Clinical Study

Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions

Michael C. Fu, BS[a], Rafael A. Buerba, BA[a], William D. Long, III, MD[a],
Daniel J. Blizzard, MD, MHS[b], Andrew W. Lischuk, MD[c], Andrew H. Haims, MD[c],
Jonathan N. Grauer, MD[a],*

[a]Department of Orthopaedics and Rehabilitation, Yale University School of Medicine, 800 Howard Ave., 1st Floor, New Haven, CT 06519, USA
[b]Department of Orthopaedic Surgery, Duke University School of Medicine, 200 Trent Drive, Box 3000, Durham, NC 27710, USA
[c]Department of Diagnostic Radiology, Yale University School of Medicine, PO Box 208042, New Haven, CT 06520, USA

Received 9 September 2013; revised 18 February 2014; accepted 4 March 2014

Abstract

BACKGROUND CONTEXT: Magnetic resonance imaging (MRI) is frequently used in the evaluation of degenerative conditions in the lumbar spine. The relative interrater and intrarater agreements of MRI findings across different pathologic conditions are underexplored, as most studies are focused on specific findings.

PURPOSE: The purpose of this study was to characterize the interrater and intrarater agreements of MRI findings used to assess the degenerative lumbar spine.

STUDY DESIGN: A retrospective diagnostic study at a large academic medical center was undertaken with a panel of orthopedic surgeons and musculoskeletal radiologists to assess lumbar MRIs using standardized criteria.

PATIENT SAMPLE: Seventy-five subjects who underwent routine lumbar spine MRI at our institution were included.

OUTCOME MEASURES: Each MRI study was assessed for 10 lumbar degenerative findings using standardized criteria. Lumbar vertebral levels were assessed independently, where applicable, for a total of 52 data points collected per study.

METHODS: T2-weighted axial and sagittal MRI sequences were presented in random order to the four reviewers (two orthopedic spine surgeons and two musculoskeletal radiologists) independently to determine interrater agreement. The first 10 studies were reevaluated at the end to determine intrarater agreement. Images were assessed using standardized and pilot-tested criteria to assess disc degeneration, stenosis, and other degenerative changes. Interrater and intrarater absolute percent agreements were calculated. To highlight the most clinically important MRI disagreements, a modified agreement analysis was also performed (in which disagreements between the lowest two severity grades for applicable conditions were ignored). Fleiss kappa coefficients for interrater agreement were determined.

RESULTS: The overall absolute and modified interrater agreements were 76.9% and 93.5%, respectively. The absolute and modified intrarater agreements were 81.3% and 92.7%, respectively. Average Fleiss kappa coefficient was 0.431, suggesting moderate overall agreement. However, when stratified by condition, absolute interrater agreement ranged from 65.1% to 92.0%. Disc hydration, disc space height, and bone marrow changes exhibited the lowest absolute interrater

---

agreements. The absolute intrarater agreement had a narrower range, from 74.5% to 91.5%. Fleiss kappa coefficients ranged from fair-to-substantial agreement (0.282–0.618).

CONCLUSIONS: Even in a study using standardized evaluation criteria, there was significant variability in the interrater and intrarater agreements of MRI in assessing different degenerative conditions of the lumbar spine. Clinicians should be aware of the condition-specific diagnostic limitations of MRI interpretation. © 2014 Elsevier Inc. All rights reserved.

Introduction

Degenerative conditions of the lumbar spine are ubiquitous in modern society [1]. Failing conservative management, magnetic resonance imaging (MRI) is a noninvasive and radiation-free imaging modality that is frequently considered for this population. Speed and image quality have continued to evolve for this imaging modality, but limitations remain.

The interpretation of MRI studies is subject to variability. This may be because of variations in the nomenclature [2,3]. Analogous to clinical medicine, there is no single-established validated grading scheme for many radiographic findings. However, there are also variations inherent to the assessment of resultant images. A study interpreted as "severe" stenosis may be read as "moderate" or perhaps "mild" by another reviewer [4]. Though much of the clinical practice of spine surgery is based on the correlation of clinical symptomatology and imaging findings, the importance of these variabilities in MRI interpretation and nomenclature cannot be ignored.

Most studies evaluating the interpretation of lumbar MRI pathologies have focused on various specific grading scales. For example, studies have examined the diagnostic characteristics of MRI with regard to conditions such as spinal cord compression in acute traumatic injury [5], disc abnormalities [6–10], end-plate signal (Modic) changes [11,12], lumbar spinal stenosis [4,13], and disc herniation [14,15]. There are several studies that have examined a handful of spinal conditions simultaneously [16–18].

Considering the reported variability in assessing specific lumbar conditions by MRI, it can be expected that this variation would exist between different pathologies in a standardized comparison. Nonetheless, we believe physicians and patients may underappreciate these inherent variabilities in MRI interpretation despite the widespread use of this imaging modality [4,16,18]. The purpose of our study was to examine the interrater and intrarater agreements of MRI in the evaluation of 10 degenerative conditions of the lumbar spine, with a panel of orthopedic spine surgeons and musculoskeletal radiologists.

Methods

Patient sample

The patient population for this study was drawn from our institution's radiology database of patients who underwent lumbar spine MRI in 2010 by our Department of Musculoskeletal Radiology. Exclusion criteria included prior lumbar instrumentation or fusion. There were no changes in imaging equipment or technique over the study period. The patients were sorted in chronological order based on the imaging study date, and the first 75 patients were included in our study based on a priori power calculations, a perceived clinically relevant difference of 15% in interrater or intrarater agreement, an alpha error level of 5%, and a statistical power of 80%. Approval was obtained from our institution's Human Investigations Committee.

MRI criteria and assessment

For each subject, sagittal and axial T2-weighted sequences were evaluated based on specific criteria for stenosis, herniation, and degenerative changes. Disc anatomy and pathology were assessed from the L1/L2 to L5/S1 disc spaces. Vertebral changes were assessed from the L1 to L5 vertebrae. Bone marrow changes (at any level) and the absence or presence of transitional vertebrae were also noted. In total, 52 data points were collected per study (10 conditions were assessed at each of 5 vertebral levels and 2 conditions assessed on an overall level). The specific pathologies assessed with the corresponding criteria and grades of severity are listed in Table 1.

The criteria and severity grades were developed and pilot tested by the authors of the study, based on clinical experience and similar studies in the literature that used graded assessment scales [18], with the purpose of examining the diagnostic performance of MRI in the lumbar spine. All imaging studies were obtained on one of the three Siemens (Siemens Medical Solutions USA, Inc., Malvern, PA, USA) MRI scanners: Verio (3.0 T), Avanto (1.5 T), or Esprit (1.5 T). The images were reviewed with Synapse digital radiography software version 3.2.1 (Fujifilm, Tokyo, Japan).

Sagittal and axial T2-weighted sequences for each subject were evaluated independently by four reviewers (two orthopedic spine surgeons and two musculoskeletal radiologists). One of the orthopedic spine surgeons was a fellowship-trained attending surgeon with approximately 10 years of experience, whereas the other was a spine surgery fellow with several years of prior experience as an attending. The two musculoskeletal radiologists were both fellowship-trained attending physicians with about 13 and 3 years of experience, respectively. The subjects were presented in a random order to the evaluators. The first 10