European Journal of Radiology 114 (2019) 45–50

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Qualitative versus quantitative lumbar spinal stenosis grading by machine learning supported texture analysis—Experience from the LSOS study cohort

Florian A. Huber[a,*,1], Shanon Stutz[a,1], Ilaria Vittoria de Martini[a], Manoj Mannil[a], Anton S. Becker[a], Sebastian Winklhofer[b], Jakob M. Burgstaller[c], Roman Guggenberger[a]

[a] Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, Zurich, Switzerland
[b] Department of Neuroradiology, University Hospital Zurich, Zurich, Switzerland
[c] Horten Centre for Patient Oriented Research And Knowledge Transfer, University of Zurich, Zurich, Switzerland

ARTICLE INFO

Keywords:
Lumbar spinal stenosis
Machine learning
Texture analysis
Magnetic resonance imaging

ABSTRACT

Purpose: To investigate and compare the reproducibility and accuracy of qualitative ratings and quantitative texture analysis (TA) in detection and grading of lumbar spinal stenosis (LSS) in magnetic resonance imaging (MR) scans of the lumbar spine.

Materials and methods: From a nationwide multicenter and multidisciplinary lumbar stenosis outcome study (LSOS) register 82 patients, undergoing MR scans of the lumbar spine due to clinical indication of spinal claudication, with a single level central or lateral severe LSS were included. In total 343 transaxial T2-weighted images of the lumbar spine were included from one to five levels (L1 to S1) per patient. One expert radiologist serving as reference standard rated LSS grade according to a standard four-point (normal to severe) as well as to an eight-point Schizas grading scale. DICOM data were then rescaled to a defined pixel size. Two independent readers performed qualitative ratings analogous to expert reader in addition to TA of spinal canals by manually placing two regions of interest (ROI) per image reflecting qualitative scales: (1) dural sac only (2) inner contour of the spinal canal including epidural fat and bilateral recesses. Interreader agreements of qualitative and quantitative parameters were assessed by Cohen's Kappa (κ) and intraclass correlation (ICC), respectively. TA feature reduction was performed by ICC threshold > 0.75. Remaining features were analyzed with machine learning algorithms (Weka 3 tool) for correlation with LSS grades using 10-fold cross validation.

Results: Qualitative ratings showed only moderate reproducibility for both LSS classification systems but high correlation with cut-off cross-sectional area (CSA) < 130mm^2 for severe spinal stenosis. In quantitative TA of both ROIs, machine learning analysis with a decision tree classifier revealed higher performances for LSS grading compared to qualitative assessments using the reference CSA cut-off, respectively.

Conclusion: Qualitative LSS grading independent of classification system shows moderate reproducibility. TA with machine learning offers highly reproducible quantitative parameters that increase accuracy for severe LSS detection with minor impact of grading score and CSA border definition.

1. Introduction

Central lumbar spinal stenosis (LSS) is defined as a condition of narrowing of the central spinal canal and/or the lateral recesses with consequent compression of nerves originating from the lumbar spine. It is frequently seen in older patients and results in the majority of cases from degenerative changes [1]. However, the pathogenesis of the condition is multifactorial and may be triggered by herniation of intervertebral discs, tumor masses, trauma or osteoporotic bone architecture [2,3]. Over the last years, degenerative causes of severe LSS have become the number one indication for spinal surgery in elderly individuals [4,5] and thus imaging plays an increasingly important role in identification of severe LSS.

Magnetic resonance imaging (MR) is generally indicated if specific clinical LSS symptoms, such as neurogenic claudication, are present [6]. However, diagnosis is still challenging, as consensus has not yet been found regarding criteria nor severity of LSS. In a systematic review, Steurer et al. have investigated the spectrum of both qualitative and quantitative imaging criteria for LSS quantification [7]. According to this study, in addition to inter-study heterogeneity of stenosis labeling,

[*] Corresponding author at: University Hospital Zurich, Institute of Diagnostic and Interventional Radiology, Rämistrasse 100, 8091 Zurich, Switzerland.
E-mail address: florian.huber@usz.ch (F.A. Huber).
[1] These authors are co-first authors.

https://doi.org/10.1016/j.ejrad.2019.02.023
Received 9 October 2018; Received in revised form 19 December 2018; Accepted 18 February 2019

Table 1
Severity of lumbar spine stenosis (LSS) according to standard score (StSc) and to Schizas score (SchSc) grades.

n = 337	normal		mild		moderate		severe	
StSc	66 (19.6 %)		191 (56.7 %)		26 (7.7 %)		54 (16.0 %)	
SchSc	normal	A1	A2	A3	A4	B	C	D
	64 (19.0 %)	75 (22.3 %)	77 (22.8 %)	43 (12.8 %)	1 (0.3 %)	21 (6.2 %)	48 (14.2 %)	8 (2.4 %)

there is also a poor interreader agreement for qualitative criteria in LSS. This is also emphasized in several other publications addressing the need for standardized and reproducible stenosis grading systems [8–11].

Furthermore, in order to overcome inconsistencies from subjective grading scales several studies have investigated the value of quantitative measures for LSS grading. Among those the cross-sectional area (CSA) of the spinal canal measured on transaxial images seems to deliver a decent diagnostic performance for severe LSS detection using a cut-off area of 130 mm^2 [12–14]. However, there are different definitions of the CSA of the stenosed segment, e.g. CSA encircling the dural sac vs. the entire spinal canal including epidural fat and bilateral recesses. Despite some qualitative LSS grading systems that have addressed that ambiguity by introducing additional subcategories for epidural fat effacement [15], LSS grading based on quantitative CSA measurements does not only depend on contour definitions but can be additionally influenced by transaxial angulation bias, inter-scanner and/or scan protocol heterogeneity.

Texture analysis (TA) as part of recent ambitions to extract quantitative information from medical images makes use of previously (manually or automatically) segmented image data and evaluates quantifiable, continuous parameters with high reproducibility [16]. As increasing LSS severity leads to increased crowding of the cauda equina with the lumbar nerves in the spinal canal, significant texture patterns may be associated with this process. TA parameters may potentially be less susceptible to image and CSA measurement heterogeneity and help to define LSS severity in a more robust way obviating inaccuracies and inconsistencies between different human readers [17,18].

The aim of this study was to investigate and compare the reproducibility and accuracy of qualitative ratings and quantitative texture analysis (TA) in detection and grading of lumbar spinal stenosis (LSS) in MR scans of the lumbar spine, using a quantitative reference standard (CSA).

2. Materials & methods

2.1. Study population

From a nationwide multicenter and multidisciplinary study register (Swiss Lumbar Stenosis Outcome Study, LSOS) 82 patients who received MR scans due to clinical indications of spinal claudication with a single level severe central (including bilateral recesses) LSS were identified. Additional inclusion criteria were predefined during inclusion into the LSOS cohort, such as age over 50 years, uni- and/or bilateral neurogenic claudication, no spinal operation or infiltration preceding imaging studies and ability of patient to give informed consent. Exclusion criteria of the cohort were cauda equina syndrome requiring urgent surgery, acute fractures, infections, peripheral arterial disease, prior epidural injections and severe scoliosis of the lumbar spine (> 15°). Furthermore, all patients with severe neuroforaminal stenoses were excluded in order to optimally homogenize the study cohort with regard to LSS cause.

The cohort study was approved by the responsible local ethics committee (Ethics Committee Zurich) and performed in accordance with the Declaration of Helsinki.

2.2. Image extraction & expert rating

From the 82 MR scans of the lumbar spine, 343 images were used for further analysis. Imaging data of the lumbar canal included at least one up to five transaxial T2-weighted MR image stacks (L1 to S1) of each individual patient. L1/2 and L2/3 was not available for all patients (full stacks of 37 patients), which is why only 343 images were included. Representative images at the intervertebral discs covering the spinal canal and the lateral recesses were identified and then exported from the image series as single DICOM images, respectively. All retrieved images were rescaled to the "finest" spacing of 0.39 mm per pixel with bicubic method, using the Imaging Processing Toolbox for Matlab (Version R2016b, The MathWorks, Inc., Natick, MA, USA).

2.3. Qualitative analysis

One expert radiologist with over ten years of musculoskeletal (MSK) experience, who was blinded to image-related patient data, rated the images for LSS according to a four-point standard score (StSc, 0 = normal, 1 = mild, 2 = moderate, 3 = severe, [9,19]) as well as to the eight-point Schizas score (SchSc; see Table 1, [15]) further sub-classifying epidural fat effacement in severe stenosis description. Two junior readers (two years of MSK experience each), blinded to clinical patient information, to each other's and to the expert's grading, performed qualitative readout in analogy to the expert readout. All images were presented in random order.

2.3.1. Interreader agreement

Interreader agreement of all junior's and expert's qualitative scores including dichotomized StSc (2PStSc, grades 0–2 vs. 3) and SchSc (2PSchSc, differentiating normal-B from C–D) were calculated using Cohen's Kappa (κ). Levels of agreement were defined for specific κ-values (< 0 = poor, 0–.2 = slight, .21–.4 = fair, .41–.6 = moderate, .61–.8 = substantial and .81–1 = almost perfect agreement, respectively), according to Landis & Koch [20].

2.3.2. Validation of reference standard CSA

In order to compare our study cohort and acquired measurements to recommended reference values, qualitative expert scores of LSS for both grading scores (StSc and SchSc) and regions of interest (ROIs, small and large, placed for TA measurements; see below) were validated against dichotomous quantitative reference standard (CSA 130 mm^2) for severe LSS. In addition, CSA values of both ROIs were validated against the qualitative expert categorization of LSS, using 2PStSc and 2PSchS. CSA cutoff values (γ-values) with optimal sensitivity and specificity were derived to differentiate severe from non-severe LSS based on qualitative expert scores.

2.4. Texture analysis

After thorough instruction both junior readers performed TA by manually placing two ROIs reflecting different qualitative scores on each image: encircling (1) the dural sac only and (2) the inner contour of the spinal canal including epidural fat and bilateral recesses (Fig. 1). The segmentation was performed with the open-source software MaZda (Version 4.6) [21]. All automatically exported parameters (concordant with default software preferences and performing normalization

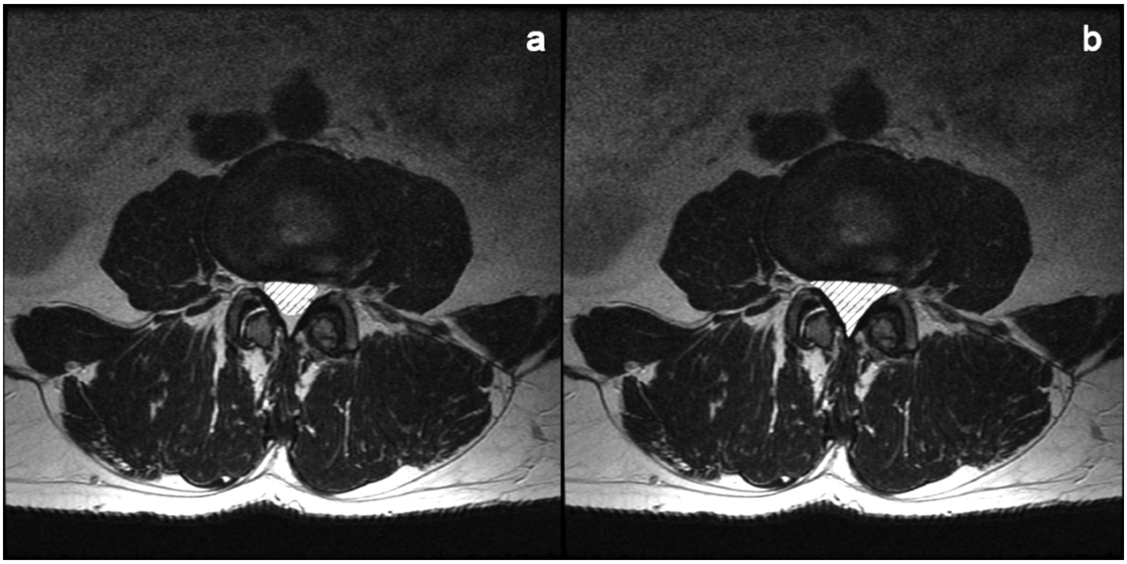

Fig. 1. Transverse T2-weighted image of the lumbar spine at level L3/4. Illustration of the two regions of interest (ROIs) differing in size as used for cross sectional area (CSA) and texture analysis (TA) calculations: the left image (a) represents the small ROI delineating the dural sac, right image (b) shows the large ROI including epidural fat and bilateral recesses.

Table 2
Interreader Cohen's Kappa for different qualitative scores.

	inter-junior	junior- expert
StSc	.325	.355
SchSc	.606	.541
2PStSc	.752	.728
2PSchSc	.827	.739

with ± 3 sigma) were included into primary feature output and further edited in Excel and SPSS (MS Office 2016, Microsoft Corporation, Redmond, WA, USA; IBM SPSS Statistics 25, IBM Corporation, Armonk, NY, USA).

2.5. Feature selection & statistical analysis

In concordance with literature, CSA values of both (small and large) ROIs were dichotomized to a cutoff CSA of 130 mm^2 as reference standard [12–14], indicating severe LSS below that value. TA data originating from both ROIs were tested separately. A total of 304 TA parameters were received and tested for interreader agreement between the two junior readers using intraclass correlation (ICC). Consequently, parameters with ICC-values below 0.75 were discarded in order to base analysis on perfect interreader agreement (according to Cicchetti [22]). Data from the better performing junior reader were used for further analysis, in order to exclude a bias due to insufficient radiological experience. Due to the use of CSA as outcome parameter, we decided to primarily exclude "area" as TA parameter and furthermore apply this rule to all highly redundant area-derived features (Pearson's $r > .995$) to minimize area-dependence of TA. TA parameters were analyzed using machine learning and several classification tools for correlations with severe LSS grades based on reference standard. Data analysis was performed with 10-fold cross validation (no split data and training sets) using the open-source tool Weka (Waikato Environment for Knowledge Analysis, Version 3, University of Waikato, Hamilton, New Zealand) [23,24]. For all statistical tests, a p-value of ≤ 0.05 was considered significant.

3. Results

The average patient age was 74.84 ± 9.22 years (mean ± standard deviation, SD), ranging from 53 to 94 years. A total of 186 images were included from 46 male vs. 157 images from 36 female individuals.

3.1. Qualitative analysis

3.1.1. Interreader agreement

Results from the expert reading showed a distribution of LSS scores as shown in Table 1. Qualitative ratings showed fair (StSc) and moderate (SchSc) interreader agreements between the two junior readers and comparing the better, i.e. more experienced, resident against the expert, respectively. The highest interreader agreements were "substantial" for the dichotomized 2PStSc and 2PSchSc (see Table 2).

3.1.2. Validation of reference standard CSA

Testing of qualitative expert ratings against dichotomous reference standard (CSA 130 mm^2) showed higher area under the curve (AUC) of receiver operator characteristic (ROC)-curve for the larger ROI (AUC = .916 and .799 for StSc and SchSc) as compared to the smaller ROI (AUC = .837 and .722 analogously, see Fig. 2) with better performance of the StSc compared to SchSc in both approaches. Respective performance descriptors are given in Table 3.

CSA-values from both small and large ROIs tested against qualitative expert categorization of severe vs. non-severe LSS showed high AUCs for both 2PStSc (AUC = .892 and .910) and 2PSchS (AUC = .932 and .916), as demonstrated in Fig. 3. γ-values were comparable for all calculations (.660 – .756) and showed best results for matching of larger CSA with 2PSchSc ($\gamma = .756$, CSA = 121.53 mm^2). Apart from CSA of small ROIs vs. 2PSchSc (96.66 mm^2, $\gamma = .735$), all optimal CSA cutoff values were found around 130 mm^2 (121.53–135.65).

3.2. Texture analysis

Interreader evaluation between junior readers revealed 263 TA parameters with excellent agreement. These parameters included 23 area-based features, 8 histogram features, 20 run-length matrix derived parameters, 5 autoregressive based model based features as well as 193 co-occurrence matrix parameters and 14 Haar wavelet features, according to the feature categorization by Szczypinski et al [21]. Interparameter correlation was perfect for all area-derived parameters of the same feature category ($r > .995$, $p < .001$) and ICC showed almost

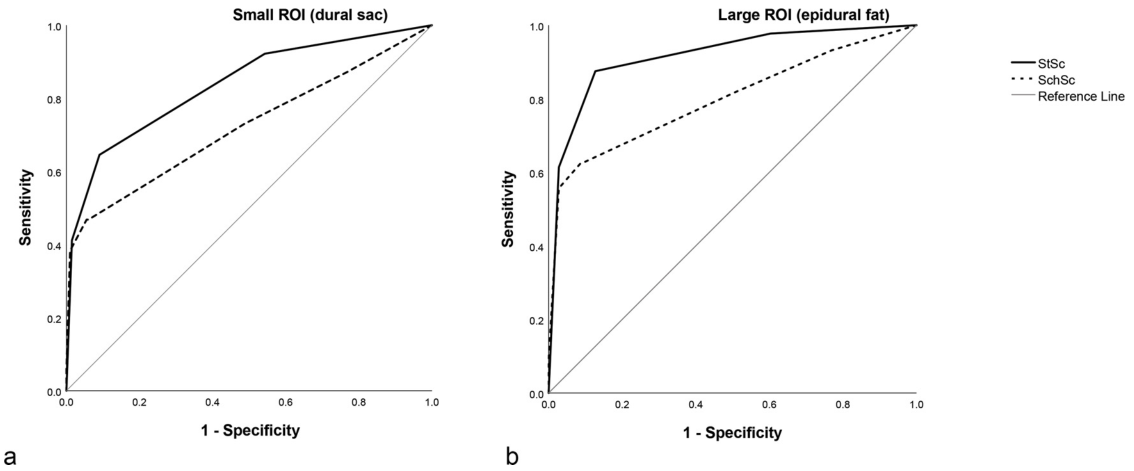

Fig. 2. Receiver operator characteristics (ROC) curves for expert standard score (StSc) and Schizas score (SchSc) ratings for prediction of severe lumbar spine stenosis (LSS) based on quantitative reference standard (cut-off cross sectional area, CSA 130 mm^2) show better performance for StSc ratings but with similar area under curve (AUCs) between (a) small and (b) large regions of interest (ROIs).

Table 3
Diagnostic performance (in %) of quantitative standard (cross-sectional area, CSA 130 mm^2) vs. qualitative expert rating as reference (two-point standard and Schizas score, 2PStSc and 2PSchSc), as well as of texture analysis (TA) vs. quantitative reference standard.

	CSA vs. 2PStSc		CSA vs. 2PSchSc		TA vs. CSA	
	Small ROI	Large ROI	Small ROI	Large ROI	Small ROI	Large ROI
Sensitivity	83.49	70.64	96.43	87.50	94.33	94.32
Specificity	78.63	95.30	69.69	86.41	96.53	98.04
PPV	64.54	87.50	38.30	55.68	95.00	94.32
NPV	91.09	87.45	99.01	97.25	96.06	98.04

perfect interreader agreement for “area” (ICC = .91). Exclusion of all area-based parameters, further analysis and feature selection was eventually conducted with a remainder of 240 TA parameters.

A decision tree classifier (C4.5 algorithm, J48 decision tree, ten-fold cross validation) was applied for solving a classification problem, i.e. dichotomized cutoff CSA above or below 130 mm^2, and revealed significant and strong correlation between TA parameters and the quantitative reference standard. The two decision trees for both ROIs consisted of different single parameters (see Fig. 4b) but did not notably differ in terms of accuracy. ROC curves were almost perfect with better performance descriptors compared to qualitative ratings, e.g. high values of AUC, positive predictive value (PPV) and sensitivity (0.962, 0.971 and 0.971 vs. 0.940, 0.956 and 0.956 for small and large ROI, respectively; Fig. 4a).

4. Discussion

Numerous authors have already attempted to standardize LSS grading suggesting qualitative as well as quantitative scoring schemes, e.g. subjective evaluation of MR features vs. manual geometric measurements [15,25,26]. However, evidence is still weak as to which scoring system offers best reproducibility and correlates best with clinical symptoms [9]. Interreader agreement for qualitative LSS scoring systems in this study showed κ-values of 0.36 and 0.54 for StSc and SchSc among both juniors and an expert reader, respectively. Agreement between both juniors was somewhat higher (0.33 and 0.61) with a small tendency for SchSc but still weak to moderate despite adequate training prior to the study readout. Although κ-values markedly increased above 0.7 when comparing dichotomized scores, i.e. severe vs. non-severe stenosis, perfect agreement and absolute consistency should be the ultimate goal for identification of severe LSS that may have to undergo surgery.

According to literature and in line with these results, qualitative criteria in LSS grading are associated with insufficient interreader agreement. Therefore, quantitative LSS scores such as CSA measurements of the spinal canal have been suggested. The total CSA of the spinal canal as a simple descriptor of LSS grade shows good interreader agreement as also shown in our study (ICC = .91), but is susceptible to different CSA border definitions, image heterogeneity and transaxial slice orientation bias. Nonetheless, using a CSA cutoff value of 130 mm^2

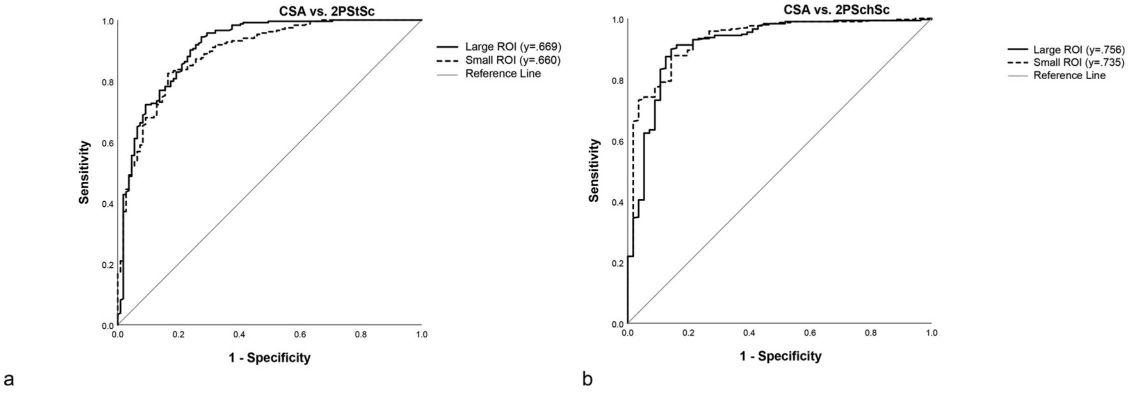

Fig. 3. Receiver operator characteristics (ROC) curves for cross sectional areas (CSAs) of both regions of interest (ROIs) for prediction of severe lumbar spine stenosis (LSS) based on dichotomized qualitative expert ratings (two-point standard and Schizas score, 2PStSc and 2PSchSc in a and b).

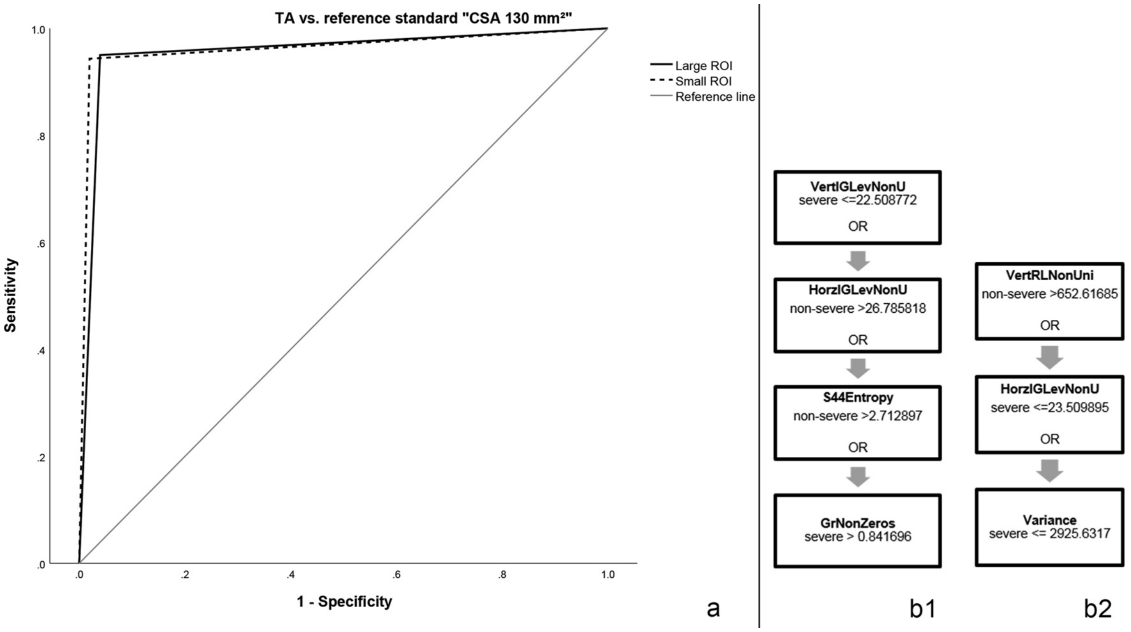

Fig. 4. a) Receiver operator characteristics (ROC) curves demonstrating strong performance of both texture analysis (TA) decision trees (small and large region of interest, ROIs) for severe lumbar spine stenosis (LSS) detection based on quantitative reference standard cutoff cross-sectional area, CSA 130 mm^2, using b) different decision trees from TA for the small (b1) and large (b2) ROI, respectively.

appears to be a simple and valid method to reliably report on LSS in daily clinical route, superior to qualitative assessment.

When dichotomizing both ROIs to a threshold reference CSA of 130 mm^2, StSc gradings of an expert reader - despite lower reproducibility- showed slightly better performance compared to SchSc for severe LSS detection. The larger ROI delivered higher AUC values, indicating better identification of relevant stenosis when including epidural fat and bilateral recesses in the CSA delineation. When testing the measured CSA of ROIs against dichotomized expert qualitative LSS scores, high AUC values indicated good and similar diagnostic performance for both scores (StSc and SchSc). Respective γ–values for optimal diagnostic performance confirmed a CSA of 130 mm^2 as a robust reference standard for severe LSS detection, independent of scoring method and in line with literature.

Given daily clinical routine, we intended to quantify and characterize central LSS of clinically symptomatic patients on a heterogeneous set of MR images from different sites with a robust set of TA parameters. These would potentially offer high reproducibility on the one hand and low dependence on CSA variance on the other. We were also interested in the effect of epidural fat inclusion on TA analysis. As only the Schizas score takes into account the persistence of epidural fat in a subcategory of severe LSS (category C vs. D) our analysis was performed for both ROIs reflecting StSc and SchSc using a common CSA threshold of 130 mm^2 as reference standard to define severe LSS.

Results showed significant correlations of certain TA features with reference standard CSA when analyzed with a machine-learning supported approach. 10-fold cross validation strongly supports the robustness of our findings and suggests potential extrapolation of findings to other cohorts. Surprisingly, TA correlations were found for both ROIs, encircling either the dural sac or the spinal canal including epidural fat and lateral recesses. Despite different decision tree algorithms, the highly comparable performance may be due to the negligible effect inclusion of epidural fat has on certain TA parameters and thus LSS grading. This also puts into question the usefulness of SchSc grades C and D for TA. Diagnostic performance descriptors increased significantly when using TA for detection of severe LSS with sensitivity and specificity of 94% and 98%, respectively when compared to standard CSA measurements using expert qualitative scores as reference (Table 3).

There are limitations of this study. All image data were retrieved from a cohort of different institutes of a large nationwide study on patients with symptoms of LSS. Therefore, study conclusion may be specific to the imaging parameters used in this cohort. However, since we achieved to define robust TA parameters for severe LSS identification we would argue that study findings may similarly apply to other cohorts. TA could probably reveal additional correlating parameters with clinical outcomes, if the retrieved images had been acquired at one single MR scanner. Additionally, due to the spread of LSS variations in our multi-center cohort and having performed a blinded, randomized readout, we considered all images as single cases. Therefore, no statistical adaption for multi-levelling has been applied. We did not investigate the clinical impact of TA analysis from MR images on patients with severe LSS, neither did we explicitly test the influence of scan heterogeneity on our results, but given the high reproducibility and diagnostic performance, we assume that an increase in diagnostic reliability can be expected.

5. Conclusion

Qualitative LSS grading independent of classification system shows moderate reproducibility of grading scores, whereas quantitative CSA measurements seems to be an easier and more reliable method to report LSS in clinical routine. Furthermore, TA offers highly reproducible quantitative parameters that may additionally increase the accuracy for severe LSS detection with minor impact of grading score and CSA border definition.

Conflicts of interest

The authors have no conflicts of interest.

References

[1] J.N. Katz, M.B. Harris, Clinical practice. Lumbar spinal stenosis, N. Engl. J. Med. 358 (8) (2008) 818–825.

[2] D.K. Binder, M.H. Schmidt, P.R. Weinstein, Lumbar spinal stenosis, Semin. Neurol. 22 (2) (2002) 157–166.

[3] M.C. Battie, A. Ortega-Alonso, R. Niemelainen, K. Gill, E. Levalahti, T. Videman, J. Kaprio, Lumbar spinal stenosis is a highly genetic condition partly mediated by

disc degeneration, Arthritis Rheumatol. 66 (12) (2014) 3505–3510.

[4] R.A. Deyo, D.T. Gray, W. Kreuter, S. Mirza, B.I. Martin, United States trends in lumbar fusion surgery for degenerative conditions, Spine (Phila Pa 1976) 30 (12) (2005) 1441–1445 discussion 1446-7.

[5] R.A. Deyo, S.K. Mirza, B.I. Martin, W. Kreuter, D.C. Goodman, J.G. Jarvik, Trends, major medical complications, and charges associated with surgery for lumbar spinal stenosis in older adults, JAMA 303 (13) (2010) 1259–1265.

[6] S. Hall, J.D. Bartleson, B.M. Onofrio, H.L. Baker Jr., H. Okazaki, J.D. O'Duffy, Lumbar spinal stenosis. Clinical features, diagnostic procedures, and results of surgical treatment in 68 patients, Ann. Intern. Med. 103 (2) (1985) 271–275.

[7] J. Steurer, S. Roner, R. Gnannt, J. Hodler, C. LumbSten Research, Quantitative radiologic criteria for the diagnosis of lumbar spinal stenosis: a systematic literature review, BMC Musculoskelet. Disord. 12 (2011) 175.

[8] G. Andreisek, J. Hodler, J. Steurer, Uncertainties in the diagnosis of lumbar spinal stenosis, Radiology 261 (3) (2011) 681–684.

[9] G. Andreisek, M. Imhof, M. Wertli, S. Winklhofer, C.W. Pfirrmann, J. Hodler, J. Steurer, Z. Lumbar Spinal Stenosis Outcome Study Working Group, A systematic review of semiquantitative and qualitative radiologic criteria for the diagnosis of lumbar spinal stenosis, AJR Am. J. Roentgenol. 201 (5) (2013) W735–46.

[10] N. Mamisch, M. Brumann, J. Hodler, U. Held, F. Brunner, J. Steurer, Z. Lumbar Spinal Stenosis Outcome Study Working Group, Radiologic criteria for the diagnosis of spinal stenosis: results of a Delphi survey, Radiology 264 (1) (2012) 174–179.

[11] J. Steurer, A. Nydegger, U. Held, F. Brunner, J. Hodler, F. Porchet, K. Min, A.F. Mannion, B. Michel, C. LumbSten Research, LumbSten: the lumbar spinal stenosis outcome study, BMC Musculoskelet. Disord. 11 (2010) 254.

[12] N.F. Bolender, N.S. Schönström, D.M. Spengler, Role of computed tomography and myelography in the diagnosis of central spinal stenosis, J. Bone Joint Surg. Am. 67 (2) (1985) 240–246.

[13] M. Mariconda, R. Fava, A. Gatto, C. Longo, C. Milano, Unilateral laminectomy for bilateral decompression of lumbar spinal stenosis: a prospective comparative study with conservatively treated patients, J. Spinal Disord. Tech. 15 (1) (2002) 39–46.

[14] N.S. Schonstrom, N.F. Bolender, D.M. Spengler, The pathomorphology of spinal stenosis as seen on CT scans of the lumbar spine, Spine (Phila Pa 1976) 10 (9) (1985) 806–811.

[15] C. Schizas, N. Theumann, A. Burn, R. Tansey, D. Wardlaw, F.W. Smith, G. Kulik, Qualitative grading of severity of lumbar spinal stenosis based on the morphology of the dural sac on magnetic resonance images, Spine (Phila Pa 1976) 35 (21) (2010) 1919–1924.

[16] M.G. Lubner, A.D. Smith, K. Sandrasegaran, D.V. Sahani, P.J. Pickhardt, CT texture analysis: definitions, applications, biologic correlates, and challenges, Radiographics 37 (5) (2017) 1483–1503.

[17] R.A. Lerski, L.R. Schad, R. Luypaert, A. Amorison, R.N. Muller, L. Mascaro, P. Ring, A. Spisni, X. Zhu, A. Bruno, Multicentre magnetic resonance texture analysis trial using reticulated foam test objects, Magn. Reson. Imaging 17 (7) (1999) 1025–1031.

[18] D. Jirak, M. Dezortova, M. Hajek, Phantoms for texture analysis of MR images. Long-term and multi-center study, Med. Phys. 31 (3) (2004) 616–622.

[19] G.Y. Lee, Y.L. Guen, J.W. Lee, W.L. Joon, H.S. Choi, S.C. Hee, K.J. Oh, O. Kyoung-Jin, H.S. Kang, S.K. Heung, A new grading system of lumbar central canal stenosis on MRI: an easy and reliable method, Skeletal Radiol. 40 (8) (2011) 1033–1039.

[20] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (1977) 159–174.

[21] P.M. Szczypinski, M. Strzelecki, A. Materka, A. Klepaczko, MaZda—a software package for image texture analysis, Comput. Methods Programs Biomed. 94 (1) (2009) 66–76.

[22] D. Cicchetti, Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instrument in Psychology, (1994).

[23] Appendix B - the WEKA workbench, in: I.H. Witten, E. Frank, M.A. Hall, C.J. Pal (Eds.), Data Mining (Fourth Edition), Morgan Kaufmann, 2017, pp. 553–571.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newsl. 11 (1) (2009) 10–18.

[25] J.D. Lurie, A.N. Tosteson, T.D. Tosteson, E. Carragee, J.A. Carrino, J. Carrino, J. Kaiser, R.T. Sequeiros, A.R. Lecomte, M.R. Grove, E.A. Blood, L.H. Pearson, J.N. Weinstein, R. Herzog, Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis, Spine (Phila Pa 1976) 33 (14) (2008) 1605–1610.

[26] R.J. Herzog, J.A. Kaiser, J.A. Saal, J.S. Saal, The importance of posterior epidural fat pad in lumbar central canal stenosis, Spine (Phila Pa 1976) 16 (6 Suppl) (1991) S227–S233.