Reliability and variability in the interpretation of lumbar high intensity zone

Shun-Wu FAN¹, Xiang-Qian FANG¹, Yun-Jian LIU², He-Jun YU¹, Yin-Jiang LU³, Chao LIU¹

¹Zhejiang University Faculty of Medicine, Sir Run Run Shaw Hospital, Department of Orthopedics, Zhejiang Province, China
²Wenzhou Medical College, Lishui Central Hospital, Department of Orthopedics, Zhejiang Province, China
³Shangyu People’s Hospital, Department of Orthopedics, Zhejiang Province, China

Objective: The aim of this study was to evaluate the reliability of high intensity zone (HIZ) and to assess discrepancy in the interpretation, as well as investigate the effects of parameters of HIZ on interobserver variation.

Methods: Four spine surgeons made independent observations on lumbar magnetic resonance imaging (MRI) from 207 consecutive patients from 3 institutions. The κ statistic was used to characterize inter- and intraobserver reliability for visual assessments of HIZ. The corresponding MRI was provided to 2 additional spine surgeons for quantitative measurements. The parameters of HIZ, including signal intensity (SI) and area ratio (HIZ%), were used to assess the interobserver variation of HIZ.

Results: The overall interobserver agreement for visual assessments was substantial (κ=0.62 at L4–5 and 0.61 at L5–S1), and intraobserver agreement was excellent (κ=0.84 at L4–5 and 0.86 at L5–S1). Of 93 observed HIZ, 17 instances (18.3%) were agreed upon by all visual observers. The SI with full agreement was significantly brighter than all the others (p<0.01). The HIZ% with 2 agreements was significantly smaller than those with 4 agreements (p=0.04) and 3 agreements (p=0.03). Although fewer observers with consensus were associated with smaller HIZ%, the difference was not significant (p>0.05).

Conclusion: The reliability in the interpretation of HIZ was sufficient for spine surgeons with differing levels of experience. This study highlighted that signal intensity was the primary cause of variability in visual observation.

Keywords: High intensity zone; lumbar spine; magnetic resonance imaging; reliability; variability.

Level of Evidence: Level IV Diagnostic Study

Magnetic resonance imaging (MRI) is a widely used imaging modality in the diagnosis of intervertebral disc pathology. Reliable assessment of disc abnormalities from MRI is important to provide diagnosis, to influence therapeutic decision-making, and to study the prognostic role of imaging features. However, consensus in rating the majority of MRI findings is often difficult to achieve.[1,2] Discrepancy in interpretation can mislead clinicians, as well as reduce the usefulness of those findings.[3–5] Hence, diagnostic imaging studies should focus on not only revealing agreement but also investigating disagreement.[6]

Since lumbar high intensity zone (HIZ) on T2-weighted MRI was first described by Aprill and Bogduk...
in 1992,[7] considerable interest has surrounded this diagnostic finding. Some investigators[8–10] believed that HIZ was a valuable marker for painful and ruptured discs, though others disagreed with this conclusion.[11,12] Although the reliability of HIZ has been documented in the literature (Table 1), previous studies may have been limited by focusing on several findings[13–18] and having fewer observers.

Furthermore, little attention has been paid to the variability in the interpretation of HIZ.

In order to avoid discrepancy and improve diagnostic consistency, it is essential to identify observer variability by using appropriate methods.[4] Without precise and rigorous methods of measurement, it is difficult to clarify interobserver variation in the interpretation of HIZ. Indeed, the detection of HIZ by the naked eye is an imprecise assessment and may be a limitation for its meaningful clinical use.[20] Recently, using cerebrospinal fluid (CSF) as a reference, a series of quantitative measurements of HIZ has been established.[21] This objective method shows advantages in minimizing artifacts caused by absence of magnetic field homogeneity and allowing comparison parameters of HIZ among different images. Thus, a rigorous investigation of interobserver variation of HIZ is available.

The primary purpose of the present study was to assess the reliability in the interpretation of HIZ among 4 observers from different institutions, while the secondary aim was to disclose discrepancy between observers and investigate the effect of parameters of HIZ on interobserver variation.

Patients and methods

This prospective multicenter study was conducted in 3 institutions located in different geographic regions. As part of a larger project on the diagnostic process for patients with low back pain (LBP), the recruitment period of this study was 1 year, from June 2009 to May 2010. The protocol was approved by the ethics com-

---

**Table 1.** Inter- and intraobserver reliability in the interpretation of HIZ in the literature (κ values).

<table>
<thead>
<tr>
<th>First authors</th>
<th>Study years</th>
<th>MR images</th>
<th>Location</th>
<th>Observers</th>
<th>Age range (years)</th>
<th>Male/female ratio</th>
<th>Initial rater agreement</th>
<th>Intra-rater agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aprill/1992</td>
<td>Westsap/1998</td>
<td>67 LBP subjects</td>
<td>20–50</td>
<td>0.6 T</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Weishaupt/1998</td>
<td></td>
<td>72 LBP subjects</td>
<td>20–50</td>
<td>1.0 T</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Smith/1998</td>
<td></td>
<td>111 subjects</td>
<td>20–50</td>
<td>0.34 T</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Carrino/2009</td>
<td></td>
<td>111 subjects</td>
<td>20–50</td>
<td>0.34 T</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Zook/2011</td>
<td></td>
<td>71 surgically treated subjects</td>
<td>20–50</td>
<td>1.0 T</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Hancock/2012</td>
<td></td>
<td>170 LBP subjects</td>
<td>20–50</td>
<td>1.0 T</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Berg/2012</td>
<td></td>
<td>554 subjects</td>
<td>20–50</td>
<td>1.0 T</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>

HIZ: High intensity zone; CI: Confidence interval; LBP: Low back pain; AF: Annular fibrosis; NR: No report; SPORT: spine patient outcomes research trial. *The intrarater agreement was based on 40 images. **The interrater agreement was based on 126 images.
Patients were recruited from orthopedic outpatient departments during the survey period. Patients who had at least 6 months' duration of LBP with or without radiculopathy were included in the study. Those with previous spine surgery, ongoing psychiatric illness, aged <18 years or ≥70 years, or who were pregnant were excluded. Additionally, based on MRI findings, cases were excluded if they had spinal fracture, scoliosis with >15° curvature, cauda equina syndrome, infection, or neoplasm. Written informed consent was obtained from all patients.

Lumbar spine MRI was acquired from the 3 participating institutions. The MRI systems included 2 Sonata units (1.5 T, Siemens, Erlangen, Germany) and 1 Signa HDx unit (1.5 T, General Electronic, Milwaukee, WI, USA). As a result of using different MRI systems, a variety of imaging techniques were performed (TR/TE: 420–560/12–14 in T1-weighted sagittal images; TR/TE: 2300–3000/100–127 in T2-weighted sagittal and axial images; 90° flip angle, 255×512 matrix size, 260×260 mm field of view, 3.0–4.0-mm section thickness, and 0.5–1.0-mm intersection gap).

Visual assessments of HIZ were performed by 4 spine surgeons (a consultant, a locum consultant, a senior fellow, and a chief resident) from different institutions. Observer experience in reading spine MRI ranged from 5 to 27 years. Each observer independently evaluated the images without knowing name, sex, age, or other background data (e.g., where the image was obtained) of the patients. To assess intraobserver reliability of visual assessments, a random subsample of 40 images was selected and re-evaluated at least 3 months after the initial reading. The observers were not allowed to access the original readings when conducting the second evaluation.

Each image was examined and scored for the presence or absence of HIZ at separate lumbar levels. In this study, HIZ was defined as a lesion with high intensity (white) signal on T2-weighted MRI located in the posterior, posterolateral, and lateral annular fibrosus.\(^8,22\)

All data were entered into a database at a centralized coordination office. The eligible images with HIZ from each individual observer’s report were provided to 2 additional spine surgeons for quantitative measurements.

Prior to quantitative measurements commenced, ImageJ software (Version 1.43, National Institute of Health, Bethesda, MD, USA) was installed on computers. Two additional spine surgeons, who were blinded to the purpose of this study, used this software to survey the region of interest (ROI) on each eligible image. The last 10 images were re-evaluated 2 weeks later so that intraobserver agreement data of quantitative measurements could be obtained. All dimension measures were made using freehand areas (Figure 1), and full details of the protocol have been described previously.\(^21\) The parameters of ROI, including the areas of HIZ and corresponding disc, and the signal intensities of HIZ and CSF were computed automatically by ImageJ software. The CSF-adjusted signal intensity of HIZ (SICSF-HIZ) was calculated as the ratio of the signal intensity of HIZ to that of CSF. The area proportion of HIZ (HIZ%) was the ratio of the area of HIZ to that of the corresponding disc.

The reliability of visual assessments and quantitative measurements was calculated by \(\kappa\) statistics and intraclass correlation coefficient (ICC) formula,\(^3,1\) respectively. The \(\kappa\) or ICC values of 0–0.2 indicated slight agreement, 0.21–0.4 fair agreement, 0.41–0.6 moderate agreement, 0.61–0.8 substantial agreement, and 0.81–1 excellent agreement.\(^23\) Since very low or high prevalence of the events could lead to artifactual agreement, reliability analysis was performed only when HIZ was reported with prevalence between 10% and 90% across all observers at each level.\(^24\) Student’s t-test was used to make comparisons between groups for continuous variables. Fisher’s exact test or \(\chi^2\) test was used to evaluate categorical variables. Statistical analysis was performed using SPSS software (version 16.0, SPSS Inc., Chicago, IL, USA). The significance level was set at \(p<0.05\).

**Results**

During the survey period, 213 consecutive eligible patients were identified. Four patients did not undergo MRI because of claustrophobia. Two patients were ex-
cluded because their images did not have an appropriate scale for quantitative measurements. Thus, a total of 207 patients with 1035 lumbar levels were included. There were 109 females and 98 males aged from 20 to 69 years (45.6±10.8 years); 80–91 HIZ were evaluable (Table 2). Due to a mean prevalence of <10% in the sample (n=207), HIZ at upper lumbar segments (L1–2, L2–3, and L3–4) were excluded for the reliability analysis. Overall interobserver agreement of visual assessments was substantial, with $\kappa$ of 0.62 (95% confidence interval [CI] [0.56–0.69] at L4–5 and 0.61 (95% CI 0.55–0.67) at L5–S1. As there were 4 visual observers (observers A, B, C, and D), this resulted in 6 unique observer pairs: AB, AC, AD, BC, BD, and CD. Pairwise agreement was moderate to substantial (Table 3).

Forty images were randomly selected to evaluate the intraobserver reliability. The intraobserver reliability was excellent at L4–5 ($\kappa$=0.84; 95% CI 0.72–0.96) and at L5–S1 ($\kappa$=0.86; 95% CI 0.75–0.97), and it was consistent between the 4 observers (Table 3).

Inter- and intraobserver reliability for the quantitative measurements was summarized in Table 4. At both L4–5 and L5–S1, there was excellent interobserver agreement for area of HIZ, HIZ%, and SICSF-HIZ, and somewhat higher reliability for area of disc. As expected, intraobserver ICC was slightly higher than interobserver ICC.

The results of the quantitative measurements made by each of the 2 quantitative observers (observers E and F) are illustrated in Table 5. There was no statistically significant difference in readings between the 2 observers. Thus, the average measures were used in the following investigation. The area of HIZ at L5–S1 was significantly smaller than the area of HIZ at L4–5 ($p=0.04$). Other parameters of ROI—including area of disc, HIZ%, and SICSF-HIZ—were also slightly lower

### Table 2. The prevalence of HIZ reported by 4 observers in visual assessments.

<table>
<thead>
<tr>
<th></th>
<th>Observer A</th>
<th></th>
<th>Observer B</th>
<th></th>
<th>Observer C</th>
<th></th>
<th>Observer D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>L1–2</td>
<td>3</td>
<td>1.45</td>
<td>2</td>
<td>0.97</td>
<td>2</td>
<td>0.97</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td>L2–3</td>
<td>2</td>
<td>0.97</td>
<td>3</td>
<td>1.45</td>
<td>3</td>
<td>1.45</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td>L3–4</td>
<td>18</td>
<td>8.70</td>
<td>17</td>
<td>8.21</td>
<td>20</td>
<td>9.66</td>
<td>21</td>
<td>10.14</td>
</tr>
<tr>
<td>L4–5</td>
<td>36</td>
<td>17.39</td>
<td>28</td>
<td>13.53</td>
<td>32</td>
<td>15.46</td>
<td>33</td>
<td>15.94</td>
</tr>
<tr>
<td>L5–S1</td>
<td>32</td>
<td>15.46</td>
<td>30</td>
<td>14.49</td>
<td>34</td>
<td>16.43</td>
<td>32</td>
<td>15.46</td>
</tr>
<tr>
<td>Overall</td>
<td>91</td>
<td>8.79</td>
<td>80</td>
<td>7.73</td>
<td>91</td>
<td>8.79</td>
<td>90</td>
<td>8.70</td>
</tr>
</tbody>
</table>

HIZ: High intensity zone. Values were the number of HIZ (%) out of a total of 207 patients at each level. Overall prevalence was the number of HIZ (percentage) in a total of 1035 lumbar levels.

### Table 3. Inter- and intraobserver agreement of visual assessments measured by statistic.

<table>
<thead>
<tr>
<th></th>
<th>Interobserver reliability (n=207)</th>
<th>Intraobserver reliability (n=40)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall (95% CI)</td>
<td>AB</td>
</tr>
<tr>
<td>L4–5</td>
<td>0.62 (0.56–0.69)</td>
<td>0.56</td>
</tr>
<tr>
<td>L5–S1</td>
<td>0.61 (0.55–0.67)</td>
<td>0.55</td>
</tr>
</tbody>
</table>

CI: Confidence interval. The data of overall reliability were computed by using 1000 bootstrapped samples.

### Table 4. Intra- and interobserver reliability of quantitative measurements (ICC).

<table>
<thead>
<tr>
<th></th>
<th>L4–5 (n=47)</th>
<th>L5–S1 (n=46)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intra-ICC</td>
<td>Inter-ICC (95% CI)</td>
</tr>
<tr>
<td></td>
<td>Observer E</td>
<td>Observer F</td>
</tr>
<tr>
<td>Area of disc</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>Area of HIZ</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>HIZ%</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>$S_{\text{CSF-HIZ}}$</td>
<td>0.92</td>
<td>0.96</td>
</tr>
</tbody>
</table>

HIZ: High intensity zone; CSF: Cerebrospinal fluid; CI: Confidence interval; ICC: Intraclass correlation coefficient. HIZ%: The area proportion of HIZ (area of HIZ/area of disc); $S_{\text{CSF-HIZ}}$: The CSF-adjusted signal intensity of HIZ (signal intensity of HIZ/signal intensity of CSF).
at L5–S1, though the differences were not significant.

A total of 93 HIZ were reported in visual assessments, including 47 at L4–5 and 46 at L5–S1. Of these, full agreement was achieved in 17 instances (18.3%). There were 52 (55.9%) and 16 HIZ (17.2%) agreed upon by 3 and 2 visual observers, respectively. The remaining 8 HIZ (8.6%) were reported by only 1 visual observer. As shown in Figure 2, the SICSF-HIZ with 4 agreements was 64.13±8.99%, which was significantly higher than all others (p<0.01). Additionally, there was statistically significant difference of the SICSF-HIZ between HIZ, with 3 and 1 agreements (55.67±8.55% versus 48.42±3.76%, p=0.02).

The area of HIZ with 2 agreements was smallest, which was found to be significantly different from the area of HIZ with 4 agreements (p=0.01) and 3 agreements (p=0.04). The HIZ% with 2 agreements was significant less than that with 4 agreements (p=0.03) and 3 agreements (p=0.02).

### Table 5: Comparison of quantitative measurements between observers E and F

<table>
<thead>
<tr>
<th>Measurement</th>
<th>L4–5</th>
<th>L5–S1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observer E (n=47)</td>
<td>Observer F (n=47)</td>
</tr>
<tr>
<td>Area of disc (mm²)</td>
<td>470.26±77.82</td>
<td>467.62±80.06</td>
</tr>
<tr>
<td>Area of HIZ (mm²)</td>
<td>6.97±2.93</td>
<td>6.25±2.23</td>
</tr>
<tr>
<td>HIZ%</td>
<td>1.51±0.65</td>
<td>1.36±0.51</td>
</tr>
<tr>
<td>SICSF-HIZ (%)</td>
<td>56.02±10.00</td>
<td>57.22±10.63</td>
</tr>
</tbody>
</table>

HIZ: High intensity zone; CSF: Cerebrospinal fluid; ICC: Intraclass correlation coefficient. HIZ%: The area proportion of HIZ (area of HIZ/area of disc); SICSF-HIZ: The CSF-adjusted signal intensity of HIZ (signal intensity of HIZ/signal intensity of CSF). *Compared average measurements between L4–5 and L5–S1.

Fig. 2. Of 93 reported HIZ in visual assessments, there were 17 HIZ with 4 agreements, 52 with 3 agreements, 16 with 2 agreements, and 8 with 1 agreement. (a) The SICSF-HIZ with 4 agreements was significantly higher than all the others (p<0.01). The SICSF-HIZ with 3 agreements was significantly higher than that with 1 agreement (p=0.02). (b) There was no significant difference of area of disc between different agreements. (c) The area of HIZ with 2 agreements was significant smaller compared with that with 4 agreements (p=0.01) and 3 agreements (p=0.02). (d) The HIZ% with 2 agreements was significant less than that with 4 agreements (p=0.04) and 3 agreements (p=0.03).
ments (p=0.02). Again, HIZ% with 2 agreements was significantly lower than HIZ% with 4 agreements (p=0.04) and 3 agreements (p=0.03).

In order to investigate the effect of signal intensity of HIZ on interobserver variation, HIZ were classified as 29 marked hyperintense (SICSF-HIZ >60%), 38 moderate hyperintense (50–60%), and 26 mild hyperintense (<50%). As shown in Figure 3, 14 (48.3%) marked HIZ were agreed upon by 4 visual observers, which was significantly higher than that in the moderate group ($\chi^2=14.16$, p<0.01; OR=10.89, 95% CI=2.72–43.54). There were no mild hyperintense HIZ with full agreement. Moreover, HIZ agreed upon by at least 3 visual observers in the marked hyperintense group were significantly different from those in the moderate group ($\chi^2=6.27$, p=0.02, OR=10.00, 95% CI 1.20–83.42) and mild ($\chi^2=15.66$, p<0.01, OR=28.00, 95% CI 3.30–237.43).

Analogously, HIZ were also be classified by area ratio as 34 large (HIZ%>1.5%), 32 middle (1.0–1.5%), and 27 small (<1.0%). Although the incidence of HIZ with full agreement was higher in large area ratio, the difference was not significant (Figure 3b).

Discussion

In terms of reliability and variability in the interpretation of HIZ, this study showed that inter- and intraobserver agreement was substantial or excellent. The $\kappa$ value of the overall interobserver agreement was 0.62 and 0.61 at L4–5 and L5–S1, respectively. The $\kappa$ value for individual visual observers was >0.80.

The interpretation of images depends on the criteria of finding, the frequency of the abnormality, the heterogeneity of population, and the size of sample. There may be variation in interpretation due to the observers. It may be difficult to obtain consensus when images are interpreted by a small number of observers. Furthermore, if observers work together in a research setting, this may also lead to an informal agreement in the diagnostic criteria and, therefore, result in under- or overestimation of the concordance. Finally, it is well recognized that different specialists may have their own response bias and preference in interpretation.

In our study, 207 eligible patients with chronic LBP were recruited from 3 outpatient departments. The inclusion criterion of HIZ was an expanded definition, which has been widely used in practice. In order to eliminate artifactual agreement beyond chance, only those HIZ with a mean prevalence between 10% and 90% at each spine level were included in the analysis. Additionally, the visual observers in this study were 4 spine surgeons from different institutions, with differing levels of clinical experience. They did not receive additional pre-test training or instruction, and were asked to interpret images as they would at their routine practice. With the above strengths, we therefore considered that the substantial interobserver agreement and excellent intraobserver agreement found in this study (Table...
3) could be representative of typical interpretations of HIZ by spine surgeons.

Although inter- and intraobserver agreement of HIZ is acceptable, it should be noted that discrepancy in the interpretation exists (Table 1). Smith et al. found that less than half of HIZ (44.1%) was agreed upon by 2 radiologists. In our study, however, only 18.3% of HIZ reached full agreement. This level of agreement was in marked contrast to previous studies. Since there were 4 spine surgeons in this study, we speculated that multiple specialists may be responsible for this discrepancy.

It has been suggested that discrepancy in the interpretation could be to some degree evitable by identifying systematic differences of interobserver variation. To our knowledge, however, no data are available regarding interobserver variation of HIZ. In this study, signal intensity (SICSF-HIZ) and area ratio (HIZ%) were used to assess the interobserver variation. The HIZ with full agreement had the greatest SICSF-HIZ, followed by those with 3 and 2 agreements (Figure 2); thus, our results support that the higher the HIZ signal intensity, the more visual observers agree upon it. In order to evaluate the effects of SI, the brightness of HIZ was divided into 3 grades. This showed that nearly half of marked HIZ reached consensus by 4 visual observers, but none of mild HIZ were in full observer agreement (Figure 3). Although fewer observers with consensus were associated with smaller area ratio of HIZ, the difference was not significant. Analogously, even when HIZ% was divided into 3 different area ratios, the distribution of HIZ with full agreement had no significant difference (Figure 3b).

Therefore, the results of this study imply that the SI of HIZ strongly influences interobserver variation. Indeed, as pointed out by Bogduk, the SI was the basic characteristic of HIZ, meaning that failure to distinguish the brightness of HIZ will not only increase the discrepancy in the interpretation but also decrease the diagnostic value in the clinical utility.

In conclusion, we conducted a prospective multicenter observational study in patients with chronic LBP to evaluate the reliability and variability of HIZ. The reliability was sufficient for spine surgeons with differing levels of clinical experience. However, interobserver variation was a general problem in the interpretation, which was caused mainly by the brightness of HIZ. During observation, it should therefore be recognized that HIZ is a more intense signal than any other area in the fibrous annulus.

Conflicts of Interest: No conflicts declared.

References
imaging findings. Radiology 2009;250:161–70.


