Inter- and Intraobserver Reliability in Radiographic Assessment of Degenerative Disk Disease

JASON ZOOK, MD; MLADEN DJURASOVIC, MD; CHARLES CRAWFORD III, MD; KELLY BRATCHER, RN; STEVEN GLASSMAN, MD; LEAH CARREON, MD, MSc

abstract

Clinicians use descriptive classification systems when treating patients with low back pain as an adjunct to surgical decision making. Magnetic resonance imaging (MRI) changes, including Modic changes, the presence of a high-intensity zone, and internal disk desiccation, are commonly used descriptors. The question remains whether different clinicians interpret these terms similarly. This study evaluated the inter- and intraobserver reliability of commonly used MRI classifications in patients presenting with low back pain.

Sixty-six patients who underwent lumbar spine fusion surgery at a single multiphysician spine specialty practice for degenerative disk disease were identified. For each surgical level, the following MRI variables were determined independently by 3 fellowship-trained spine surgeons: presence or absence of high-intensity zone and/or internal disk desiccation, presence and classification of disk herniation, Modic grade, and disk height. Each surgeon reviewed the same set of MRI studies a second time at least 2 weeks from the first reading. Inter- and intraobserver reliability was determined using multiobserver Kappa coefficients. Intraobserver reliability ranged from 0.563 to 0.988, with greatest agreement in determining disk height. The greatest interobserver agreement was for determining Modic changes (0.819).

Controversy remains on the criteria for diagnosing degenerative disk disease. In patients presenting with low back pain diagnosed with degenerative disk disease, the inter- and intraobserver reliability with use of several common MRI diagnostic tools was substantial. These data imply that clinicians interpret these findings in a reproducible fashion and interpret these terms similarly.
Low back pain is the second most common reason cited for physician visits, and almost 1 in 4 adults experienced back pain in the past 3 months.\textsuperscript{1} Based on patient symptoms, clinical examination, and radiographic imaging, some of these patients will be diagnosed with degenerative disk disease.

Several radiographic changes have been identified on magnetic resonance imaging (MRI) and are postulated to be related to disk degeneration. These include Modic changes,\textsuperscript{2,3} loss of disk height,\textsuperscript{4,5} presence of annular tears or high-intensity zones,\textsuperscript{6} disk protrusions and herniations, and internal disk desiccation.\textsuperscript{7} Although several authors have shown that Modic I changes are associated with low back pain,\textsuperscript{8-10} other authors have shown no association between the presence of Modic changes and pain provocation during diskography.\textsuperscript{11,12} Contrasting evidence exists that Modic I changes are\textsuperscript{13} or are not\textsuperscript{14} associated with segmental instability. Similarly, the value of the presence of a high-intensity zone on MRI is unclear, with some studies showing that it is a reliable marker for outer annular disruption\textsuperscript{15} and others showing that it is present in asymptomatic individuals as well.\textsuperscript{16,17} There is also scant evidence on the use of loss of disk height as a predictor of outcomes after surgery in patients with low back pain.\textsuperscript{5,18}

Among clinicians, these radiographic markers are interpreted differently and may be applied to clinical decision making inconsistently. Although several studies have previously looked at the reproducibility of some of these descriptors in small groups and individually,\textsuperscript{18-22} to our knowledge none have compared all of these multiple diagnostic criteria simultaneously. Inconsistency exists in previous studies as to the amount of preparation and education reviewers received, potentially artificially elevating reliability results.\textsuperscript{23} Due to the ubiquitous nature of low back pain, there exists a need for an efficient, reproducible means of objectively identifying and accurately reproducing radiographic descriptors of disk degeneration. The goal of this study was to evaluate the reliability and reproducibility of several commonly used radiographic descriptors used in the diagnosis of degenerative disk disease in the lumbar spine.

**Materials and Methods**

A series of 71 patients who underwent lumbar fusion at a single multiphysician spine specialty practice between 2002 and 2006 were identified. The preoperative diagnosis for all patients was degenerative disk disease with or without radiculopathy. All patients had complete preoperative MRI studies (both T1- and T2-weighted images in multiple planes) available for review, either via in-house imaging system or on computer disks from outside facilities. Three fellowship-trained spine surgeons (M.D., C.C., S.G.) reviewed the imaging on the following characteristics: internal disk desiccation, disk height, annular tear, high-intensity zone, disk protrusion and herniation, and Modic changes. Definitions for these criteria were based on those previously provided by a combined task force\textsuperscript{24} and provided to all 3 evaluators. In an effort to reproduce clinical practice, no formal instruction or training was provided to the evaluators, nor was there any pre-test discussion between them regarding definitions and use of terminology.

For internal disk desiccation, annular tear, high-intensity zone, and disk herniation, the evaluator simply identified the presence or absence of such characteristics. For disk herniation, they were asked to classify (if applicable) the herniation as protruded, extruded, or sequestered. In the case of disk height, evaluators were asked to measure the height of treated disk levels at the highest point mid-disk on the T2 image as well as the immediate-adjacent nonoperative disk based on a midsagittal MRI slice. The evaluators were not instructed as to which particular image to use, to better simulate true clinical situations. Modic’s classification system was used to classify degenerative changes, if present. Characterization as type 1, 2, or 3 was based on accepted descriptors.\textsuperscript{4,5} Measurements were performed independently of the other investigators. The measurements were taken on 2 separate occasions, spaced at least 2 weeks apart. Inter- and intraobserver reliability was calculated using the multirater kappa statistic (with corresponding 95% confidence interval) based on intraclass correlations.\textsuperscript{25}

**Results**

Analysis of data demonstrated almost universally good kappa levels for inter- and interobserver reliability. Interobserver reliability ranged from 0.385 to 0.988 (Table 1), and interobserver levels from 0.510 to 0.922 (Table 2). The highest interobserver reliability was disk desiccation (0.922) and the lowest was disk disruption (0.510). Only 1 category in 1 observer had an intraobserver reliability <0.600 (disk disruption; rater 1, 0.385). Otherwise, all other intraobserver ratings were >0.600, some has high as 0.988.

**Discussion**

Controversy exists as to the appropriate clinical indications for surgical treatment of diskogenic back pain. One reason for this may be a lack of consistently used criteria for diagnosing degenerative disk disease. Another may be the presence of degenerative changes in many asymptomatic individuals.\textsuperscript{26} Although surgical decision making is ultimately based on clinical findings, incorporating and using radiographic selection criteria that are universally understood and applied among clinicians may serve as a basis for the development of more reliable treatment paradigm in surgical decision making.

This study attempted to evaluate various radiographic morphologic characteristics used currently in the diagnosis of degenerative disk disease in a clinical subset of patients who had already undergone surgical fusion. The findings of good to excellent reliability in interpretation of
lumbar MRI have been reported previously, but to our knowledge no previous study has attempted to look at multiple variables simultaneously in a surgically treated population. Furthermore, by removing pre-test education of evaluators, we hoped to more reliably reproduce a “real-life” clinical scenario.

In all categories of measurement, there was at least moderate agreement, and in 2 categories almost perfect agreement, which is higher than previous studies. Carrino et al reported good interobserver agreement for disk degeneration (0.55), Modic changes (0.59), and posterior or high-intensity zones (0.44). For Modic changes, they found a kappa of 0.43 for intraobserver agreement; good and moderate agreement was found. This is despite training with sample images and an in-person meeting to review and refine the groups’ definitions. Additionally, they did not look specifically at disk height.

Mulconrey et al achieved good agreement looking at disk degeneration (0.773) with 4 observers evaluating 17 patients (80 levels). Jones et al read 50 MRIs (1 level per image) and found interobserver agreement to be very good (0.85) and intraobserver agreement to be good or very good (0.71-1.00). However, only 1 interobserver agreement value was reported, and there was no mention of pair-wise comparisons. Smith et al have previously evaluated the diagnosis of high-intensity zones. Two readers evaluated 175 MRIs. No repeat readings were performed. Moderate interobserver agreement for high-intensity zones (0.57) was reported.

To our knowledge, ours is the only study to assess disk height in this subset of surgically treated patients. The intraobserver reliability demonstrated high numbers with all 3 clinicians. Interestingly, the interobserver reliability was not as high (0.584). This is likely secondary to 2 factors: (1) no formal instruction was given to providers, which meant that they were free to use any means/technique to measure height; and (2) unlike a yes/no dichotomous response in the categories of disk desiccation and high-intensity zones, disk height is a continuous variable, making achieving substantial interobserver reliability more difficult. As the intraobserver kappa values were high, the implication is that no matter what technique is used (ruler, digital ruler, or digital radiography system), a given provider can reliably use disk height as a measurement tool. Further study will look into disk height as it correlates with clinical outcome in this same group of patients.

A relative strength of our study is the sole use of treating clinicians and providing them with pre-test education. The thought is that this provides a more realistic real-world scenario that can be applied to clinical practice in a nonacademic setting.

Our study has limitations. There was no consistency in terms of imaging source/quality. Although several patients had single-tau inversion recovery sequence imaging, which can help rule out infection, not all patients had these available, and they could not be evaluated uniformly. Patients frequently had MRIs performed at the practice’s hospital-based MRI, but also presented with MRIs (print and/or disk) from outside hospitals. It has been reported previously that the strength of the MRI magnet affects observer assessment of imaging, but it has also been shown recently that even with open 2-T magnets, acceptable reliability can be achieved. Moreover, we feel that the variability of image quality more closely resembles a typical clinical practice, and therefore the kappa values obtained with these give more relative strength to the findings. The size of the canal was not evaluated. A protruded disk might compress a root, whereas a patient with large canal an extruded disk might be asymptomatic. Thus, nerve root compression, displacement, and distortion might correspond to the clinically most useful information.

Another weakness of the study is the inclusion of only 71 surgically treated levels. This represents a 4-year period of surgery at a busy 6-surgeon practice. Practically, to obtain a larger cohort, a multicenter study would be more pragmatic. In addition, because this is a retrospective review on a surgically treated population, there is an inherent selection bias toward more impressive radiographic findings.
Although only 71 treated levels were evaluated, all patients were repeat evaluated, which is in contrast to previous studies where intraobserver follow-up was <100%. Our study group had only 3 evaluators, all from the same institution, which has been identified previously as a potential weakness. We acknowledge that having several evaluators would add strength to interobserver reliability findings, and may have artificially elevated our kappa values. Other studies have included musculoskeletal and/or neuroradiologists. Although this would provide an additional level of expertise to the study, we feel that it would detract from the applicability of the results to those practitioners who do not have access to subspecialized radiologists. Also, we used dichotomous descriptors (presence of absence of internal disk desiccation, for example) versus a continuous variable (disk height). This may have contributed to our high kappa values as well.

Ultimately, a descriptor system’s strength is limited by its reproducibility. Additionally, it should correlate with a pathologic disease state and ideally correlate with intervention-based outcomes. It appears that several radiologic criteria are reproducible. Now that statistical agreement in various descriptors has been established, we hope to expound further on this data and look at predictors of clinical outcome in this same cohort.

REFERENCES

29. Jensen TS, Sorensen JS, Kjaer P. Intra- and interobserver reproducibility of vertebral


