Assessment of Magnetic Resonance Imaging in the Diagnosis of Lumbar Spine Foraminal Stenosis—A Surgeon's Perspective

Naftaly Attias, MD,* Anne Hayman, MD,† John A. Hipp, PhD,‡ Philip Noble, PhD,‡ and Stephen I. Esses, MD§

Study Design: A 2-part cadaveric study.

Objectives: Part 1: To assess the reliability of a semiquantitative classification system for grading lumbar spine foraminal stenosis and the variability in magnetic resonance imaging (MRI) examinations carried out in different institutions. Part 2: to assess the difference between the foraminal measurements carried out on MRI images and on cadaveric specimens.

Summary of Background Data: There are limited data to suggest that MRI examinations are sensitive or specific for the diagnosis of lumbar spine foraminal stenosis. The effect of the variability in techniques, used by different imaging centers, is not well understood.

Methods: Three fresh, frozen human lumbar spines were examined with 3 different clinical MRI systems. Eight spine surgeons used the MRI images to grade the foramina, on the basis of a previously suggested semiquantitative classification. In addition, the dimensions of all foramina were measured using software available on each of the MRI systems. The dimensions of the specimens' foramina were then measured in situ, graded, and compared with the MRI measurements.

Results and Conclusions: There was poor intraobserver agreement using the semiquantitative grading system. The parameters associated with the grade of stenosis assigned to the foramen were as follows: (1) the observer doing the grading, (2) the place it was imaged, and (3) the location of the foramen. There was poor correlation between measurements of the foramina carried out on MRI and the specimens.

Key Words: MRI, lumbar spine foramen, foraminal stenosis, foramen measurement

(*J Spinal Disord Tech* 2006;19:249–256)

Jenis and An[1] identified lumbar spine foraminal stenosis (LSFS) as an important pathologic entity to be recognized in the patient with radicular symptoms. The authors estimated that the incidence of LSFS is between 8% and 11%, and concluded, as did Czervionke et al,[2] that magnetic resonance imaging (MRI) may be the preferred study for evaluating the symptomatic patient with LSFS.

Although MRI is widely used, and is considered by many as an appropriate tool for studying spine pathologies, there are limited data to suggest that MRI examinations are sensitive or specific for the diagnosis of LSFS. In addition, although the variability in techniques used by different imaging centers, and different radiologists, is widely recognized by surgeons, the effect of this variability on assessing foraminal stenosis is not well understood. Furthermore, the surgeons, usually, do not have control over the final quality, or quantity of images provided to them.

Glenn[3] subjectively used the words "normal, slight, mild, moderate, and severe" to estimate the size of a lesion in the spine. This system of evaluation is widely used by radiologists, and was recently investigated by Speciale et al.[4] The authors analyzed the results of a blinded rating by 7 observers of MRI scans performed on 15 patients with lumbar stenosis. Their findings indicated only a fair level of agreement among the observers. Their average interobserver κ score was 0.26, but showed variability within different specialties: radiologists—0.40, neurosurgeons—0.21, and orthopedic surgeons—0.15. Their average intraobserver κ score was 0.11. The authors suggested that further studies should be carried out to evaluate a consensus-based, standardized MRI classification, aimed at improving agreement among observers.

Recently, Wildermuth et al[5] introduced a semiquantitative classification system for grading LSFS, on the basis of MRI findings. The severity of LSFS was graded "0" if the foramina was without pathology, "1" if slight foraminal stenosis was present, "2" if marked foraminal stenosis was present, and "3" to indicate advanced stenosis (Table 1). They acquired images of the lumbar spine of symptomatic patients in flexion and extension positions, using an "open" MRI, and used this system to grade the changes in the size of the foramina in the 2 positions. This system, to the best of our knowledge,

TABLE 1. Wildermuth's MRI Grading System for Grading Lumbar Spine Foraminal Stenosis

Grade 0	Normal foramina [normal dorsolateral border of the intervertebral disk and normal form at the foraminal epidural fat (oval or inverted pear shape)]
Grade 1	Slight foraminal stenosis and deformity of the epidural fat, with the remaining fat still completely surrounding the exiting nerve root
Grade 2	Marked foraminal stenosis, with epidural fat only partially surrounding the nerve root
Grade 3	Advanced stenosis with obliteration of the epidural fat

is the only system available today that is MRI based, and was chosen for this reason to be used in this study, even though it was not validated in clinical trials.

To address the above-mentioned concerns, a 2-part study was conducted to investigate the use of MRI in diagnosing LSFS. In the first part, MRIs of 3 cadavers, carried out in 3 centers, were studied by 8 spine surgeons, and the purpose of this part was to assess the reliability of Wildermuth semiquantitative classification system for grading LSFS, and the variability in MRI examinations carried out in different institutions. In the second part, measurements from those MRIs were compared with measurements obtained from the cadaveric macroscopic slices, and the purpose of this part was to assess the differences between the foraminal measurements carried out on MRI images and on cadaveric specimens.

MATERIALS AND METHODS

Definition

The foramina were defined according to Jenis and An[1] and Delamarter et al[6] as a vertical interpedicular zone incorporating portions of the lateral recess and exit zone as described by Lee et al[7] (Fig. 1).

Part 1

Three fresh, frozen human lumbar spines specimens, containing L1 to S1 vertebrae, were obtained from the anatomic gifts program at Baylor College of Medicine, Houston, TX. Two were of men aged 69 and 80 years, and 1 was of a woman aged 81 years. A cuff of soft tissue was left around the spine when harvested. Under fluoroscopic imaging, rigid plastic tubes of 2 mm diameter were inserted bilaterally through the vertebra to mark the pedicles of the proximal and distal vertebrae. The spine specimens were placed in a clear plastic bag and then vacuum sealed. The specimens were kept frozen at −20°C, thawed for 8 hours before the scheduled MRI examinations, and deep frozen again, immediately after the examinations.

These spine specimens were examined by 3 clinical MRI facilities, each with a different type of MRI scanner:

MRI no. 1—1.5 T scanner (Signa, GE Medical Systems, Milwaukee, WI).

MRI no. 2—0.3 T open MRI scanner (AIRIS II Premium, Hitachi, Tokyo, Japan).
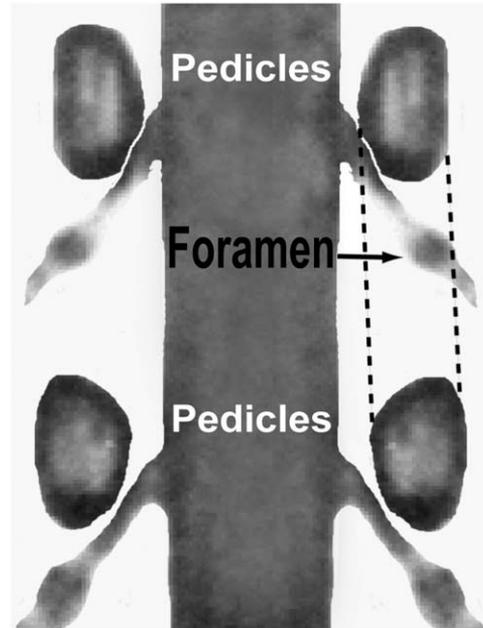
FIGURE 1. The foramen is the vertical interpedicular zone as suggested by Jenis et al.[1]

MRI no. 3—1.5 T scanner (Magneton Symphony, Siemens, Erlangen, Germany).

The technologists were asked to examine the spine section, using techniques consistent with those that they would have used in clinical practice. The investigator was careful not to influence the imaging sequences that they used. The only exception was the suggestion to use extremity coils to perform the examinations owing to the small size of the specimens. Their protocol included T1-weighted and T2-weighted sagittal scan cuts, T1-weighted coronal scan cuts, and T1-weighted axial scan cuts. The images from the MRI examination were recorded on film. Each specimen was examined using each of the 3 MRI scanners.

The complete set of MRI images was submitted to 8 observers for assessment and grading. The observers were all orthopedic surgeons with experience in spine surgery (most of them completed a "spine fellowship"). They were recruited to the study if the majority of their practice entailed treating patients with spine diseases. They reviewed 9 sets of MRI scans, blinded to the type of MRI scanner and to the specimen examined. Each set of MRI scans was hung on a light board in a separate room, and the observers moved from room to room, at their convenience, until they had completed their data collection sheets. Each observer was instructed to assess the foramina of L2-L3, L3-L4, L4-L5, and L5-S1 bilaterally in each set of MRI scans, and to grade the severity of foraminal stenosis on a 4-point scale based on the Wildermuth et al[5] grading system: *Grade 0*, *Grade 1*, *Grade 2*, and *Grade 3*. Before the study, the observers were given a short lecture explaining the Wildermuth et al grading system, and their data collection sheets contained

a copy of the grading system, as it appeared in the original article. The observers were also asked to choose their preferred MRI sequences for the assessment of foraminal stenosis.

Part 2

MRI Measurements

The foramen measurements were obtained on one of the MRI consoles by using the MRI manufacturer's software (Magneton Symphony, Siemens, Erlangen, Germany), and by using "Rad Work standard 5.1" software (Applicare Medicare Imaging, Zeist, The Netherlands) for the other 2 scanners. The measurements were obtained by one of us (N.A.), from the images of the sagittal sections that showed the smallest foraminal dimension (Fig. 2). An effort was made to obtain reproducible, comparable values. The parameters that were measured were as follows:

1. Foraminal height was defined as the maximum distance between the inferior margin of the pedicle of the superior vertebra and the superior margin of the pedicle of the inferior vertebra (Fig. 2).
2. Superior foraminal width was defined as the anterior-posterior width measured in the horizontal plane. The superior width was the maximum width in the superior parts of the foramen (Fig. 2).
3. Middle foraminal width was defined as the anterior-posterior width measured in the horizontal plane. The middle width was the width in the central part of the foramen measured at the level of the middle height of the disc (Fig. 2).
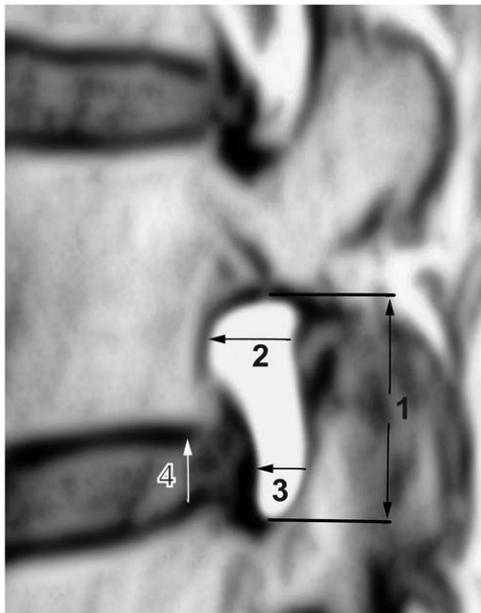
FIGURE 2. Measurement carried out on cadaver and MRI. 1, Foraminal height; 2, superior foraminal width; 3, middle foraminal width; 4, posterior disc height.

4. Posterior disc height was defined as the shortest distance between the adjacent superior and inferior end plates, measured in the planes of the posterior surfaces of the adjacent vertebral bodies (Fig. 2).

Specimen Measurements

After the MRI examinations were completed, fluoroscopy was used to place the specimens in a rectangular plastic mold that was filled with fast-setting plastic liquid, so that the sides of the rectangular mold were coplanar with the primary anatomic planes. Using a band saw with a thin, sharp blade, the frozen blocks containing the specimens were sectioned along a plane that was parallel to the longitudinal plane that passed immediately to one side of the preintroduced tubes. A second cut was made to obtain a 5-mm-thick slab that included those tubes. The foraminal region of the slab was measured using an electronic caliper (precise to 0.1 mm).[8,9] The following parameters were measured and defined as above: foraminal height, superior foraminal width, middle foraminal width (Fig. 2).

One of the authors graded the foramina, using the same Wildermuth et al's grading system as the observers did previously when assessing images. These measurements served as the gold standard to assess the qualitative observations, reported by the observers who reviewed the same specimens' MRI scans. They also served to assess the measurements carried out on the MRI images (Fig. 6).

Statistical Analysis

The agreement between observers (agreement beyond chance) was measured using the κ coefficient statistic. Multirater, multicategory κ statistics were calculated, using a commercial software package (Stata Version 5.0, College Station, TX). Landis and Koch[10] guidelines for the interpretation of the strength of agreement for the κ statistic were used: 0.81 to 1.00 (almost perfect), 0.61 to 0.80 (substantial), 0.41 to 0.60 (moderate), 0.21 to 0.40 (fair), 0.00 to 0.20 (slight), and less than 0.00 (poor). Analysis of variance (ANOVA) was used, as an overall screening statistic, to identify the variables that were significantly associated with the grading. The best outcome would have been that the only variable that affects the grading is the specific foramen being assessed: all observers would assign the same grade no matter where the foramen was imaged. Using the data describing the degree of stenosis of each specimen, as it was physically assessed by one of the authors who examined the cadaveric 5-mm slabs, it was possible to treat the grading system by Wildermuth et al[5] as a formal diagnostic test and to calculate the normal battery of measures defining its predictive performance. These included false positive and false negative rates, sensitivity, and specificity. A false positive was defined as an observer rating that was more than 1 grade from what was observed in the physical section through the foramen. For example, an observer grade of 1 was considered the same as an actual grade of 0 or 2, but would be considered a

false negative, if the actual grade was 3. Similarly, an observer grade of 2 would be considered a false positive, if the actual grade was 0. Thus, we allowed for a modest amount of over-reading or under-reading. Sensitivity was defined as the rate of detection of patients with stenosis; specificity, as the rate of correct affirmation that a patient does not have stenosis; and accuracy, as the rate of patients' assignment to groups or grades correctly describing their degree of stenosis.

RESULTS

Part 1

Each of the 3 specimens was examined using 3 different MRI scanners. The thickness of the scans in MRI scanners 1 and 3 was 4 mm, and 5 mm in MRI scanner 2. Each one of the observers reviewed 9 sets of MRI scans, and in total assessed 72 foramina. Two hundred twenty-two foramina were classified as grade 0 (39%), 201 as grade 1 (35%), 127 as grade 2 (22%), and 26 as grade 3 (4%). On the basis of paired t tests, there was a nearly significant ($P = 0.15$) trend ,where the grade of stenosis was slightly lower when assessed using images from MRI scanner number 3 versus the other MRI scanners. The level of intraobserver agreement, when observers assessed LSFS, using images from the 3 different scanners, was assessed using κ statistics. The results were compared between scanner 1 [1.5 T (Signa, GE Medical Systems, Milwaukee, WI)] and scanner 2 [0.3 T open MRI (AIRIS II Premium, Hitachi, Tokyo, Japan)], between scanner 1 and scanner 3 [1.5 T (Magneton Symphony, Siemens, Erlangen, Germany)], and between scanner 2 and scanner 3. There was slight to fair intraobserver agreement: κ 0.20 between scanner 1 and scanner 2; κ 0.21 between scanner 1 and scanner 3; κ 0.13 between scanner 2 and scanner 3.

The level of interobserver agreement when observers assessed LSFS was slight: κ grade 0 = 0.13; grade 1 = −0.01; grade 2 = 0.06; grade 3 = 0.11; combined = 0.07. Most of the observers underestimated the grade of LSFS as compared with the grading done on the specimens.

The ANOVA tests that we used to identify the variables associated with the grade of stenosis assigned to each foramen indicated that the observer doing the grading, the place it was imaged, and the location of the foramen were all highly ($P < 0.005$) associated with the grade assigned to the foramen.

The majority of the observers (75%) indicated T1 sagittal sequence as the preferred method to assess the foramen.

Part 2

MRI Measurements

The mean foraminal height was 14.78 mm (range 23.53 to 7.06 mm, SD = 4.53). The mean proximal foraminal width was 7.73 mm (range 12 to 3 mm, SD = 1.77). The mean middle foraminal width was 3.79 mm (range 8.95 to 1 mm, SD = 1.68). The mean

posterior disc height was 5.55 mm (range 10 to 2.28 mm, SD = 1.63). The correlation between MRI measurements of the foramina and the actual dimensions of the foramina measured in the specimens was poor overall, but twice as good when carried out using the 1.5 T (Magneton Symphony, Siemens, Erlangen, Germany) scanner than the other 2 scanners (Fig. 3).

Analyzing all the MRI T1 sagittal images available to us, we identified an average of 1.19 images (range 1 to 2) that depicted an individual foramen.
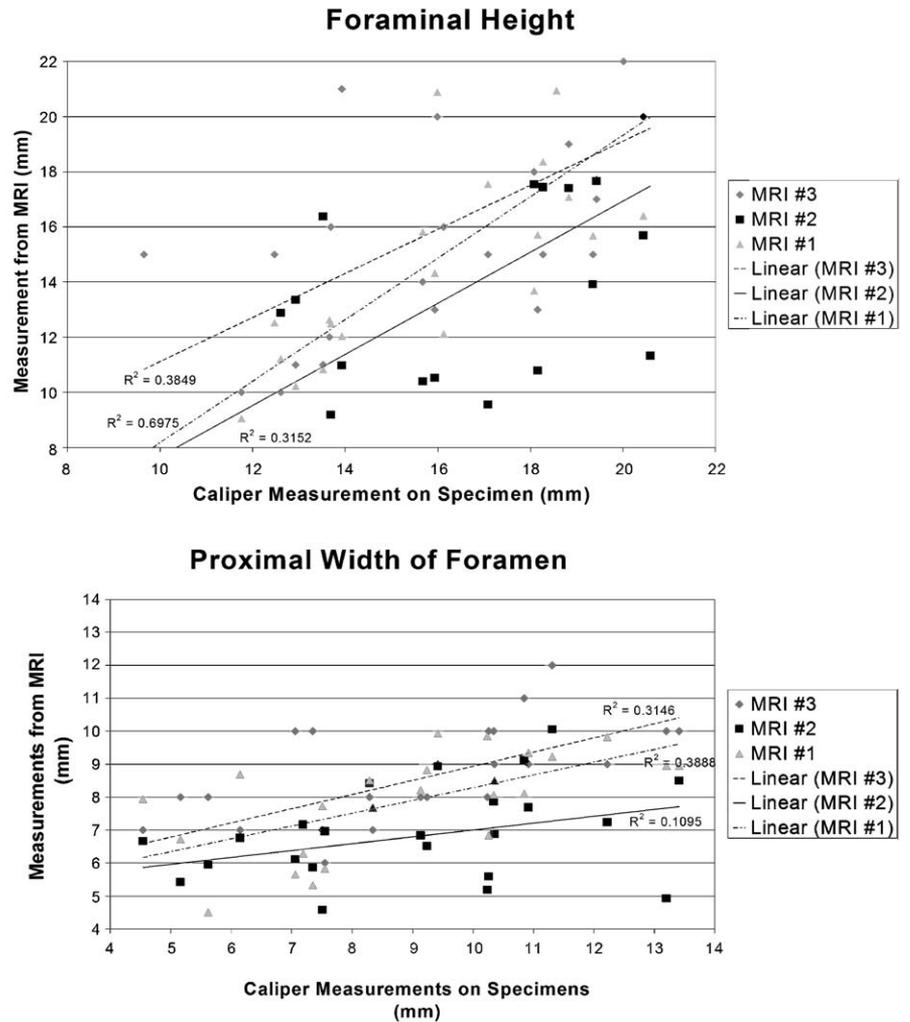
Specimen Measurements

Twenty-four foramina were measured, and graded according to the Wildermuth et al classification.[5] Eight foramina (33%) were graded as grade 0, 3 (12.5%) as grade 1, 10 (42%) as grade 2, and 3 (12.5%) as grade 3. The mean foraminal height was 16.12 mm (range 20.44 to 9.65 mm, SD = –3.11). The mean proximal foraminal width was 8.99 mm (range 13.41 to 4.54 mm, SD = − 2.43). The mean middle foraminal width was 4.76 mm (range 10.2 to 1.11 mm, SD = − 2.80). Using data describing the degree of stenosis of each specimen, as it was physically assessed by one of the authors who examined the cadaveric 5 mm slabs, it was possible to treat the grading system by Wildermuth et al[5] as a formal diagnostic test, and to calculate the normal battery of measures, defining its predictive performance: the rate of false-positive values, false-negative values, sensitivity, specificity, and accuracy are depicted in Table 2. Many of these values were low.

DISCUSSION

Jenis and An[1] defined the lumbar intervertebral foramen as a vertical interpedicular zone incorporating portions of the lateral recess and exit zone as described by Lee et al.[7] We chose to use their definition because we believe that it is the most logical one to use, assuming that MRI is the main modality used, to diagnose lumbar spine pathologies.

Wildermuth et al[5] studied the effect of flexion, and extension positions, on the size of the lumbar foramina, using an "open MRI". After the authors concluded that they could not measure accurately the size of the lumbar foramina using images acquired in those positions, they devised a qualitative grading system to grade the severity of foraminal stenosis. They did not specify the rational behind their grading system, and it is yet to be validated in clinical studies. However, to the best of our knowledge, this is the only grading system published in the English literature that classified the severity of lumbar foraminal stenosis using MRI as the main diagnostic tool. Using this classification system for grading LSFS, the results of our study demonstrated a slight interobserver agreement. Furthermore, comparing the data describing the degree of stenosis in the foramina of each specimen, as it was physically assessed by one of the authors, and the data obtained from the MRI images, and calculating the normal battery of measures defining predictive

Foraminal Height

Proximal Width of Foramen

FIGURE 3. Comparison of the measurements carried out on the cadavers, and on the MRI images. A, Foraminal height measurements; B, superior foraminal width measurements.

performance, the classification was found to have a low specificity and sensitivity. These results are in support of the results of Speciale et al[4] who studied observers' variability in the assessment of lumbar spinal stenosis on MRI, using the Glenn[3] classification, and found a low level of agreement between the observers, especially if they were orthopedic surgeons ($\kappa = 0.15$). However, these results are in contrast to those of Wildermuth et al[5] who found a good interobserver agreement ($\kappa = 0.62$) between 2 observers who graded foraminal stenosis using their classification. However, Wildermuth et al[5] calculated only those foramina considered to be sufficiently visible (39 to

49 out of 300 were not) by both observers. In the present study, no distinction was made between good or bad images, and all images were graded. We had 8 observers blinded to the specimens and to the radiology center; and we did not try to influence the technologist to obtain the best image. Our observers were exclusively surgeons who, on the basis of the study by Speciale et al,[4] have the lowest κ score. This setup was devised to simulate a real life situation, where most surgeons have no control over the quality of the images they receive, or the way they are acquired by the various MRI centers; yet, they are the ones who have to make operational decisions based,

TABLE 2. Using Data Describing the True Degree of Stenosis of Each Specimen, the Wildermuth's MRI Grading System was Treated as a Formal Diagnostic Test, and the Normal Battery of Measures Defining its Predictive Performance Were Calculated

Purpose of Exam	False Positive (%)	False Negative (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
Stenosis detection	25	54	69	54	64
Stenosis grading	25	54	69	54	35
Detection of grades 1, 2	28	48	35	84	57
Detection of grade 3	85	12	6	96	84

among other things, on those images. We think that if the same rules would have been applied by Wildermuth et al,[5] their results might have been different.

One could argue that these findings were the result of the surgeons classifying most of the foramina as having little or no stenosis. This could effect all correlations, because the data were heavily weighted toward the low grades of stenosis. However, this distribution of stenosis was similar to the distribution found by Wildermuth et al[5] using the same classification. The majority of foramina in their study was classified as grades 0 and 1 (0—53%, 1—34%).

The κ values were calculated using a 4-grade scale, using Stata, a commercial software package. The κ statistics we provided described the level of agreement between observers for the classification of stenosis in each of the 4 different levels. The interpretation of these values was that, in general, there was slight agreement between observers for any of the 4 grades, although there was slightly better agreement in classifying stenosis as grades 0 or 3 compared with classifying stenosis as grades 1 and 2. If we calculated κ statistics for each individual observer, we would have to compare observer 1 with 2, 1 with 3, 5 with 8, etc. This would results in 64 κ values and these results would have limited external validity.

ANOVA was used as an overall screening statistic to identify the variables that were significantly associated with the grading. The best outcome would have been that the only variable that affects the grading is the specific foramen being assessed: all observers would assign the same grade, no matter where the foramen was imaged. We found that the variables associated with the grade of stenosis assigned to each foramen were (a) the observer doing the grading, (b) the place it was imaged, and (c) the location of the foramen. This observation raised the question of how each observer defined stenosis in his or her mind. If an observer was focused on the width, he/she might have obtained a different grade, compared with an observer who was primarily focused on the height. Unfortunately, the present study did not provide us with an answer to this question. Furthermore, a more rigid classification of foraminal stenosis might be needed to standardize the way surgeons and radiologists classify this disease.

To the best of our knowledge, there is no direct comparison in the English literature of lumbar spine foramina measurements carried out on cadavers with those carried out using MRI images. Tien et al[9] measured the neural foramina diameters on thin, sagittal fast gradient-echo and spin-echo images, through the neural foramina of a fresh human cervical spine specimen; and compared the results with direct measurements carried out with the aid of calipers, on the specimen at the corresponding locations. Their results indicated that the fast gradient-echo images showed more apparent narrowing than the spin-echo images. In all cases, the neural foramina seemed smaller on the MRI images than in the specimen. Furthermore, the absolute distortion of 7 neural foramina was constant on the fast gradient-echo
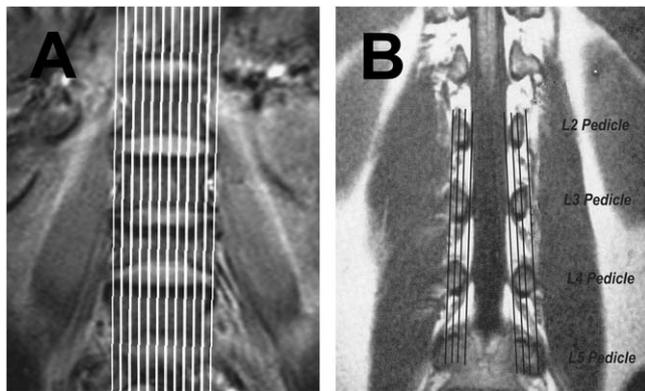
images; therefore, the relative distortion was inversely proportional to the size of the neural foramen, ranging from 9.5% to 25% in the upper cervical region. They concluded that artifactual accentuation of spinal canal narrowing and foraminal stenosis, especially in the upper cervical region, might result in false interpretations of MR images. Others authors,[11,12] studying MRI of the cervical spine, found variability in the results of MRI measurements, when compared with those of CT myelography or agar-saline spine phantom. Yousem et al[11] found, in 52 out of 90 patients, that the AP diameter of the cervical cord, as measured on MRI, was less than that measured on CT, with a difference as large as 2.3 mm. The present study results are in agreement with Tien et al[9] results. There were differences between the measurements carried out on MRI images, using the manufacturer's suggested software, and the actual measurement carried out in the cadavers. We are not sure why it is so, and the study does not provide us with a clear answer to that question. One possible reason for this discrepancy could be the fact that we used cadavers and not actual patients. This argument was addressed in another study[13] and the authors concluded that the use of cadavers is appropriate in this type of study. Furthermore, it is difficult to compare the results of foraminal measurements carried out on MRI and on the corresponding patients undergoing surgery, because the foramina would not be visualized or measured in surgery in the way they are visualized and measured in sagittal MRI slices.

Another reason could be that the sagittal cuts in the cadavers were not identical to the sagittal slice obtained with the MRI. We suggest the following explanation for this observation. In the present study, the majority of the foramina was represented by only 1 sagittal slice (average 1.19, range 1 to 2) in the MRI images. We found that the protocols presently used, by the institutions sampled by us, are based on sagittal slices reproduced from a 90° grid of lines, 4 or 5 mm apart, drawn on 1 coronal section (Fig. 5A). As the pedicle width in the lumbar spine has been reported to be in the range of 7 to 13 mm, and as the lumbar spine resembles more the shape of a cone than the shape of a tube, it is easy to understand why the foramen would be depicted by only 1 slice that would randomly represent different parts of a foramen. Because our sagittal cuts in the cadavers were planed to pass through the center of the foramen, bisecting the pedicles, and executed using fluoroscopy guidance, we assume that our cuts represented the actual dimensions of the foramen, in contrast to the sagittal cuts carried out by the MRI, that could randomly represent the smallest, or the largest part of a foramen. This explanation suggests that a clinician using MRI might not receive enough information about the shape and size of a foramen. We think that to be able to assess correctly the grade of stenosis in a foramen, assuming that the place that the nerve root is maximally compressed is where the foramen is narrowest, as was advocated by Hasegawa et al,[14] the clinician should be provided with the maximum information available (Figs. 4–6).
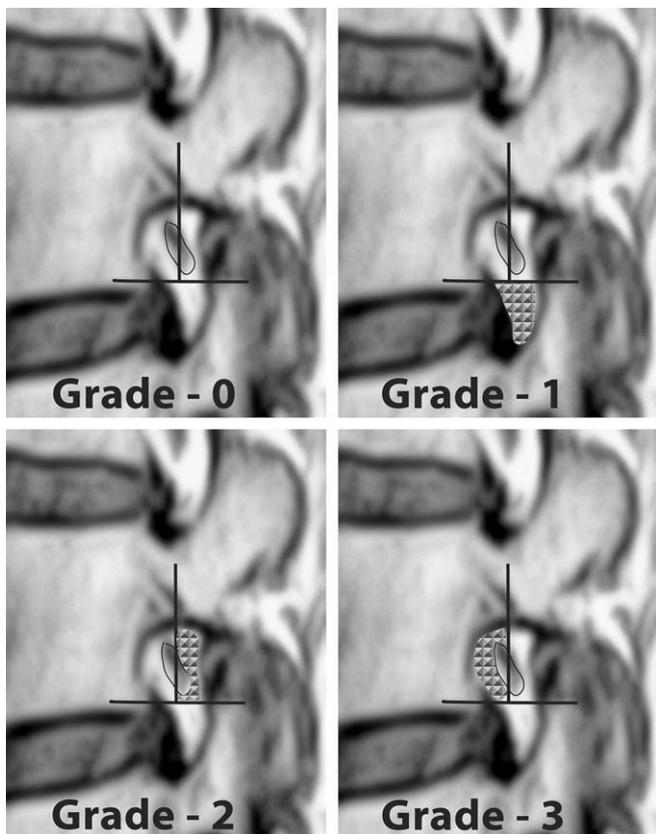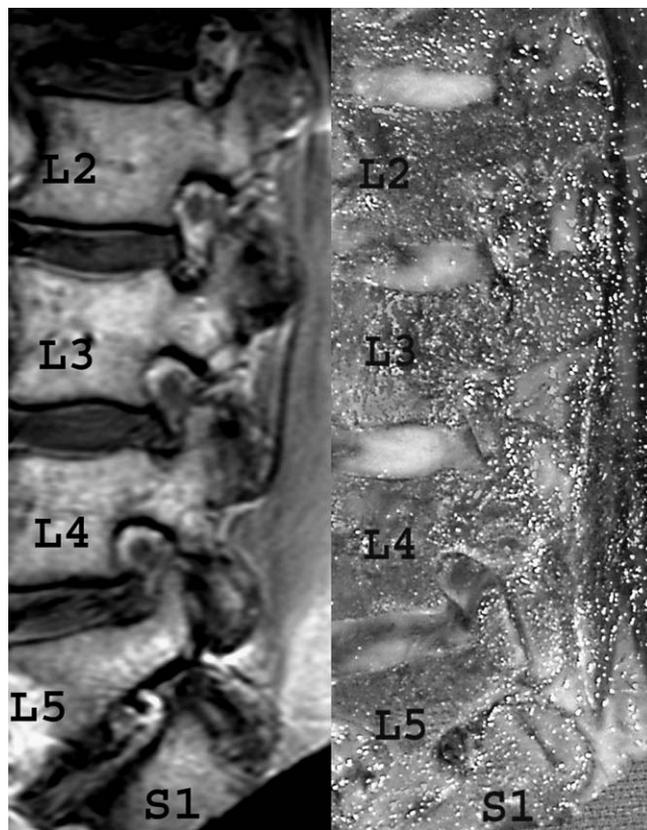
FIGURE 4. New classification system for grading foraminal stenosis, based on T1-weighted MRI. A, Grade 0; B, grade 1; C, grade 2; D, grade 3.

FIGURE 6. An example of an MRI sagittal scan of one of the specimens and the corresponding macroscopic cadaveric slab.

We think that 1 sagittal cut is not enough to provide this information; hence, we are suggesting a protocol that changes the way the sagittal sections are made, to reflect, as much as possible, the true anatomy of the intervertebral foramen. We suggest using a T1 coronal section, splitting the frame in 2, and guiding the sagittal sections parallel to the pedicle as suggested in Figure 5B. As the

pedicle width is in the range of 7 to 13 mm, we think that a thinner cut of 3 mm is necessary to get more sagittal cuts through the foramen, even though some of the clarity will be lost. We think that standardizing the way MRIs are carried out, the same way that mammography was standardized, is important, to help the clinician make the right diagnosis and plan the right treatment. Furthermore, it will assist greatly in MRI-based research.

One could suggest that a CT scan with sagittal reformatted images might be more precise in detecting bony foraminal stenosis. However, as surgeons are interested in bony and soft tissue narrowing of the foramina, the CT might not be optimal to achieve this goal. Furthermore, Beers et al,[15] studying the results of lumbar spine foraminal measurements carried out by 5 experienced radiologists on sagittaly reformatted images of 10 patients' CT scans, found statistically significant differences, that, in a number of cases, could have led to disagreement over whether stenosis was present. Preliminary data of an unpublished study, which analyzed the data obtained from 100 MRIs of the lumbar spine, performed in our institution, using the above-suggested guidelines indicated that approximately 50% more images of foramina could be obtained, when compared with the amount of foramina seen in the images obtained using the routine sagittal protocol.[16]

FIGURE 5. Coronal templates to prepare sagittal cuts in MRI. A, The usual way a technician will prepare sagittal cuts; B, the proposed way to prepare sagittal cuts.

Considering all the above results and the recommendations of Speciale et al,[4] we also suggest a new classification system for foraminal stenosis. We based the classification on sagittal T1 MRI images (Fig. 4) that depict well the fat surrounding the nerve root. The foramen can be divided into 4 quadrants by a vertical line that bisects the nerve root and foramen, and by a horizontal line that passes approximately 1 mm below the nerve root and bisects the foramen: *Grade 0* is defined as a normal foramen similar to description by Wildermuth et al[5] and Stephens et al[17]; *Grade 1* is defined as obliteration of the fat in posterior and/or anterior lower quadrant; *Grade 2* is defined as obliteration of the fat in the posterior upper quadrant; and *Grade 3* is defined as obliteration of the fat in the anterior upper quadrant. The rationale behind this classification is that the lower quadrant of the foramen, whatever the shape of the foramen, is always below the nerve root in this classification, hence being less likely to cause clinical symptoms. The 2 upper quadrants are divided equally into 2, grade 2 being pressure on the nerve root from posterior structures and grade 3 from anterior structures, a situation we think is more difficult to treat surgically as most surgeons use the posterior approach for decompression of the spine. We believe that this classification can give a common tool for reporting foraminal stenosis that avoids a possible bias by observers that might focus solely on the width, or on the height of the foramen, to assess a foramen.

There are, in our opinion, various advantages and disadvantages in using cadavers in the present study. The advantages are as follows: (1) The same specimen could be used in 3 MRI facilities and, theoretically, any change in the results could be attributed to the change in location, as no other variable was changed. (2) It was easy to mobilize a cadaver specimen between 3 facilities. (3) The same tools used to study the MRI images could be used to study the cadaveric 5-mm thick slabs, enabling a direct comparison between the two.

The disadvantages are as follows: (1) Only 3 cadavers were used in this study. More cadavers would have given our measurement much more statistical power, but because the κ values were so low we think the results are important. (2) Using the same cut in the cadaveric foramen and in MRI could theoretically prove to be more precise in comparing the results of the 2, but it would have introduced a variable, different from the real life scenario, we wanted to simulate. (3) The Wildermuth et al[5] classification was not validated in clinical studies. (4) There were 3 variables associated with the grade of stenosis assigned to each foramen, and this fact made it difficult to suggest specific recommendations, to standardize the way MRIs are performed. More studies are needed to address the variability between different institutions, and the optimal MRI sequences needed to examine the lumbar foramen.

In summary, in this cadaveric model, the Wildermuth classification system was found to have a poor interobserver reliability. There were significant differences between foraminal measurements carried out on MRI and on the cadavers. The parameters associated with the grade of stenosis assigned to the foramen were (1) the observer doing the grading, (2) the place it was imaged, and (3) the location of the foramen.

REFERENCES

1. Jenis L, An H. Spine update: lumbar foraminal stenosis. *Spine*. 2000;25:389–394.
2. Czervionke LF, Berquist TH. Musculoskeletal imaging update, part I. *Radiol Clin North Am*. 1997;28:583–616.
3. Glen WV. Tomography multiplaner reformatted (CT/MRI) examination of the lumbar spine. In: Genant H, Chafetz N, Helms C, eds. *Computed Tomography of the Lumbar Spine*. Emeriville, California: University of California Printing Department; 1982;87.
4. 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;10:1082–1086.
5. Wildermuth S, Zanetti M, Duewell S, et al. Lumbar spine: quantitative and qualitative assessment of positional (upright flexion and extension) MR imaging and myelography. *Radiology*. 1998;207: 391–398.
6. Delamarter RB, McCulloch JA. Microdiscectomy and microsurgical spinal laminotomies. In: Frymoyer JW, ed. *The Adult Spine: Principles and Practice*. Philadelphia: Lippincot-Raven Publishers; 1997;1961–1988.
7. Lee C, Rauschning W, Glenn W. Lateral lumbar spinal canal stenosis: classification, pathologic anatomy, surgical decompression. *Spine*. 1988;13:313–320.
8. Kandziora F, Schulze-Stahl N, Khodadadyan-Klostermann C, et al. Screw placement in transoral atlantoaxial plate systems: an anatomical study. *J Neurosurg*. 2001;95(Suppl 1):80–87.
9. Tien RD, Buxton RB, Schwaighofer BW, et al. Quantitation of structural distortion of the cervical neural foramina in gradient-echo MR imaging. *J Magn Reson Imaging*. 1991;1:683–687.
10. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
11. Yousem DM, Janick PA, Atlas SW, et al. Pseudoatrophy of the cervical portion of the spinal cord on MR images: a manifestation of the truncation artifact? *AJNR Am J Neuroradiol*. 1990;11:373–377.
12. Ros L, Mota J, Guedea A, et al. Quantitative measurements of the spinal cord and canal by MR imaging and myelography. *Eur Radiol*. 1998;8:966–970.
13. Nowicki B, Haughton V. Neural foraminal ligaments of the lumbar spine: appearance at CT and MR imaging. *Radiology*. 1992;183: 257–264.
14. Hasegawa T, An H, Haughton V, et al. Lumbar foraminal stenosis: critical heights of the intervertebral discs and foramina. *J Bone Jt Surg [Am]*. 1995;77:32–38.
15. Beers GJ, Carter AP, Leiter BE, et al. Interobserver discrepancies in distance measurements from lumbar spine CT scans. *AJR Am J Roentgenol*. 1985;144:395–398.
16. Attias N, Crowder T, Connell M, et al. Assessment of a new MRI protocol for the study of the lumbar foramina. Unpublished study.
17. Stephens M, Evans J, O'Brien J. Lumbar intervertebral foramens: an in vitro study of their shape in relation to intervertebral disc pathology. *Spine*. 1991;16:525–529.