Magn Reson Med Sci 2020; 19; 207–215
doi:10.2463/mrms.mp.2019-0079

Published Online: September 24, 2019

MAJOR PAPER

Interobserver Reliability When Classifying MR Imaging of the Lumbar Spine: Written Instructions Alone Do Not Suffice

Ulf Krister Hofmann[1*], Ramona Luise Keller[2], Marco Gesicki[1,3], Christian Walter[1], and Falk Mittag[1]

Purpose: Numerous classification systems have been proposed to analyze lumbar spine MRI scans. When evaluating these systems, most studies draw their conclusions from measurements of experienced clinicians. The aim of this study was to evaluate the impact of specific measurement training on interobserver reliability in MRI classification of the lumbar spine.

Methods: Various measurement and classification systems were assessed for their interobserver reliability in 30 MRIs from patients with chronic lumbar back and sciatic pain. Two observers were experienced spine surgeons. The third observer was an inexperienced medical student who, prior to the study measurements, in addition to being given the detailed written instructions also given to the surgeons, obtained a list of 20 reference measurements in MRI scans from other patients to practice with.

Results: Excellent agreement was observed between the medical student and the spine surgeon who had also created the reference measurements. Between the two spine surgeons, agreement was markedly lower in all systems investigated (e.g., antero-posterior spinal canal diameter intraclass correlation coefficient [ICC] [3.1] = 0.979 vs. ICC [3.1] = 0.857).

Conclusion: These data warrant the creation of publicly available standardised measurement examples of accepted classification systems to increase reliability of the interpretation of MR images.

Keywords: *facet joint degeneration, lumbar spinal stenosis, magnetic resonance imaging, neuroforaminal stenosis, Schizas' spinal stenosis classification*

Introduction

Lumbar spinal stenosis is characterised by a narrowing of the spinal canal, the lateral recess, or the neural foramen. The reasons for this condition are usually decreased height of the intervertebral disc with protrusion of the degenerative disc tissue into the canal, a hypertrophic flaval ligament that also bulges into the canal because of the decreased intervertebral height, and, in many cases, obtrusion of degenerate facet joints with osteophyte formation and capsular hypertrophy. This condition is the most frequent indication for spinal surgery in the lumbar region in patients older than 65 years.[1,2] While a clinical examination and reporting of the characteristic symptoms are essential in these patients, over the past two decades, the trend has been the base indications for surgery on radiographic imaging. Hence, standardised radiographic tools are needed to compare different treatment strategies to advance knowledge and improve performance in spine surgery. Although numerous classification systems for lumbar spinal stenosis have been developed as a means of objectifying the morphological observations, their applicability remains disputed and clear recommendations are not available. There also remains a lack of consensus about the classifications to use for observed anatomic abnormalities. Quantitative measurements are frequently taken, such as the antero-posterior diameter of the sagittal canal or the neural foramen,[3,4] but often they are not applied uniformly and sometimes they are even used in an individually modified way. With respect to the cut-off values, the discussion is still ongoing. Next to the fact that absolute values seem to be difficult to apply worldwide given the high heterogeneity of the

[1] Department of Orthopaedic Surgery, University Hospital of Tübingen, Hoppe-Seyler-Strasse 3, Tübingen D-72076, Germany
[2] Faculty of Medicine, Julius-Maximilians University of Würzburg, Würzburg, Germany
[3] Praxis Dres. Falck and Gesicki, Tübingen, Germany

[*] Corresponding author, Phone: +49-7071-29-86685, Fax: +49-7071-29-4091, E-mail: ulf.hofmann@med.uni-tuebingen.de

*This work has been presented as an oral presentation at the 12th German Spine Conference/Annual Meeting of the German Spine Society, 30th November to 2nd December 2017, Stuttgart, Germany.

Received: January 25, 2019 | Accepted: July 27, 2019

physis, the degree of neural compression that leads to particular symptoms has also never been established.[5] For these reasons, in recent years, qualitative instead of quantitative measurements have been proposed, for e.g., by Lee et al.[6] or by Schizas et al.[7] for spinal stenosis, or by Lee et al.[8] for neuroforaminal stenosis. In a recent consensus meeting to identify core radiological parameters to describe lumbar stenosis, qualitative parameters were also preferred over quantitative read-out.[9] Although qualitative classifications might control the variance in the physis, they could nonetheless be more prone to inter and intraobserver discrepancies and, again, their pathological value in terms of valid prediction of symptoms is yet to be evaluated. In view of the rising numbers of spine surgery in particular, more efforts are needed to implement a uniformly accepted classification system for lumbar degenerative pathologies.

A crucial prerequisite that such a classification system needs to meet before it can be implemented is high intra and interobserver reliability. But again, under what conditions do we report these reliabilities? Most data in the literature are derived from the initial descriptions of the classification, which bear a strong bias and originate from observers who usually spent much time working on this specific measurement problem. With respect to the overall experience of the observer, its effect is still a matter of controversy. Although the level of experience has been described as not correlating with a better kappa score when performing the Modic classification,[10] Fayad et al.[11] described a higher repeatability between experienced observers. Generally, interobserver reliability is considered lower between clinicians of different specialties or when readers are not subspecialised and do not work together (reviewed by Andreisek et al.[3]). The true impact on reliability of training in advance using a set of reference measurements is rarely evaluated.

The aim of this study was to evaluate the impact of specific measurement training on interobserver reliability in MRI classification of the lumbar spine. To this end, we looked at inter-rater agreement for various quantitative measurements and qualitative classifications for lumbar spinal stenosis. We analyzed inter-rater reliability between two experienced observers (spine surgeons) who were using the same standardised instructions. We then compared these results with those obtained from a medical student after she had practiced in accordance with measurement examples provided by one of the spine surgeons. Our hypothesis was that the training provided is more important than professional experience with respect to achieving high inter-rater reliabilities for the measurements performed.

Materials and Methods

Study design

We analyzed MR images from 30 patients with chronic lumbar back and sciatic pain who attended our department for evaluation of indications for surgery. Various parameters were used to evaluate either spinal canal stenosis, neuroforaminal stenosis, or facet joint degeneration by three observers (two experienced spine surgeons from the same institution and one medical student). Exclusion criteria were spondylolisthesis and a neoplastic or infectious aetiology.

A list was generated with detailed instructions for the measurements in accordance with the literature, including information on which sequences to use; for some parameters, the information from the literature was amended when further instructions were deemed necessary for reasons of clarity and better specification. If image material was available in the original literature, this was also added to the measurement instructions.

One of the two experienced spine surgeons familiarised himself with the different classifications and measurement protocols, and then measured 20 motion segments (functional spinal units), the results of which were later made available to the medical student as reference measurements. This patient group was independent from the patients included for the final measurements. The medical student received a brief tutorial on MRI interpretation of the lumbar spine, and then was handed the written instructions for the measurement technique. Thereafter, the data set of the reference measurements was given to the student to practice measuring on the same MRIs and adjust her measurement technique to obtain similar results as denoted in the reference measurements. Only after this practice with the given reference measurement values was completed were the study measurements performed by the student.

The other experienced spine surgeon received the detailed specified written measurement instructions only, without access to the reference measurements database. Only one predefined motion segment of the lumbar spine was measured per patient. If an observer deemed a measurement not to make sense because of image quality or poorly defined measurement reference points, no measurement was performed.

Full departmental, institutional, and local ethical committee approval were obtained before commencement of the study (project number 503/2016BO2).

Measurements performed

Spinal Canal Stenosis

(a) Quantitative Measurements

The sagittal diameter (e.g., Verbiest[12]; Ullrich et al.[13]; Herzog et al.[14]; Kalichman et al.[15]) of the spinal canal was measured as the mid-sagittal diameter of the thecal sac. Four different measurement sites were selected in accordance with the literature and measured in axial T_2 images: (1) at mid-height of the inferior vertebra of the motion segment in question, (2) at mid-height of the superior vertebra of the motion segment (Fig. 1a), (3) at the level of the intervertebral disc (in the case of reduced disc height, sometimes only images of the plane directly above and below the disc were recorded; in this case, the image with a narrower dural sac was measured) (Fig. 1b), and finally (4) at its narrowest height in axial T_2 images.

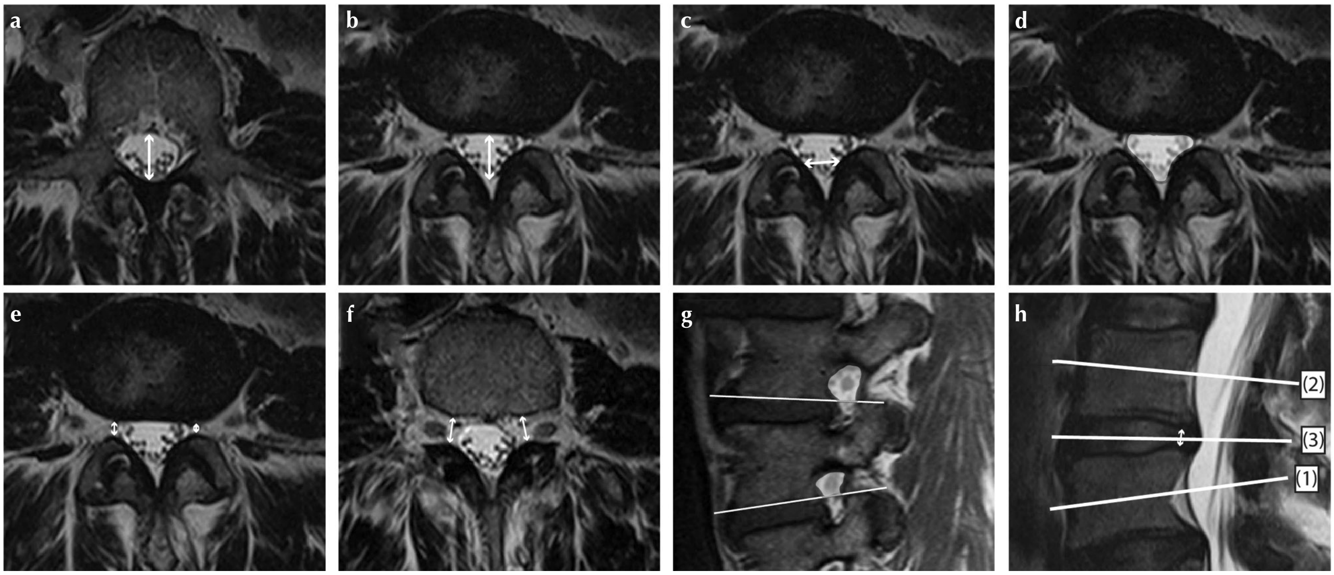


Fig. 1 Quantitatively measured spinal canal parameters. (**a**–**f**) Axial T_2, (**g**) sagittal T_1, and (**h**) sagittal T_2 images. Sagittal diameter of the spinal canal was measured as the mid-sagittal diameter of the thecal sac at mid-vertebrae level (**a**) or at the height of the intervertebral disc (**b**). The measurement levels are also depicted in blue lines in the sagittal T_2 image (**h**): (**1**) at mid-height of the inferior vertebra of the motion segment in question, (**2**) at mid-height of the superior vertebra of the motion segment, and (**3**) at the level of the intervertebral disc (in the case of reduced disc height, sometimes only images of the plane directly above and below the disc were recorded; in this case, the image with a narrower dural sac was measured). (**c**) Ligamentous interfacet distance[14] between the inner surface of the flaval ligaments on a line connecting the ventral joint space of the facet joints at the level of the intervertebral disc. (**d**) The smallest cross-sectional area of the dural sac[16–18] at the level of the intervertebral disc. The lateral border at the level of the neuroforamina was extrapolated from the images above and below. The small arrows depict the antero-posterior diameter of the neural foramen at the level of the intervertebral disc (**e**) and at the level where the root can be seen to traverse it[19] (**f**). (**g**) The cross-sectional area of the neural foramen[20] below the pedicle. No space below the red line parallel to the lower end plate was included in area measurements because of the cranial course of the nerve root. (**h**) Posterior disc height at the central position of the spinal canal in the axial plane.[21]

The ligamentous interfacet distance[14] was measured in axial T_2 images between the inner surface of the flaval ligaments on a line connecting the ventral joint space of the facet joints at the level of the intervertebral disc. If two such images were available, the narrower distance was measured (Fig. 1c).

The smallest cross-sectional area of the dural sac in T_2 axial images was measured within the motion segment in question, i.e., at the level of the intervertebral disc.[16–18] The stenosis ratio,[17] i.e., the cross-sectional area of the motion segment divided by that of the stable segment at pedicle height, was, however, not calculated. With respect to the lateral border of that area at the level of the neuroforamina, no specifications are given in the original literature and thus the hypothetical margin was visually extrapolated from the images above and below (Fig. 1d).

(b) Qualitative Measurement

Using axial T_2 images, we performed qualitative spinal stenosis measurements in accordance with the method of Schizas et al.,[7] taking into account the cerebrospinal fluid/rootlet content and the presence of epidural fat. Grade A is thereby characterized by the ample presence of cerebrospinal fluid. In grade B, the rootlets occupy the whole of the dural sac but they can still be distinguished from one another. When individual rootlets can no longer be seen, spinal stenosis is considered grade C, and in grade D, even the epidural fat posterior to the dural sac is lost. In contrast to the original description, categories A1–4 were combined into one category A (an almost identical classification was suggested by Lee et al.[6]).

Neuroforaminal Stenosis

(a) Quantitative Measurements

The minimum antero-posterior diameter of the neural foramen was measured in axial T_2 images at the level where the location of the intervertebral disc was confirmed in sagittal T_2 images. If there were several images that included the intervertebral disc, the most cranial was used because of the relative cranial transition of the nerve root through the foramen (Fig. 1e).

In accordance with the technique of Beers et al.,[19] the width of the foramen was measured in axial T_2 images for its minimum antero-posterior diameter in the axial plane where the root can be seen to traverse it (Fig. 1f).

The minimum cross-sectional area of the foraminal zone was measured on sagittal T_1 images below the pedicle as suggested by Sipola et al.[20] Because the nerve root is located more cranially than the lower end plate, no space below the line parallel to the lower end plate was included in area measurements (Fig. 1g).

The posterior disc height was measured, as proposed by Hasegawa et al.,[21] in sagittal T_2 images after the central position of the spinal canal in the axial plane was identified (Fig. 1h).

(b) Qualitative Measurement
Using sagittal T_1 images, we qualitatively classified neuroforaminal stenosis in accordance with the system suggested by Lee et al.[8] and Kurogi et al.[22] Stenosis was classified into grades 0–3 at the narrowest point at the medial margin of the pedicle in the subpedicular zone. If there were ambiguous results in T_1, T_2 images were additionally analyzed.

For all neuroforaminal measurements, both the right and left sides were analyzed.

Facet Joint Degeneration, Qualitative Measurement
Facet joint degeneration was classified in accordance with the technique of Pathria et al.[23] and Weishaupt et al.[24] T_2 axial images were analyzed for joint space, osteophyte formation, bone erosions, subchondral cyst formation, and joint hypertrophy and graded from 0 to 3. Grade 0 is allocated when a normal facet joint space (2–4 mm) is present. Narrowing of the joint space <2 mm and/or small osteophyte formation or articular process hypertrophy is classified as grade 1. Grade 2 degeneration means a clear narrowing of the facet joint space and/or moderate osteophytes and/or moderate hypertrophy of the articular process and/or mild subarticular bone erosions. End-stage facet joint degeneration with loss of the facet joint space and/or large osteophytes and/or severe hypertrophy of the articular process and/or severe subarticular bone erosions and/or subchondral cysts are classified as grade 4.

In addition, the presence of intra-articular fluid was denoted in axial T_2 images.

Imaging software and MRI scans

MR images were acquired with the 3 Tesla Siemens Skyra, or the 1.5 Tesla Siemens Aera, Avanto, and Espree (Siemens Healthcare, Erlangen, Germany). Imaging was performed angled to the disc level. All MRIs were available in digital form and analyzed with a centricity picture archiving and communication systems (PACS) Centricity RA1000 workstation (GE Healthcare, Barrington, IL, USA) on an RadiForce RS110 48 cm class Color LCD screen (Eizo Corporation, Ishikawa, Japan). Quantitative measurements were performed using the software-integrated measurement tools.

Statistical analysis

Distributions of variables within the study groups were assessed by histograms. For interval scaled variables, agreement was calculated as the intraclass correlation coefficient (ICC) 3.1 as two-way mixed, absolute agreements with single measures between the three observers and as direct comparison between each pair of observers. For ordinal and nominal scaled variables, Cohen's ordinal or nominal kappa was calculated, as appropriate. According to Landis and Koch, a kappa of 0–0.2 was considered slight agreement, 0.21–0.4 fair, 0.41–0.6 moderate, 0.61–0.8 substantial, and 0.81–1 excellent agreement.[25] Confidence intervals were formed using analysis of variance methods for estimating intraclass correlations.[26] Graphic illustration of the results was performed using bar diagrams for nominal scaled variables, heat maps for ordinal scaled variables, and Bland–Altman plots for interval scaled variables. Statistical analysis was conducted with IBM SPSS Statistics, version 22 (IBM Corporation, Armonk, NY, USA). For reporting of results, the medical student is labeled as observer 1, the spine surgeon responsible for training the other two as observer 2, and the trained spine surgeon as observer 3.

Results

Quantitative spinal canal parameters showed excellent interobserver agreements across all three observers with an ICC (3.1) of 0.889 for the sagittal diameter at the narrowest point of the motion segment, an ICC (3.1) of 0.961 for the ligamentous interfacet distance, and an ICC of 0.888 for the cross-sectional area. This sagittal diameter measured at other levels, however, showed only substantial agreement (Table 1, Fig. 2a–2a″ and Supplementary Fig. 1). Neuroforaminal quantitative measurements showed much weaker correlations, with only a fair correlation of 0.369 for the foraminal width measured at the level of the intervertebral disc and a moderate ICC (3.1) of 0.557 for the width in accordance with the method of Beers et al.[19] The best interobserver agreement was found for the neuroforaminal cross-sectional area with an ICC (3.1) of 0.712 (Table 2). A similar agreement was found for the posterior disc height (0.701) (Table 2 and Fig. 2b–2b″). An interesting feature observed in all of these measurements was the generally excellent agreement between observers 1 and 2, whereas the correlation between observer 3 and the other two observers was much weaker. The same observation could be made for the nominally scaled facet joint fluid, which showed an excellent kappa of 0.833 between observers 1 and 2, while agreement between observers 1 and 3 was 0.474 and that between observers 2 and 3 was 0.595 (Table 3). For the ordinally scaled qualitative variables, excellent agreement was observed between observers 1 and 2, ranging from a kappa of 0.873 for facet joint degeneration to 0.909 for spinal stenosis, in accordance with the classification of Schizas et al.,[7] while the correlation between the other observer combinations was only moderate or good (Table 3, Fig. 3 and Supplementary Fig. 2). With the exception of the sagittal spinal canal diameter, the qualitative spinal stenosis classification of Schizas et al.,[7] and intra-articular fluid determination, all variables had at least one patient who was not measured by at least one observer due to insufficient morphological discrimination in the MRI of the necessary landmarks.

Table 1 Intraclass correlation coefficient (ICC) 3.1 for spinal canal measurements in axial T_2 images

Observer	ICC (3.1)	95% CI
Sagittal diameter of the spinal canal at the narrowest point of the motion segment		
Three observers (*n* = 29)	0.889	0.733–0.951
Observers 1 and 2 (*n* = 30)	0.979	0.957–0.990
Observers 1 and 3 (*n* = 29)	0.834	0.477–0.935
Observers 2 and 3 (*n* = 29)	0.857	0.482–0.947
Sagittal diameter of the spinal canal at the level of the intervertebral disc		
Three observers (*n* = 30)	0.700	0.337–0.865
Observers 1 and 2 (*n* = 30)	0.980	0.958–0.990
Observers 1 and 3 (*n* = 30)	0.582	0.010–0.826
Observers 2 and 3 (*n* = 30)	0.601	0.018–0.837
Sagittal diameter of the spinal canal at mid-vertebrae level above the stenosis at the intervertebral disc		
Three observers (*n* = 30)	0.805	0.404–0.925
Observers 1 and 2 (*n* = 30)	0.973	0.944–0.987
Observers 1 and 3 (*n* = 30)	0.601	0.025–0.836
Observers 2 and 3 (*n* = 30)	0.738	0.001–0.916
Sagittal diameter of the spinal canal at mid-vertebrae level below the stenosis at the intervertebral disc		
Three observers (*n* = 29)	0.715	0.374–0.871
Observers 1 and 2 (*n* = 29)	0.974	0.945–0.988
Observers 1 and 3 (*n* = 30)	0.568	0.023–0.814
Observers 2 and 3 (*n* = 29)	0.638	0.061–0.856
Ligamentous interfacet distance as suggested by Carrino et al.[27]		
Three observers (*n* = 29)	0.961	0.930–0.980
Observers 1 and 2 (*n* = 29)	0.975	0.948–0.988
Observers 1 and 3 (*n* = 30)	0.945	0.887–0.974
Observers 2 and 3 (*n* = 29)	0.955	0.908–0.979
Cross-sectional area of the spinal canal[24–26]		
Three observers (*n* = 28)	0.888	0.789–0.945
Observers 1 and 2 (*n* = 30)	0.972	0.942–0.986
Observers 1 and 3 (*n* = 28)	0.842	0.636–0.929
Observers 2 and 3 (*n* = 28)	0.863	0.697–0.937

Discussion

Similar to laboratory tests or histopathologic findings, imaging results are considered solid evidence by physicians, patients, and others, for e.g., insurance companies. Given the uncertainty surrounding lumbar spinal stenosis classification and its clinical relevance, solid evidence needs to be established to justify and merit the anticipated confidence in this technique. This is even more the case because reference values with respect to interobserver reliability are usually based on agreement between experts in the field, which does not reflect the reality of medical teaching during residency.

In our study, we found excellent overall interobserver reliability for some quantitative and qualitative MRI features, such as ligamentous interfacet distance, cross-sectional area, and sagittal diameter at the narrowest point of the motion segment. Nevertheless, other important parameters showed only moderate interobserver reliability between the trained medical student and the spine surgeon after standardized instructions, as well as between the two spine surgeons. Notably, values between the medical student with no clinical and surgical experience and the spine surgeon who had created the reference measurements were consistently substantial or excellent.

This study shows a relevant increase in interobserver reliability through specific training of the observer. This finding suggests that the sole use of classifications without the help of reference measurements does not lead to optimal interobserver measurements of spinal images, even in experienced clinicians. Interestingly, this aspect of interobserver

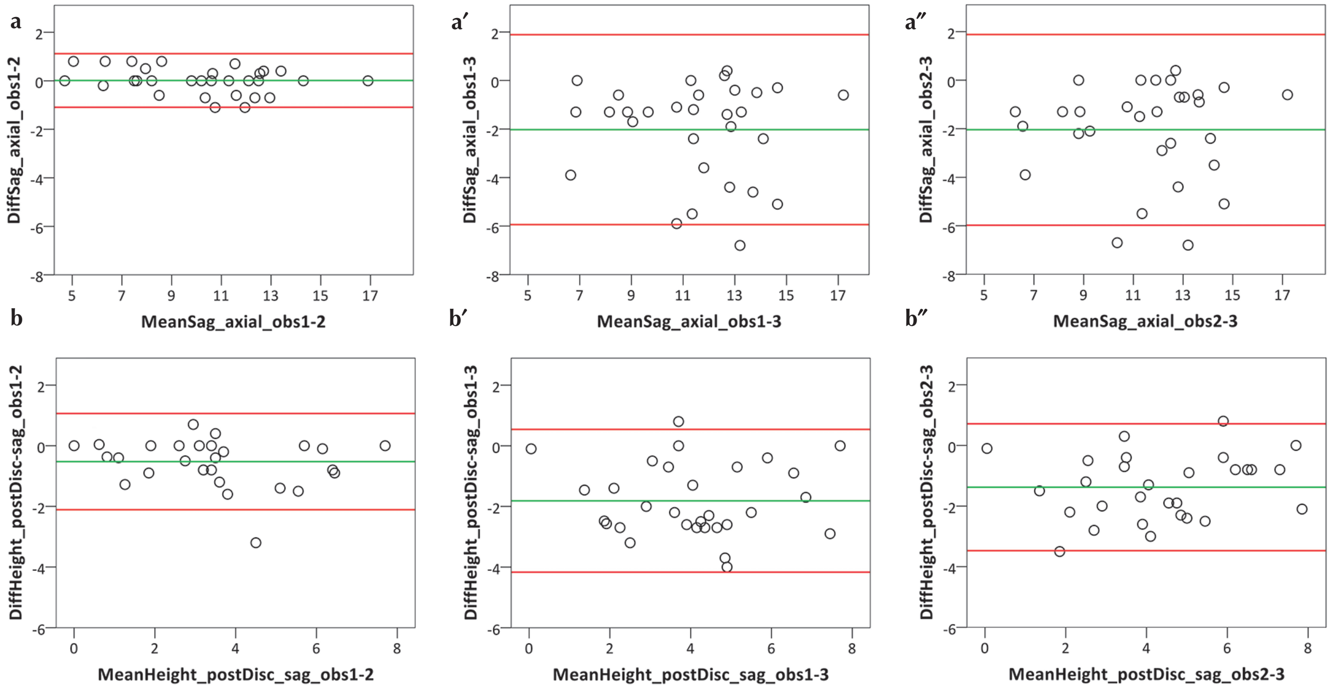


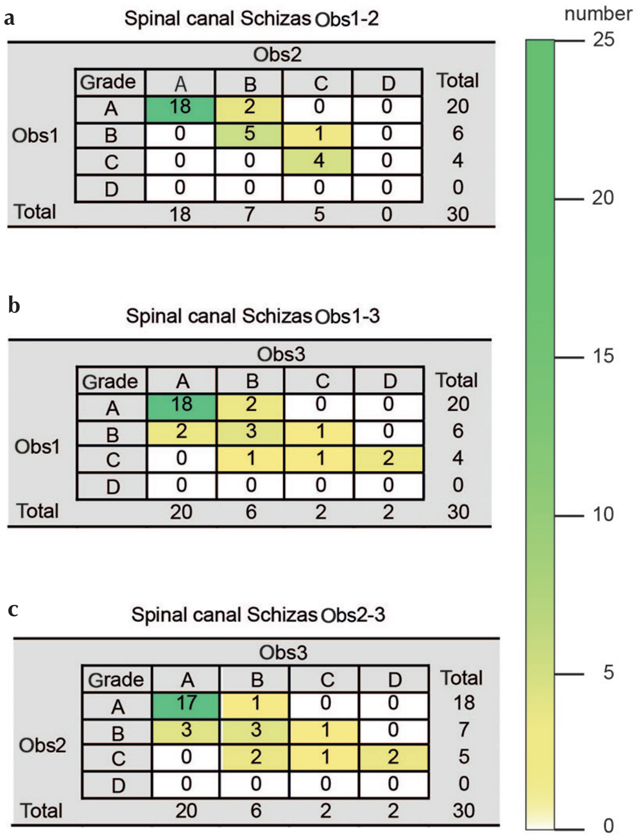
Fig. 2 Interobserver agreement displayed in the form of Bland–Altman plots. (**a**–**a″**) Agreement for the sagittal diameter of the spinal canal in the axial T_2 plane at its narrowest point, and (**b**–**b″**) for posterior disc height. (**a** and **b**) Comparison of findings of observers 1 and 2, (**a′** and **b′**) observers 1 and 3, and (**a″** and **b″**) observers 2 and 3.

Table 2 Intraclass correlation coefficient (ICC 3.1) for neuroforaminal measurements

Observer	ICC (3.1)	95% CI
Neuroforaminal width in axial T_2 images at the level of the intervertebral disc		
Three observers (*n* = 56)	0.369	0.046–0.621
Observers 1 and 2 (*n* = 57)	0.611	0.212–0.800
Observers 1 and 3 (*n* = 56)	0.302	−0.082–0.594
Observers 2 and 3 (*n* = 57)	0.240	−0.096–0.557
Neuroforaminal width in axial T_2 images as proposed by Beers et al.[19]		
Three observers (*n* = 58)	0.557	0.406–0.690
Observers 1 and 2 (*n* = 58)	0.787	0.666–0.867
Observers 1 and 3 (*n* = 58)	0.362	0.127–0.561
Observers 2 and 3 (*n* = 58)	0.530	0.297–0.699
Neuroforaminal cross-sectional area in sagittal T_1 images in accordance with the technique of Sipola et al.[20]		
Three observers (*n* = 55)	0.712	0.420–0.850
Observers 1 and 2 (*n* = 55)	0.903	0.840–0.942
Observers 1 and 3 (*n* = 56)	0.616	0.056–0.831
Observers 2 and 3 (*n* = 56)	0.672	0.173–0.853
Posterior height of the intervertebral disc in sagittal T_2 images[29]		
Three observers (*n* = 29)	0.701	0.214–0.882
Observers 1 and 2 (*n* = 29)	0.892	0.686–0.956
Observers 1 and 3 (*n* = 30)	0.550	−0.095–0.837
Observers 2 and 3 (*n* = 29)	0.700	−0.029–0.901

Table 3 Interobserver Cohen's kappa for nominally and ordinally scaled variables

Observer	Kappa
Spinal stenosis in axial T_2 images in accordance with the classification of Schizas et al.[7]	
Observers 1 and 2 (n = 30)	0.909
Observers 1 and 3 (n = 30)	0.795
Observers 2 and 3 (n = 30)	0.780
Neuroforaminal stenosis in sagittal T_1 images as proposed by Lee et al.[8]	
Observers 1 and 2 (n = 56)	0.894
Observers 1 and 3 (n = 56)	0.738
Observers 2 and 3 (n = 56)	0.741
Facet joint degeneration in axial T_2 images classified as suggested by Weishaupt et al.[24]	
Observers 1 and 2 (n = 60)	0.873
Observers 1 and 3 (n = 58)	0.457
Observers 2 and 3 (n = 58)	0.473
Intra-articular fluid of facet joints in axial T_2 images	
Observers 1 and 2 (n = 60)	0.833
Observers 1 and 3 (n = 60)	0.474
Observers 2 and 3 (n = 60)	0.595



Spinal canal Schizas Obs1-2

	Obs2					
	Grade	A	B	C	D	Total
Obs1	A	18	2	0	0	20
	B	0	5	1	0	6
	C	0	0	4	0	4
	D	0	0	0	0	0
Total		18	7	5	0	30

Spinal canal Schizas Obs1-3

	Obs3					
	Grade	A	B	C	D	Total
Obs1	A	18	2	0	0	20
	B	2	3	1	0	6
	C	0	1	1	2	4
	D	0	0	0	0	0
Total		20	6	2	2	30

Spinal canal Schizas Obs2-3

	Obs3					
	Grade	A	B	C	D	Total
Obs2	A	17	1	0	0	18
	B	3	3	1	0	7
	C	0	2	1	2	5
	D	0	0	0	0	0
Total		20	6	2	2	30

Fig. 3 Agreement for the ordinally scaled classification of spinal stenosis as suggested by Schizas et al.[7] (**a**) Comparison of findings of observers 1 and 2, (**b**) observers 1 and 3, and (**c**) observers 2 and 3.

reliability and training has seldom been investigated in the field of spine image interpretation. Carrino et al.[27] and Lurie et al.,[28] for e.g., investigated interobserver kappas between three musculoskeletal radiologists and one orthopedic surgeon with cumulative professional experience of over 87 years. Both studies describe interobserver agreements for various parameters of spinal stenosis comparable to those in the present study. They do, not, however, evaluate the different skill levels and training of the observer. Given the results of our study, we propose that before using such classification systems for medical or scientific purposes, reliability ought to be evaluated and possibly improved by adjusting the measurements to predefined standard measurements. Bearing this important aspect in mind is also essential when comparing results of different studies originating from different institutions. Leading spine organizations such as the North American Spine Society (NASS) or the Spine Society of Europe (EUROSPINE) could facilitate this process and improve inter-rater reliability by providing databases with consensus reference measurements for relevant spinal parameters.

Another important question is whether qualitative or quantitative measurements systems are superior. From a statistical standpoint, the correlation values obtained from qualitative measurements cannot be directly compared with those from quantitative measurements because the former are ordinal scaled and the latter interval scaled. Different statistical tests thus apply, which makes direct comparison impossible. Notably, however, despite the fact that qualitative measurements bear a more prominent subjective factor than simple distance or surface measurements do, the qualitative classifications also yield excellent results. At the same time, they offer additional information independent of patient size or anatomy. It thus stands to reason that, especially when trying to apply measurement systems across the globe, qualitative measurement systems are to be preferred over simple quantitative evaluation. One recent direct comparison showed no relevant difference between cross-sectional area measurement of the dural sac and morphological classification of spinal stenosis regarding their association with patient-rated outcome of spinal surgery for stenosis in 157 patients.[29] The authors did, however, spell out the advantages of morphological grading in the sense that no specific tools are required and classification can be performed in an instant. Yet, the literature on this topic is still scarce and further research is warranted.

One limitation of the present study is that intrarater reliability was not investigated; this was not, however, our goal. Furthermore, not all available classification systems could be tested.

Conclusion

For surgeons to keep providing the best medical care, continuous evaluation of established and new diagnostic and therapeutic procedures is essential. Given the complexity of our profession, such evaluation can be performed successfully only if highly reliable imaging interpretation is the standard. Our study suggests that good familiarisation by means of using reference measurements leads to much higher reliability in comparison to standardized instructions for experienced professionals. We encourage leading spine organizations such as the NASS or EUROSPINE to provide databases with consensus reference measurements for relevant spinal parameters to improve future research outcomes in the field of spinal disease and treatment.

Author Contribution

UKH designed the study, performed the statistical analyzes, and wrote the manuscript; RLK performed the measurements and helped with the statistical analyzes; MG performed the measurements; CW helped write the manuscript; and FM performed the measurements and wrote the manuscript. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

Full departmental, institutional, and local ethical committee approval were obtained before commencement of the study (project number 503/2016BO2).

Acknowledgment

We thank Barbara Every, ELS of BioMedical Editor, for English language editing.

Conflicts of Interest

The authors declare that they have no competing interests.

Supplementary Information

Supplementary Figs. 1 and 2 are available online

Supplementary Fig. 1

Interobserver agreement for the cross-sectional area of the spinal canal in T_2 axial planes, displayed in the form of Bland-Altman plots. (**A**) Comparison of findings of observers 1 and 2, (**B**) observers 1 and 3, and (**C**) observers 2 and 3.

Supplementary Fig. 2

Agreement for ordinally scaled variables displayed as heat maps. Classification of (**A**) neuroforaminal stenosis as suggested by Lee et al.,[8] and (**B**) facet joint degeneration as suggested by Weishaupt et al.[24] (**A** and **B**) Comparison of findings of observers 1 and 2, (**A′** and **B′**) observers 1 and 3, and (**A″ and B″**) observers 2 and 3.

References

1. Deyo RA, Gray DT, Kreuter W, Mirza S, Martin BI. United States trends in lumbar fusion surgery for degenerative conditions. Spine (Phila Pa 1976) 2005; 30:1441–1445; discussion 1446–1447.
2. Katz JN, Harris MB. Clinical practice. Lumbar spinal stenosis. N Engl J Med 2008; 358:818–825.
3. Andreisek G, Hodler J, Steurer J. Uncertainties in the diagnosis of lumbar spinal stenosis. Radiology 2011; 261:681–684.
4. Steurer J, Roner S, Gnannt R, Hodler J, LumbSten Research Collaboration. Quantitative radiologic criteria for the diagnosis of lumbar spinal stenosis: a systematic literature review. BMC Musculoskelet Disord 2011; 12:175.
5. Lohman CM, Tallroth K, Kettunen JA, Lindgren KA. Comparison of radiologic signs and clinical symptoms of spinal stenosis. Spine (Phila Pa 1976) 2006; 31:1834–1840.
6. Lee GY, Lee JW, Choi HS, Oh KJ, Kang HS. A new grading system of lumbar central canal stenosis on MRI: an easy and reliable method. Skeletal Radiol 2011; 40:1033–1039.
7. Schizas C, Theumann N, Burn A, et al. Qualitative grading of severity of lumbar spinal stenosis based on the morphology of the dural sac on magnetic resonance images. Spine (Phila Pa 1976) 2010; 35:1919–1924.
8. Lee S, Lee JW, Yeom JS, et al. A practical MRI grading system for lumbar foraminal stenosis. AJR Am J Roentgenol 2010; 194:1095–1098.
9. Andreisek G, Deyo RA, Jarvik JG, et al. Consensus conference on core radiological parameters to describe lumbar stenosis - an initiative for structured reporting. Eur Radiol 2014; 24:3224–3232.
10. Jones A, Clarke A, Freeman BJ, Lam KS, Grevitt MP. The Modic classification: inter- and intraobserver error in clinical practice. Spine (Phila Pa 1976) 2005; 30:1867–1869.
11. Fayad F, Lefevre-Colau MM, Drapé JL, et al. Reliability of a modified Modic classification of bone marrow changes in lumbar spine MRI. Joint Bone Spine 2009; 76:286–289.
12. Verbiest H. The significance and principles of computerized axial tomography in idiopathic developmental stenosis of the bony lumbar vertebral canal. Spine (Phila Pa 1976) 1979; 4:369–378.
13. Ullrich CG, Binet EF, Sanecki MG, Kieffer SA. Quantitative assessment of the lumbar spinal canal by computed tomography. Radiology 1980; 134:137–143.
14. Herzog RJ, Kaiser JA, Saal JA, Saal JS. The importance of posterior epidural fat pad in lumbar central canal stenosis. Spine (Phila Pa 1976) 1991; 16:S227–S233.
15. Kalichman L, Cole R, Kim DH, et al. Spinal stenosis prevalence and association with symptoms: the Framingham Study. Spine J 2009; 9:545–550.
16. Hamanishi C, Matukura N, Fujita M, Tomihara M, Tanaka S. Cross-sectional area of the stenotic lumbar dural tube measured from the transverse views of magnetic resonance imaging. J Spinal Disord 1994; 7:388–393.
17. Laurencin CT, Lipson SJ, Senatus P, et al. The stenosis ratio: a new tool for the diagnosis of degenerative spinal stenosis. Int J Surg Investig 1999; 1:127–131.
18. 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; 7:115–121.
19. Beers GJ, Carter AP, Leiter BE, Tilak SP, Shah RR. Interobserver discrepancies in distance measurements from lumbar spine CT scans. AJR Am J Roentgenol 1985; 144:395–398.
20. Sipola P, Leinonen V, Niemeläinen R, et al. Visual and quantitative assessment of lateral lumbar spinal canal

stenosis with magnetic resonance imaging. Acta Radiol 2011; 52:1024–1031.

21. Hasegawa T, An HS, Haughton VM, Nowicki BH. Lumbar foraminal stenosis: critical heights of the intervertebral discs and foramina. A cryomicrotome study in cadavera. J Bone Joint Surg Am 1995; 77:32–38.
22. Kunogi J, Hasue M. Diagnosis and operative treatment of intraforaminal and extraforaminal nerve root compression. Spine (Phila Pa 1976) 1991; 16:1312–1320.
23. Pathria M, Sartoris DJ, Resnick D. Osteoarthritis of the facet joints: accuracy of oblique radiographic assessment. Radiology 1987; 164:227–230.
24. Weishaupt D, Zanetti M, Boos N, Hodler J. MR imaging and CT in osteoarthritis of the lumbar facet joints. Skeletal Radiol 1999; 28:215–219.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159–174.
26. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86:420–428.
27. Carrino JA, Lurie JD, Tosteson AN, et al. Lumbar spine: reliability of MR imaging findings. Radiology 2009; 250:161–170.
28. Lurie JD, Tosteson AN, Tosteson TD, et al. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. Spine (Phila Pa 1976) 2008; 33:1605–1610.
29. Mannion AF, Fekete TF, Pacifico D, et al. Dural sac cross-sectional area and morphological grade show significant associations with patient-rated outcome of surgery for lumbar central spinal stenosis. Eur Spine J 2017; 26:2552–2564.