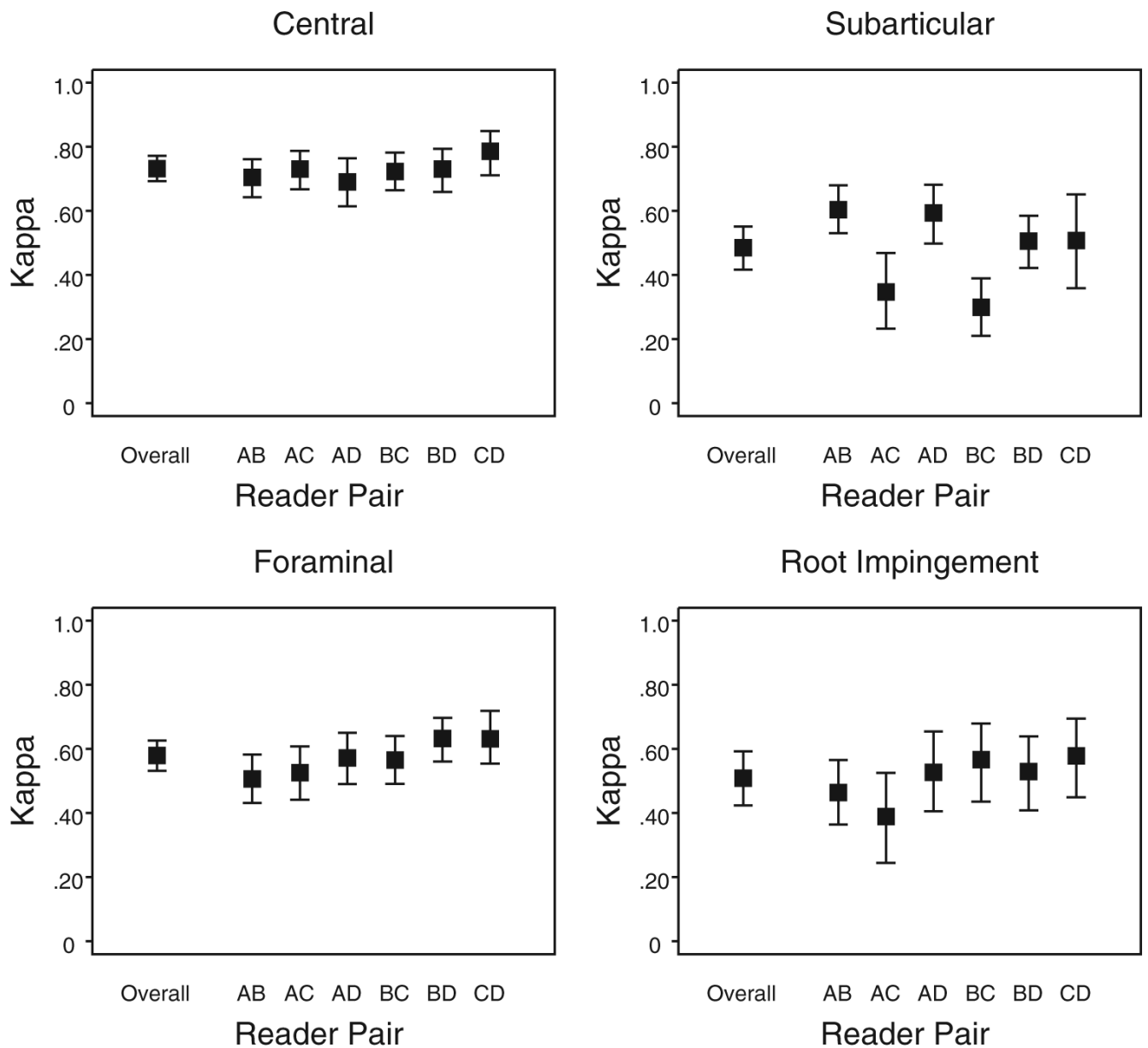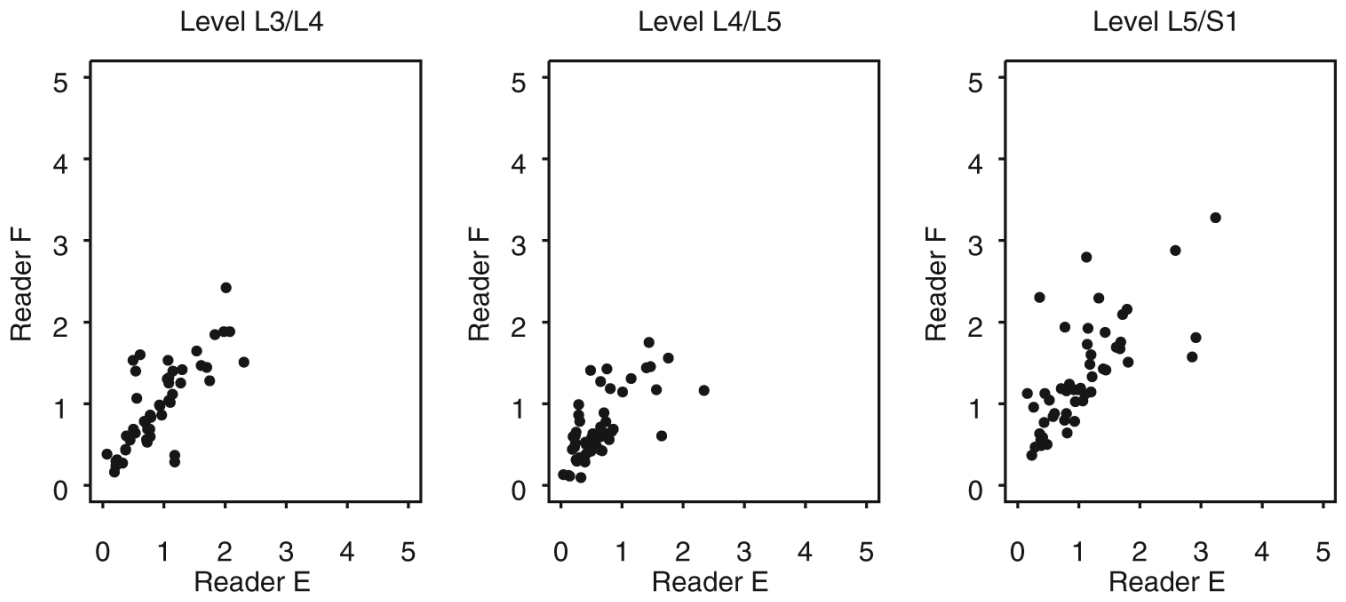Inter-reader reliability of qualitative ratings: Weighted κ, with 95% bootstrap confidence intervals.

Figure 2.
Inter-reader reliability of quantitative measurements: inter-reader comparison of thecal sac area at disc area for 52 patients.

Table 1

Subject Characteristics: Comparison of Baseline Data Among DS/SPS Patients With and Without Image Data

	Image Data		
	Yes (n = 58)	No (n = 1203)	P
Mean age (SD)	64.8 (10.7)	65.4 (11.1)	0.72
Female	34 (59)	638 (53)	0.49
Ethnicity: not Hispanic	55 (95)	1161 (97)	0.76
Race—white	46 (79)	1013 (84)	0.42
Mean body mass index (BMI), (SD)	29.2 (5.9)	29.3 (5.9)	0.90
Time since recent episode <6 mo	24 (41)	495 (41)	0.92
Pseudoclaudication-any	54 (93)	988 (82)	0.048
SLR or femoral tension	8 (14)	214 (18)	0.55
Pain radiation-any	48 (83)	938 (78)	0.48
Any neurologic deficit	34 (59)	654 (54)	0.62
Listhesis level			0.32
L3-L4	5 (9)	52 (4)	
L4-L5	26 (45)	524 (44)	
Stenosis levels			
L2-L3	6 (10)	232 (19)	0.13
L3-L4	29 (50)	642 (53)	0.71
L4-L5	56 (97)	1127 (94)	0.54
L5-S1	9 (16)	226 (19)	0.65
Stenotic levels (mod/severe)			0.90
None	2 (3)	37 (3)	
One	31 (53)	586 (49)	
Two	17 (29)	402 (33)	
Three+	8 (14)	178 (15)	
Stenosis locations			
Central	46 (79)	1070 (89)	0.042
Lateral recess	53 (91)	1016 (84)	0.21
Neuroforamen	22 (38)	436 (36)	0.90
Stenosis severity			0.98
Mild	2 (3)	37 (3)	
Moderate	23 (40)	485 (40)	
Severe	33 (57)	681 (57)	
Bodily pain (BP) score	29.4 (16.8)	31.5 (17.3)	0.36
Oswestry (ODI)	46.4 (19)	41.9 (18.2)	0.061

Values inside parentheses indicate percentages.

Table 2

Intra-reader Reliability of Qualitative Ratings: Weighted κ With 95% Bootstrap Confidence Intervals

Reader	Central Stenosis	Subarticular Zone Stenosis	Foraminal Stenosis	Root Impingement
A	0.79 (0.67-0.9)	0.78 (0.69-0.86)	0.87 (0.8-0.95)	0.8 (0.62-0.92)
B	0.78 (0.69-0.86)	0.74 (0.65-0.84)	0.8 (0.7-0.9)	0.75 (0.64-0.85)
C	0.86 (0.75-0.94)	0.66 (0.43-0.85)	0.72 (0.59-0.83)	0.82 (0.67-0.94)
D	0.85 (0.79-0.92)	0.71 (0.64-0.81)	0.7 (0.6-0.8)	0.63 (0.42-0.82)
Overall[*]	0.82 (0.78-0.87)	0.75 (0.69-0.81)	0.77 (0.72-0.82)	0.76 (0.68-0.83)

Twenty images, levels 1 through 5 included.

[*] Overall κ is computed by comparing first and second reads across all readers.

Table 3

Comparison of Quantitative Readings Between Readers E and F (1 Read From Each)

Measurement	Reader E (n = 153)	Reader F (n = 153)	P	Combined (n = 306)
Max AP canal length—ST canal	1.42 (0.4)	1.46 (0.4)	0.34	1.44 (0.38)
Max AP canal width—ST canal	1.16 (0.6)	1.17 (0.6)	0.92	1.17 (0.58)
Max AP canal length —osseous canal	2.09 (0.4)	2.09 (0.4)	0.90	2.09 (0.41)
Max AP canal width—osseous canal	2.18 (0.6)	1.97 (0.5)	<0.001	2.08 (0.54)
Osseous canal area	3.65 (1.3)	3.25 (1)	0.004	3.45 (1.19)
ST canal area	1.62 (0.9)	1.7 (0.9)	0.46	1.66 (0.89)
Thecal sac area	0.887 (0.6)	1.01 (0.6)	0.067	0.95 (0.61)
Canal area at pedicle above	3.34 (1.3)	3.06 (0.9)	0.053	3.19 (1.14)
Thecal sac area at pedicle above	1.6 (0.8)	1.57 (0.6)	0.71	1.58 (0.68)

Eighteen images, levels 3 through 5 included.

Table 4

Reliability of Quantitative Measurements: Intraclass Correlation Coefficients (ICC)

Measurement	Intra-ICC		Inter-ICC
	Reader E	Reader F	
Max AP canal length—ST canal	0.63 (0.44-0.77)	0.9 (0.83-0.94)	0.76 (0.68-0.82)
Max AP canal width—ST canal	0.92 (0.87-0.95)	0.93 (0.89-0.96)	0.9 (0.86-0.92)
Max AP canal length—osseous canal	0.7 (0.53-0.81)	0.86 (0.77-0.92)	0.65 (0.55-0.74)
Max AP canal width—osseous canal	0.9 (0.83-0.94)	0.86 (0.78-0.92)	0.68 (0.58-0.75)
Osseous canal area	0.71 (0.55-0.82)	0.88 (0.8-0.93)	0.58 (0.47-0.68)
ST canal area	0.77 (0.64-0.86)	0.94 (0.9-0.97)	0.78 (0.71-0.83)
Thecal sac area	0.86 (0.76-0.91)	0.9 (0.83-0.94)	0.76 (0.68-0.82)
Canal area at pedicle above	0.57 (0.35-0.73)	0.77 (0.63-0.86)	0.63 (0.52-0.73)
Thecal sac area at pedicle above	0.28 (0-0.51)	0.86 (0.77-0.92)	0.63 (0.52-0.72)

Intra-reader variables collected on 18 patients 3 levels each. Inter-reader variables collected on 52 images 3 levels each. ICC and 95% CI based on Shrout-Fleiss formulas.