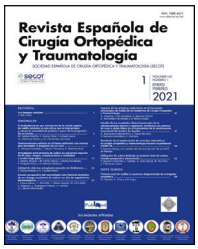
Revista Española de Cirugía Ortopédica y Traumatología 67 (2023) 175–180

Revista Española de Cirugía Ortopédica y Traumatología

www.elsevier.es/rot



ORIGINAL PAPER

Independent inter and intra-observer agreement of the Schizas's classification of degenerative lumbar stenosis: Comparison among three levels of surgical training

F. Holc[a], A. Albani-Forneris[a], G. Kido[a], S. Beltrame[a], M. Petracchi[a], M. Gruenberg[a], C. Sola[a], G. Camino-Willhuber[a,b,*]

[a] *Orthopaedic and Traumatology Department, Institute of Orthopedics ''Carlos E. Ottolenghi,'' Hospital Italiano de Buenos Aires, Buenos Aires, Argentina*
[b] *Department of Orthopaedics, University of California at Irvine, 101 The City Drive South, Orange, CA 92868, USA*

Received 29 June 2022; accepted 1 October 2022
Available online 12 October 2022

KEYWORDS
Independent agreement;
Schizas's classification;
Lumbar spinal stenosis

Abstract
Introduction and objectives: Lumbar spinal stenosis is a common age-related condition that affects the quality of life. Multiple classifications have been developed to quantify the severity of stenosis affecting comparison between studies and homogenous communication among surgeons and researchers. Even though this classification has not shown a direct clinical correlation, Schizas's classification appears to be a simple method to assess stenosis. Our objective was to evaluate the inter and intraobserver independent agreement of the Schizas's classification to assess stenosis severity. Additionally, we aimed to compare agreement among three levels of training in spine surgery.
Materials and methods: An independent inter and intra observer agreement was conducted among junior, senior orthopedic residents and attending spine surgeons. Ninety lumbar levels from 30 patients were evaluated by 16 observers. Weighted kappa agreement was used.
Results: Overall interobserver and intraobserver agreement was of 0.57 (95% CI = 0.52–0.63) and 0.69 (0.55–0.79), respectively. Interobserver agreement according to level of training yielded values of 0.53 (0.46–0.60) for junior residents, 0.61 (0.54–0.67) for senior residents and 0.67 (0.59–0.74) for attendings. Intraobserver agreement was of 0.54 (0.48–0.60) for junior, 0.60 (0.55–0.66) for senior and 0.66 (0.60–0.72) for attendings.

* Corresponding author.
E-mail address: gaston.camino@hospitalitaliano.org.ar (G. Camino-Willhuber).

https://doi.org/10.1016/j.recot.2022.10.003
1888-4415/© 2022 SECOT. Published by Elsevier España, S.L.U. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Conclusion: The Schizas's classification showed moderate interobserver and substantial intraobserver agreement. Among attending surgeons, substantial inter and intraobserver agreement was observed. The classification allowed acceptable communication among trained spine surgeons.
© 2022 SECOT. Published by Elsevier España, S.L.U. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

PALABRAS CLAVE
Acuerdo independiente;
Clasificación de Schizas;
Estenosis espinal lumbar

Acuerdo inter e intraobservador independiente de la clasificación de Schizas de estenosis lumbar degenerativa. Comparación entre tres niveles de entrenamiento quirúrgico

Resumen
Introducción y objetivos: La estenosis espinal lumbar es una condición frecuente relacionada con la edad que afecta la calidad de vida. Se han desarrollado múltiples clasificaciones para cuantificar la gravedad de la estenosis que afectan la comparación entre estudios y la comunicación homogénea entre cirujanos e investigadores. A pesar de que esta clasificación no ha mostrado una correlación directa con la clínica, la clasificación de Schizas parece ser un método simple para evaluar la estenosis. Nuestro objetivo fue evaluar el acuerdo independiente inter e intraobservador de la clasificación de Schizas en la severidad de la estenosis. Además, comparamos la concordancia entre tres niveles de formación en cirugía de columna.
Materiales y métodos: Se llevó a cabo un acuerdo independiente inter e intraobservador entre los residentes ortopédicos principiantes, avanzados y los cirujanos de columna; 90 niveles lumbares de 30 pacientes fueron evaluados por 16 observadores. Se utilizó concordancia a través del kappa ponderado.
Resultados: La concordancia global interobservador e intraobservador fue de 0,57 (IC 95% = 0,52-0,63) y 0,69 (0,55-0,79), respectivamente. La concordancia interobservador según el nivel de formación arroja valores de 0,53 (0,46-0,60) para los residentes menores, 0,61 (0,54-0,67) para los residentes mayores y 0,67 (0,59-0,74) para los asistentes. La concordancia intraobservador fue de 0,54 (0,48-0,60) para principiantes, 0,60 (0,55-0,66) para avanzados y 0,66 (0,60-0,72) para cirujanos de columna.
Conclusión: La clasificación de Schizas mostró concordancia interobservador moderada y concordancia intraobservador sustancial. Entre los cirujanos de columna, se observó un acuerdo sustancial inter e intraobservador. La clasificación permitió una comunicación aceptable entre los cirujanos de columna entrenados.
© 2022 SECOT. Publicado por Elsevier España, S.L.U. Este es un artículo Open Access bajo la licencia CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Lumbar spinal stenosis (LSS) is one of the most common degenerative conditions of the spine and the first cause of spinal surgery in people over 65 years of age.[1–3]

Clinical presentation involves neurogenic claudication,[4] that worsen with lumbar extension and improve with flexion, affecting walking activities.[5] Along with clinical symptoms, magnetic resonance imaging (MRI) has become essential for diagnostic confirmation and surgical decision-making.[6]

Cross-sectional surface area has been suggested for diagnosis and severity, this system lacks clinical correlation.[7] In 2010, Schizas et al.[8] proposed a seven-categories classification based on the morphological appearance of the dural sac in axial T2 MR images of the lumbar spine, analyzing the disposition of the cerebrospinal fluid (CSF) and the nerve roots, with the advantage of not requiring specific measurement tools and be easily applicable to daily practice. The authors reported substantial interobserver reliability in their original study.

Even though this classification has been used in other studies, only two independent agreement studies have been performed.[9–10] Another independent evaluation of this system's reliability is important to further validate its use. Therefore, the objective of this study is to perform an independent, inter and intra-observer agreement analysis. Additionally, the authors aimed to compare agreement among observers of different levels of training in spine surgery-spinal disease.

Methods

Institutional review board approval was obtained to conduct this study; informed consent was waived due to the retrospective nature of the study without risk to participants. We retrospectively selected at random preoperative lumbar MR images of 30 patients who were seen in our outpatient

clinical division between 2019 and 2020. The MRI was requested for different reasons, such as chronic low back pain, leg pain or neurogenic claudication. MRIs were excluded from analysis in patients with history or suspicion of trauma, infection, or tumor; previous lumbar surgery and lack of T2-weighted axial image at any of the L2–L3, L3–L4, or L4–L5 disk levels. An author with a high level of expertise (M.P.), who later did not participate in the classification phase of this study, selected the cases. All the selected MRI images were stored digitally (DICOM) in the database.

The Schizas's classification is a 7-grade system based on the CSF/rootlet ratio on T2-weighted axial images. Grade A (no or minor stenosis), where CSF is observed inside the dural sac and its distribution is homogeneous. In turn, this grade is divided into 4 subgroups according to the arrangement of the nerve roots, where in A1; the roots remain in the dorsal region and occupy less than half the area of the dural sac, A2; the roots remain in the dorsal region and are distributed in a horseshoe shape, A3; the roots remain in the dorsal region and occupy more than half the area of the dural sac, A4; the roots are found in the central region and occupy most of the area of the dural sac. Grade B (moderate stenosis) comprises those stenoses in which the roots occupy the entire area of the dural sac but can still be individualized and there is a certain amount of CSF between them, giving a granular appearance. In grade C (severe stenosis) the roots cannot be recognized individually, resulting in a gray homogeneous signal within the sac and the posterior epidural fat is still present and grade D (extreme stenosis) is equivalent to grade C with the difference that no epidural fat is visualized posteriorly (Fig. 1).

Image analysis was conducted by 16 orthopedists with 3 different levels of experience. Three first year post graduate residents (PGY-1) and four PGY-2 formed the first group, three PGY-3 and three PGY-4 formed the second group and three spine surgeons with at least 5 years of experience in spine surgery formed the third group.

The evaluators were trained in this classification system through an online session to discuss it and to clarify doubts before performing the assessments in order to standardize the evaluation process. Additionally, they were provided with the original article by Schizas et al.[8] to solve any doubt at the time of evaluation. All participants independently evaluated the lumbar canal stenosis grade at L2–L3, L3–L4, and L4–L5 levels. Each observer classified a total of 90 levels. In the second phase, after a time interval of 4 weeks to avoid recall bias, the same examination with a random distribution of cases was performed by all observers.

Statistical analysis

Statistical analysis was performed using Stata 16 software (StataCorp, College Station, TX).

Considering the classification developed by Schizas et al. is an ordinal variable based on severity, we decided to use the weighted kappa statistics (*wk*) for two-way agreements. Weighted kappa allows measuring agreement with multiple response levels when not all disagreements are equally important; weight was set linearly. Inter-observer agreement was determined by comparing the initial read of all the

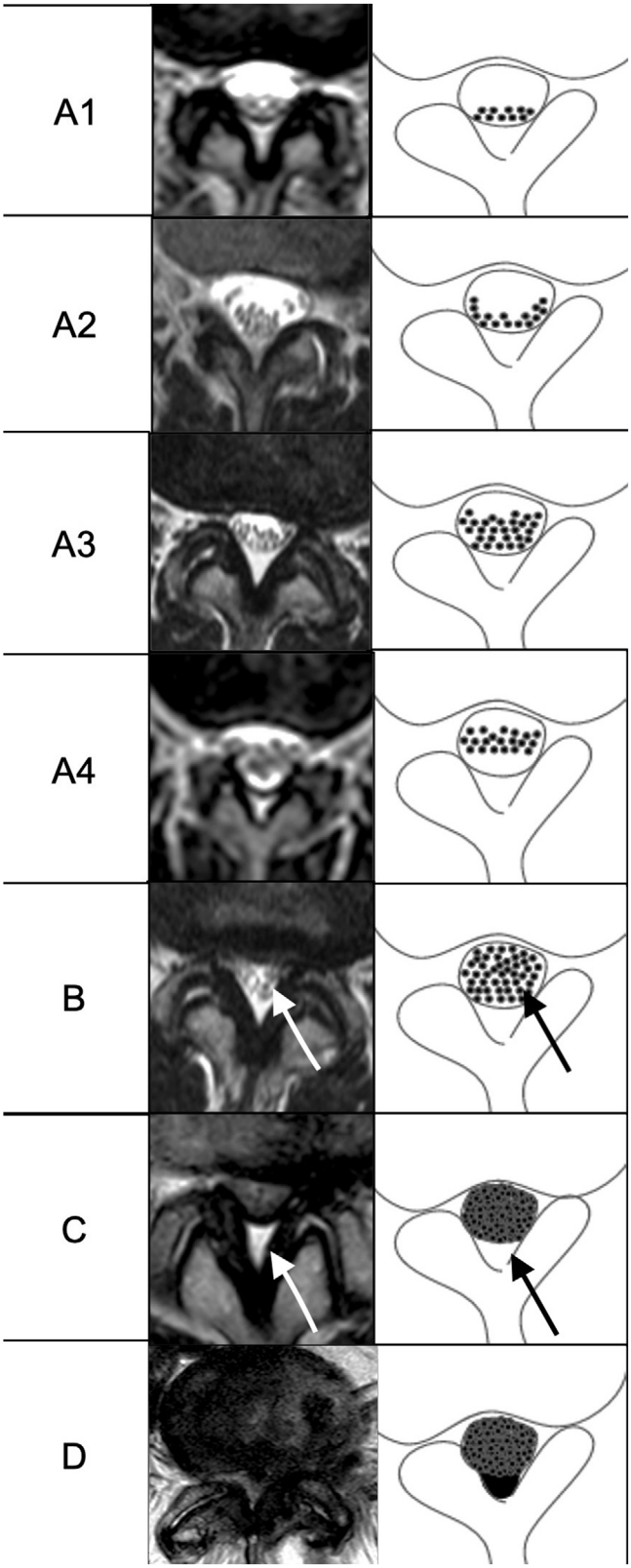


Figure 1 MRI and schematic representation of the Schizas's classification for lumbar stenosis. Arrows in B image show the individualized rootles that differentiate between B or moderate and C or severe stenosis. Arrows in C show the presence of posterior epidural fat that differentiates between C or severe stenosis and D or extreme stenosis.

Table 1 Interobserver agreement according to level of expertise.

Level of expertise	First evaluation kappa (95% CI)	Second evaluation kappa (95% CI)
PGY-1 and 2	0.53 (0.46–0.60)	0.53 (0.46–0.59)
PGY-3 and 4	0.61 (0.54–0.67)	0.60 (0.53–0.67)
Attendings	0.67 (0.59–0.74)	0.60 (0.52–0.68)

Table 2 Intraobserver agreement according to level of expertise.

Level of expertise	Kappa	(95% CI)
PGY-1 and 2	0.54	(0.48–0.60)
PGY-3 and 4	0.60	(0.55–0.66)
Attendings	0.66	(0.60–0.72)

assessors. Intra-observer agreement was calculated by comparing the same evaluator's reads between two assessments of the same patients. The two assessments were separated by a four-week interval and presented in a random sequence to avoid recall bias.

Levels of agreement for *wk* were determined as proposed by Landis et al.[11], as follows: *k* values 0.00–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement. All agreements are expressed with 95% confidence interval (CI).

Results

The overall interobserver and intraobserver agreement were of 0.57 (95% CI = 0.52–0.63) and $k = 0.69$ (0.55–0.79), respectively, corresponding to moderate and substantial agreement.

Interobserver agreement according to level of training

Interobserver agreement is illustrated in Table 1. Based on level of training, agreement for the PGY1-2 was lower (0.53) compared to PGY-3 and 4 (0.60) and attendings (0.67).

Intraobserver agreement according to level of training

Both PGY-1 and 2 and PGY-3 and 4 showed moderate agreement (0.54) compared with substantial agreement observed in attendings (0.66) (Table 2).

Interobserver agreement according to lumbar stenosis severity

Fair agreement was observed for Grades A2, A4 and B whereas Grade A1, A3, C and D showed moderate agreement (Table 3).

Table 3 Interobserver agreement according to lumbar stenosis severity.

Stenosis	Kappa	(95% CI)
Grade A1	0.59	(0.593–0.594)
Grade A2	0.26	(0.267–0.268)
Grade A3	0.41	(0.410–0.412)
Grade A4	0.20	(0.266–0.207)
Grade B	0.27	(0.279–0.28)
Grade C	0.39	(0.394–0.395)
Grade D	0.53	(0.533–0.534)

Interobserver agreement for each group according to lumbar stenosis severity

A2, A3 and D show the same concordance respectively in all groups, being fair for A2, and moderate for A3 and D. A1 and B had better agreement on the attendings group (Table 4).

Discussion

LSS is one of the most frequent diagnoses and one of the most common indications of spinal surgery. Independent agreement analyses of any classification are helpful in providing external validations that are essential to facilitate communication among physicians, standardize research terminology, and help to guide decisions making on patients.[12–13] In this study, we validated inter- and intraobserver agreements of a qualitative MRI grading system for central LSS between observers of different levels of training. In this regard, we found moderate overall interobserver and substantial intraobserver agreements (0.57 (95% CI = 0.52–0.63) and 0.69 (0.55–0.79)), respectively. Our results showed better agreement compared to the original authors of the classification.[8] Schizas et al. average inter and intraobserver kappa values were moderate (0.44) and substantial (0.65), respectively.

Ko et al.[12] in their correlation work included 5 observers: of whom two fellows with 3 months of training, a diagnostic imaging resident with 1 month of experience, and two were surgeons with more than 10 years of experience. They showed an intraclass correlation index (ICC) of 0.82–0.98, equivalent to almost perfect agreement. Regarding intraobserver correlation, they reported an ICC of 0.65–0.99, which is interpreted as moderate to excellent agreement. However, these results are not comparable with those obtained in our study, since another correlation method was used.[12]

Another correlation study carried out by Weber et al.[9] that included two neurosurgeons and two diagnostic imaging specialists, all with extensive experience in their field; obtained a substantial interobserver agreement level

Table 4 Interobserver agreement for each group according to lumbar stenosis severity.

Stenosis	PGY-1 and 2		PGY-3 and 4		Attendings	
	Kappa	(95% CI)	Kappa	(95% CI)	Kappa	(95% CI)
Grade A1	0.59	(0.518–0.608)	0.59	(0.534–0.641)	0.66	(0.544–0.783)
Grade A2	0.25	(0.209–0.299)	0.25	(0.198–0.304)	0.33	(0.209–0.448)
Grade A3	0.42	(0.374–0.465)	0.48	(0.423–0.530)	0.41	(0.288–0.526)
Grade A4	0.19	(0.144–0.234)	0.22	(0.172–0.279)	0.31	(0.190–0.429)
Grade B	0.31	(0.264–0.354)	0.27	(0.220–0.327)	0.45	(0.334–0.573)
Grade C	0.32	(0.279–0.369)	0.45	(0.393–0.500)	0.48	(0.363–0.602)
Grade D	0.51	(0.468–0.558)	0.51	(0.456–0.562)	0.59	(0.466–0.705)

$k = 0.76$ (95% CI 0.69–0.83) and an almost perfect intraobserver agreement level $k = 0.76$ at 0.96. Unlike our study, all the observers had extensive experience in the analysis of MR images and, in turn, used the abbreviated Schizas classification, since grade A subclassifications were not considered; therefore, it could affect the estimation value of the correlation, reducing a potential difference between the observations.

Lønne et al.[10] reported substantial inter-observer agreement $k = 0.65$ (CI 0.56–0.74) between two experienced radiologists using this system, and high intra-observer agreement $k = 0.78$ (CI 0.65–0.92) and $k = 0.81$ (CI 0.68–0.94), respectively. Like the study from Weber et al., the observers were experienced specialists and used the abbreviated Schizas classification; this might have a low external validity contribution beyond diagnostic imaging specialists. Furthermore, all the patients in this study were candidates for surgery, which could represent a biased selection.

When stratifying the analysis according to the level of training, we found moderate interobserver and intraobserver agreement among PGY-1 and PGY-2 residents (0.53 and 0.54, respectively) while PGY-3 and PGY-4 showed substantial interobserver agreement (0.61) and moderate (0.60) intraobserver agreement. The higher results were observed in the more expertise level of trainees (interobserver = 0.67, intraobserver 0.66). These differences could be explained by the fact that PGY-1 and PGY-2 residents are inherently exposed to fewer MRI spine cases compared to PGY-3 and PGY-4, as junior residents do not routinely evaluate patients in the spine unit at that moment of training and are only exposed to spine cases eventually during on-call activities.[13–15]

Interestingly, when comparing concordance between PYG-3 and PGY-4 residents and the attending physicians, small difference was observed, meaning that the ability to record and use this classification is essentially acquired and processed similarly in both groups. This finding supports the idea of the usefulness of this classification for communication and research purposes. Of note, it is worth mentioning that the participants in this study were not previously familiar with this classification system, which could have influenced our results.

Various studies have analyzed the reliability of classification systems among observers of different experience levels. Some of them demonstrated a positive correlation between agreement and clinical experience,[13,16,17] while others found no differences in the reliability of observers based on level of experience.[15,18]

Reliability analyzes[9,12,15,16] usually used participants of varying levels of experience and training, in an attempt to analyze the generalizability of the study emulating a more real situation of daily clinical practice. Nevertheless, to our knowledge, this is the first study that evaluates the agreement of this classification according to the clinical experience and level of training. Even though Ko et al.[12] involved observers of different levels of experience, no comparative analysis between each observer was performed.

Finally, when analyzing the agreement based on the severity of lumbar stenosis proposed by Schizas et al., we found a slight agreement in Grades A2, A4, and B, a moderate agreement in Grades A1, A3, C and D. We believe that this heterogeneous agreement could be explained by the fact that this classification has seven grades, especially considering that Grade A has four subtypes and the differences among them appear to be slight. Then, when studying the concordance for each degree of stenosis severity according to the level of expertise, we found that the extreme degrees were also the ones with the best agreement, and the experienced group presented a more homogeneous concordance between the different grades.

Our study has limitations and strengths, first, we analyzed MRI images without any clinical correlation. Whether the authors do not know if clinical information would influence the evaluation by the assessors, by including clinical findings our study would reproduce a more real clinical scenario of the daily practice.

Second, we only included participants with training in orthopedics, without participation of radiologist specialists or neurosurgery. While this could affect our results, the authors selected surgeons and residents of different training degree to better understand the influence of level of training when classifying lumbar stenosis.

Finally, the images delivered for analysis included the complete set of MRI, where each observer had to localize and evaluate the requested levels simulating a 'real world' context and allowing a more authentic assessment of the images.

Conclusion

This study independently validated the inter-and intraobserver agreement of the Schizas's morphological

classification system for central LSS among orthopedic residents of different training levels and spine surgeons. Interobserver agreement was moderate for junior residents and substantial for both senior residents and attending spinal surgeons, intraobserver agreement was substantial only for attending surgeons. Even though this classification has not shown a direct clinical correlation, Schizas's morphology-based system could allow homogeneous language communication among physicians at advanced levels of training. The ultimate utility of this classification should be evaluated in prospective large cohort studies.

Level of evidence

Level of evidence IV.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Genevay S, Atlas SJ, Katz JN. Variation in eligibility criteria from studies of radiculopathy due to a herniated disc and of neurogenic claudication due to lumbar spinal stenosis: a structured literature review. Spine. 2010;35:803–11.
2. Kalichman L, Cole R, Kim DH, Li L, Suri P, Guermazi A, et al. Spinal stenosis prevalence and association with symptoms: the Framingham study. Spine J. 2009;9:545–50.
3. Katz JN, Harris MB. Clinical practice. Lumbar spinal stenosis. N Engl J Med. 2008;358:818–25.
4. Deer T, Sayed D, Michels J, Josephson Y, Li S, Calodney AK. A review of lumbar spinal stenosis with intermittent neurogenic claudication: disease and diagnosis. Pain Med. 2019;20 Suppl 2:S32–44.
5. Katz JN, Dalgas M, Stucki G, Katz NP, Bayley J, Fossel AH, et al. Degenerative lumbar spinal stenosis. Diagnostic value of the history and physical examination. Arthritis Rheum. 1995;38:1236–41.
6. Lurie J, Tomkins-Lane C. Management of lumbar spinal stenosis. BMJ. 2016;352:h6234.
7. Schönström N, Lindahl S, Willén J, Hansson T. Dynamic changes in the dimensions of the lumbar spinal canal: an experimental study in vitro. J Orthop Res. 1989;7:115–21, 10.1002/jor.1100070116.
8. Schizas C, Theumann N, Burn A, Tansey R, Wardlaw D, Smith FW, et al. Qualitative grading of severity of lumbar spinal stenosis based on the morphology of the dural sac on magnetic resonance images. Spine. 2010;35:1919–24.
9. Weber C, Rao V, Gulati S, Kvistad KA, Nygaard ØP, Lønne G. Inter- and intraobserver agreement of morphological grading for central lumbar spinal stenosis on magnetic resonance imaging. Global Spine J. 2015;5:406–10, http://dx.doi.org/10.1055/s-0035-1551651.
10. Lonne G, Ødegard B, Johnsen LG, Solberg TK, Kvistad KA, Nygaard ØP. MRI evaluation of lumbar spinal stenosis: is a rapid visual assessment as good as area measurement? Eur Spine J. 2014;23:1320–4.
11. Landis JR, Richard Landis J, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74, http://dx.doi.org/10.2307/2529310.
12. Ko Y-J, Lee E, Lee JW, Park CY, Cho J, Kang Y, et al. Clinical validity of two different grading systems for lumbar central canal stenosis: Schizas and Lee classification systems. PLoS One. 2020;15:e0233633.
13. Yoon RS, Koerner JD, Patel NM, Sirkin MS, Reilly MC, Liporace FA. Impact of specialty and level of training on CT measurement of femoral version: an interobserver agreement analysis. J Orthop Traumatol. 2013;14:277–81, http://dx.doi.org/10.1007/s10195-013-0263-x.
14. Jin EH, Chung SJ, Lim JH, Chung GE, Lee C, Yang JI, et al. Training effect on the inter-observer agreement in endoscopic diagnosis and grading of atrophic gastritis according to level of endoscopic experience. J Korean Med Sci. 2018;33:e117, http://dx.doi.org/10.3346/jkms.2018.33.e117.
15. Niemeyer T, Wolf A, Kluba S, Halm HF, Dietz K, Kluba T. Interobserver and intraobserver agreement of Lenke and King classifications for idiopathic scoliosis and the influence of level of professional training. Spine (Phila Pa 1976). 2006;31:2103–8, http://dx.doi.org/10.1097/01.brs.0000231434.93884.c9.
16. Wing N, Van Zyl N, Wing M, Corrigan R, Loch A, Wall C. Reliability of three radiographic classification systems for knee osteoarthritis among observers of different experience levels. Skeletal Radiol. 2021;50:399–405, http://dx.doi.org/10.1007/s00256-020-03551-4.
17. Tzavellas AN, Kenanidis E, Potoupnis M, Pellios S, Tsiridis E, Sayegh F. Interobserver and intraobserver reliability of Salter–Harris classification of physeal injuries. Hippokratia. 2016;20:222–6.
18. Karamian BA, Schroeder GD, Levy HA, Canseco JA, Benneker LM, Kandziora F, et al., AO Spine Sacral Classification Group Members. The influence of surgeon experience and subspeciality on the reliability of the AO spine sacral classification system. Spine (Phila Pa 1976). 2021;46:1705–13, http://dx.doi.org/10.1097/BRS.0000000000004199.