ORIGINAL ARTICLE

Acta Orthop Traumatol Turc 2015;49(6):606–613
doi: 10.3944/AOTT.2015.14.0267

Reliability and variability in the interpretation of lumbar high intensity zone

Shun-Wu FAN[1], Xiang-Qian FANG[1], Yun-Jian LIU[2], He-Jun YU[1], Yin-Jiang LU[3], Chao LIU[1]

[1]*Zhejiang University Faculty of Medicine, Sir Run Run Shaw Hospital, Department of Orthopedics, Zhejiang Province, China*
[2]*Wenzhou Medical College, Lishui Central Hospital, Department of Orthopedics, Zhejiang Province, China*
[3]*Shangyu 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 ($\kappa=0.62$ at L4–5 and 0.61 at L5–S1), and intraobserver agreement was excellent ($\kappa=0.84$ 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

Correspondence: Chao Liu, MD. Zhejiang University Faculty of Medicine, Sir Run Run Shaw Hospital, Department of Orthopedics, Zhejiang Province, China.

Tel: +86-571-86090073 e-mail: liuchaozju@gmail.com

Submitted: July 28, 2014 **Accepted:** May 17, 2015

Available online at
www.aott.org.tr
doi: 10.3944/AOTT.2015.14.0267
QR (Quick Response) Code

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.[13,15–19] 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.

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

First authors /year	Study samples	Male/female subjects or ratio	Age range years (mean)	MRI	Location of HIZ	Observers	Inter-rater (95% CI)	Intra-rater (95% CI)	Observer agreement
Aprill/1992	67 LBP subjects	5:2	20–60	0.6 T	Posterior AF	2	NR	NR	167/168 (99.4%)
Weishaupt/1998	60 asymptomatic subjects	30/30	20–50 (35)	1.0 T	Posterior AF	2 radiologists	0.91	NR	NR
Smith/1998	72 LBP subjects	1:1	23–70 (46)	1.5 T	Posterior AF	2 radiologists	0.57 (0.44–0.70)	NR	26/59 (44.1%)
Carrino/2009	111 subjects in SPORT	54/57	(53.5±16.2)	1.5 T	Posterior, lateral and posterolateral AF	3 radiologists and 1 spine surgeon	0.44 (0.34–0.55)	0.67 (0.53–0.78)*	NR
Zook/2011	71 surgically treated subjects	NR	NR	NR	NR	3 spine surgeons	0.824	0.683–0.854	NR
Hancock/2012	30 subjects with chronic LBP and 30 subjects with acute LBP	53:47	18–50 (37)	1.5 T	Posterior AF	1 radiologist and 1 spinal neuro-surgeon	0.46 (0.22–0.69)	NR	77%
Berg/2012	170 LBP subjects	82/88	25–55 (41)	1.5 T in 150 of 170 cases	Posterior AF	3 radiologists	0.58 (0.46–0.68) at L4–5; 0.46 (0.34–0.58) at L5–S1	0.81–0.85 at L4–5; 0.60–0.88 at L5–S1#	NR
Takatalo/2012	554 subjects	233/321	20–23 (21.2)	1.5 T	Posterior AF	2 radiologists	0.89	NR	NR

HIZ: High intensity zone; CI: Confidence interval; LBP: Low back pain; AF; Annular fibrosis; NR: No report; SPORT: spine patient outcomes research trial. *The intraobserver agreement was based on 40 images. #The intraobserver agreement was based on 126 images.

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 committees of the participating institutions.

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.

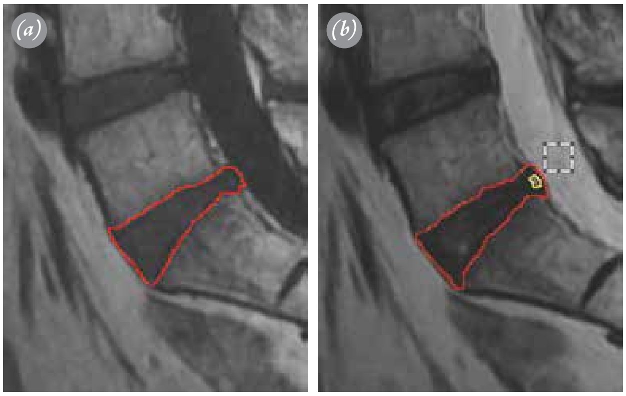

Fig. 1. The corresponding disc was outlined on **(a)** T1-weighted MRI; the segmentation overlays were copied and moved to **(b)** T2-weighted image. The white frame indicated a clean sample of CSF. The outer perimeter of HIZ was contoured with yellow line on T2-weighted MRI. [Color figure can be viewed in the online issue, which is available at www.aott.org.tr]

The reliability of visual assessments and quantitative measurements was calculated by κ statistics and intraclass correlation coefficient (ICC) formula,[3,1] respectively. The κ 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 χ^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-

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

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

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.

	Interobserver reliability (n=207)							Intraobserver reliability (n=40)				
	Overall (95% CI)	AB	AC	AD	BC	BD	CD	Overall (95% CI)	A	B	C	D
L4–5	0.62 (0.56–0.69)	0.56	0.72	0.67	0.53	0.60	0.65	0.84 (0.72–0.96)	0.84	0.88	0.84	0.80
L5–S1	0.61 (0.55–0.67)	0.55	0.68	0.67	0.52	0.58	0.68	0.86 (0.75–0.97)	0.83	0.80	0.92	0.90

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).

	L4–5 (n=47)			L5–S1 (n=46)		
	Intra-ICC		Inter-ICC (95% CI)	Intra-ICC		Inter-ICC (95% CI)
Measurement	**Observer E**	**Observer F**		**Observer E**	**Observer F**	
Area of disc	0.96	0.96	0.95 (0.84–0.97)	0.96	0.95	0.95 (0.84–0.98)
Area of HIZ	0.93	0.91	0.84 (0.72–0.90)	0.91	0.91	0.83 (0.78–0.86)
HIZ%	0.92	0.92	0.83 (0.71–0.90)	0.92	0.92	0.81 (0.74–0.87)
$SI_{CSF-HIZ}$	0.92	0.96	0.89 (0.81–0.94)	0.92	0.92	0.88 (0.79–0.92)

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); $SI_{CSF-HIZ}$: The CSF-adjusted signal intensity of HIZ (signal intensity of HIZ/signal intensity of CSF).

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 κ of 0.62 (95% confidence interval [CI] 0.56–0.69) at L4–5 and κ of 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 (κ=0.84; 95% CI 0.72–0.96) and at L5–S1 (κ=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 5. Comparison of quantitative measurements between observers E and F.

Measurement	L4–5 Observer E (n=47)	L4–5 Observer F (n=47)	p	L4–5 Average (n=94)	L5–S1 Observer E (n=46)	L5–S1 Observer F (n=46)	p	L5–S1 Average (n=92)	p*
Area of disc (mm²)	470.26±77.82	467.62±80.06	0.87	468.94±77.85	456.05±87.81	459.91±93.25	0.84	457.98±89.54	0.53
Area of HIZ (mm²)	6.97±2.93	6.25±2.23	0.19	6.61±2.54	5.62±2.34	5.55±2.29	0.89	5.58±2.27	0.04
HIZ%	1.51±0.65	1.36±0.51	0.21	1.43±0.56	1.26±0.57	1.23±0.52	0.81	1.24±0.54	0.10
$SI_{CSF-HIZ}$ (%)	56.02±10.00	57.22±10.63	0.57	56.62±9.80	55.29±10.35	55.08±8.31	0.91	55.18±823	0.45

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

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 agree-

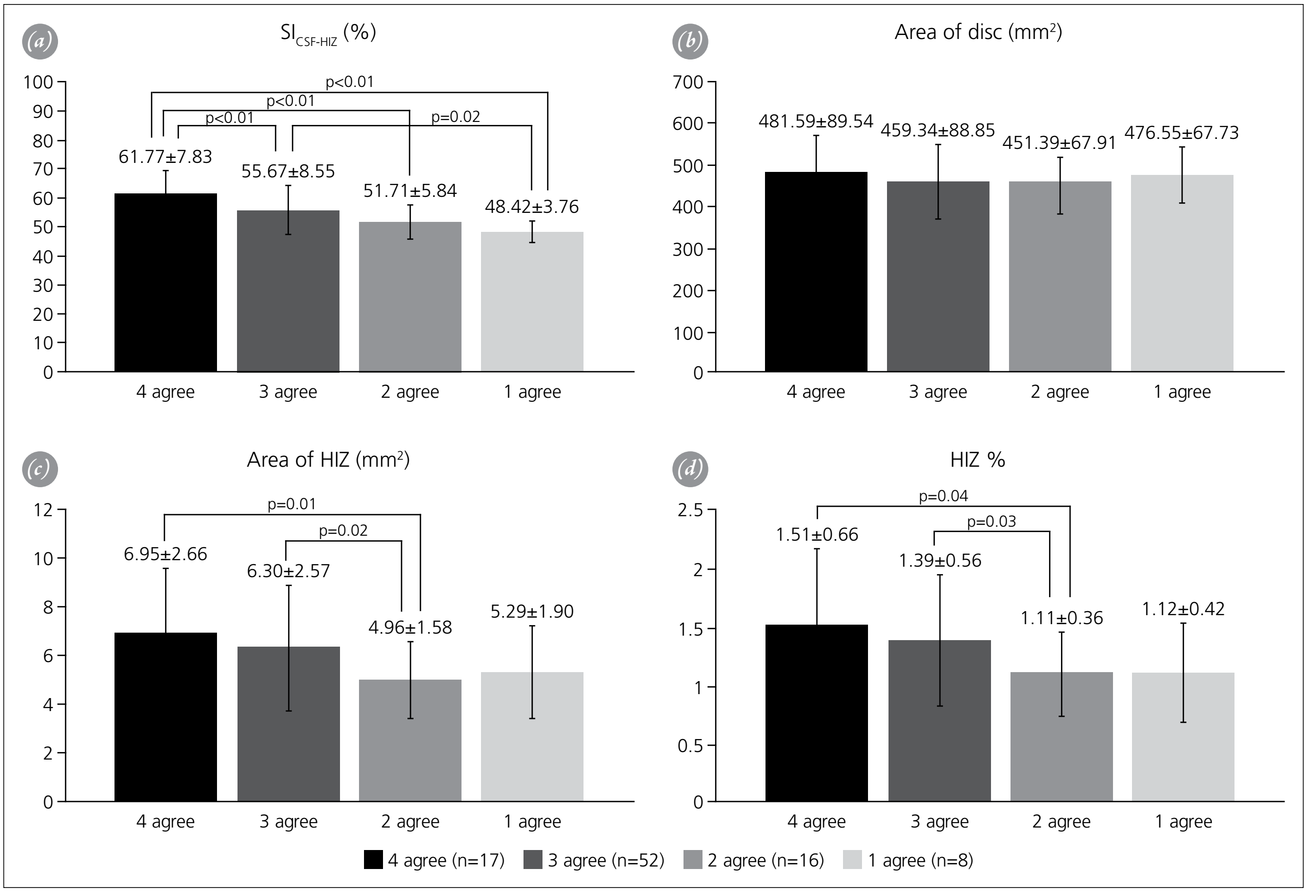

Fig. 2. Of 93 reported HIZ in visual assessments, there were 17 HIZ with 4 agreements, 52 with 3 agreements, 16 with 2 agreements, and 8 with 1 agreement. **(a)** The $SI_{CSF-HIZ}$ with 4 agreements was significantly higher than all the others (p<0.01). The $SI_{CSF-HIZ}$ with 3 agreements was significantly 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).

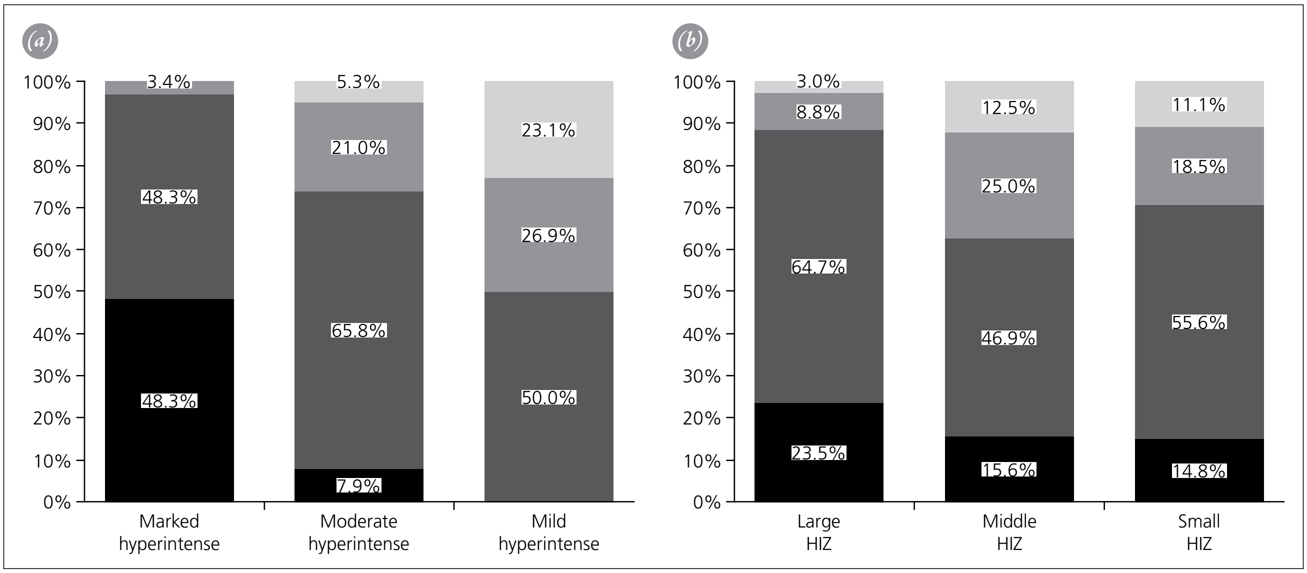

Fig. 3. The effects of signal intensity of HIZ ($SI_{CSF\text{-}HIZ}$) and area ratio of HIZ (HIZ%) on interobserver variation. **(a)** There were 14 (48.3%) HIZ with 4 agreements in marked hyperintense. It was significantly more than 3 (7.9%) HIZ with 4 agreements in moderate hyperintense (χ^2=14.16, p<0.01; OR=10.89, 95% CI=2.72–43.54). None of HIZ with 4 agreements was obtained in mild hyperintense. **(b)** The incidence of HIZ with 4 agreements was higher (23.5%) in large area ratio, but the difference was not significant when compared with that in middle and small area ratios.

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 (χ^2=14.16, p<0.01, odds ratio [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 (χ^2=6.27, p=0.02, OR=10.00, 95% CI 1.20–83.42) and mild (χ^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 κ value of the overall interobserver agreement was 0.62 and 0.61 at L4–5 and L5–S1, respectively. The κ 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.[25] There may be variation in interpretation due to the observers.[25,26] It may be difficult to obtain consensus when images are interpreted by a small number of observers.[26,27] 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.[28] Finally, it is well recognized that different specialists may have their own response bias and preference in interpretation.[4,17,25,29]

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.[19] 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.[4,17] 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,[30] 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 fibrosus annulus.

Conflics of Interest: No conflicts declared.

References

1. Brant-Zawadzki MN, Jensen MC, Obuchowski N, Ross JS, Modic MT. Interobserver and intraobserver variability in interpretation of lumbar disc abnormalities. A comparison of two nomenclatures. Spine (Phila Pa 1976) 1995;20:1257–64.
2. Raininko R, Manninen H, Battié MC, Gibbons LE, Gill K, Fisher LD. Observer variability in the assessment of disc degeneration on magnetic resonance images of the lumbar and thoracic spine. Spine (Phila Pa 1976) 1995;20:1029–35.
3. Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. Ann Intern Med 1983;99:843–8.
4. Espeland A, Korsbrekke K, Albrektsen G, Larsen JL. Observer variation in plain radiography of the lumbosacral spine. Br J Radiol 1998;71:366–75.
5. Jarvik JG, Deyo RA. Moderate versus mediocre: the reliability of spine MR data interpretations. Radiology 2009;250:15–7.
6. van Rijn JC, Klemetsö N, Reitsma JB, Majoie CB, Hulsmans FJ, Peul WC, et al. Observer variation in MRI evaluation of patients suspected of lumbar disk herniation. AJR Am J Roentgenol 2005;184:299–303.
7. Aprill C, Bogduk N. High-intensity zone: a diagnostic sign of painful lumbar disc on magnetic resonance imaging. Br J Radiol 1992;65:361–9.
8. Schellhas KP, Pollei SR, Gundry CR, Heithoff KB. Lumbar disc high-intensity zone. Correlation of magnetic resonance imaging and discography. Spine (Phila Pa 1976) 1996;21:79–86.
9. Lam KS, Carlin D, Mulholland RC. Lumbar disc high-intensity zone: the value and significance of provocative discography in the determination of the discogenic pain source. Eur Spine J 2000;9:36–41.
10. Peng B, Hou S, Wu W, Zhang C, Yang Y. The pathogenesis and clinical significance of a high-intensity zone (HIZ) of lumbar intervertebral disc on MR imaging in the patient with discogenic low back pain. Eur Spine J 2006;15:583–7.
11. Ricketson R, Simmons JW, Hauser BO. The prolapsed intervertebral disc. The high-intensity zone with discography correlation. Spine (Phila Pa 1976) 1996;21:2758–62.
12. Ito M, Incorvaia KM, Yu SF, Fredrickson BE, Yuan HA, Rosenbaum AE. Predictive signs of discogenic lumbar pain on magnetic resonance imaging with discography correlation. Spine (Phila Pa 1976) 1998;23:1252–60.
13. Weishaupt D, Zanetti M, Hodler J, Boos N. 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.
14. Carrino JA, Lurie JD, Tosteson AN, Tosteson TD, Carragee EJ, Kaiser J, et al. Lumbar spine: reliability of MR

imaging findings. Radiology 2009;250:161–70.
15. Zook J, Djurasovic M, Crawford C 3rd, Bratcher K, Glassman S, Carreon L. Inter- and intraobserver reliability in radiographic assessment of degenerative disk disease. Orthopedics 2011;34.
16. Hancock M, Maher C, Macaskill P, Latimer J, Kos W, Pik J. MRI findings are more common in selected patients with acute low back pain than controls? Eur Spine J 2012;21:240–6.
17. Berg L, Neckelmann G, Gjertsen O, Hellum C, Johnsen LG, Eide GE, et al. Reliability of MRI findings in candidates for lumbar disc prosthesis. Neuroradiology 2012;54:699–707.
18. Takatalo J, Karppinen J, Niinimäki J, Taimela S, Mutanen P, Sequeiros RB, et al. Association of modic changes, Schmorl's nodes, spondylolytic defects, high-intensity zone lesions, disc herniations, and radial tears with low back symptom severity among young Finnish adults. Spine (Phila Pa 1976) 2012;37:1231–9.
19. Smith BM, Hurwitz EL, Solsberg D, Rubinstein D, Corenman DS, Dwyer AP, et al. Interobserver reliability of detecting lumbar intervertebral disc high-intensity zone on magnetic resonance imaging and association of high-intensity zone with pain and anular disruption. Spine (Phila Pa 1976) 1998;23:2074–80.
20. Hebelka H, Hansson T. HIZ's relation to axial load and low back pain: investigated with axial loaded MRI and pressure controlled discography. Eur Spine J 2013;22:734–9.
21. Liu C, Cai HX, Zhang JF, Ma JJ, Lu YJ, Fan SW. Quantitative estimation of the high-intensity zone in the lumbar spine: comparison between the symptomatic and asymptomatic population. Spine J 2014;14:391–6.
22. Rankine JJ, Gill KP, Hutchinson CE, Ross ER, Williamson JB. The clinical significance of the high-intensity zone on lumbar spine magnetic resonance imaging. Spine (Phila Pa 1976) 1999;24:1913–20.
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
24. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005;85:257–68.
25. Madan SS, Rai A, Harley JM. Interobserver error in interpretation of the radiographs for degeneration of the lumbar spine. Iowa Orthop J 2003;23:51–6.
26. Obuchowski NA, Zepp RC. Simple steps for improving multiple-reader studies in radiology. AJR Am J Roentgenol 1996;166:517–21.
27. Beam CA, Baker ME, Paine SS, Sostman HD, Sullivan DC. Answering unanswered questions: proposal for a shared resource in clinical diagnostic radiology research. Radiology 1992;183:619–20.
28. Arana E, Royuela A, Kovacs FM, Estremera A, Sarasíbar H, Amengual G, et al. Lumbar spine: agreement in the interpretation of 1.5-T MR images by using the Nordic Modic Consensus Group classification form. Radiology 2010;254:809–17.
29. Ker M. Issues in the use of kappa. Invest Radiol 1991;26:78–83.
30. Bogduk N. Point of View: Predictive signs of discogenic lumbar pain on magnetic resonance imaging with discography correlation. Spine (Phila Pa 1976) 1998;23:1259-1260.