Journal of Digital Imaging (2021) 34:811–819
https://doi.org/10.1007/s10278-021-00462-1

ORIGINAL PAPER

The Composite Severity Score for Lumbar Spine MRI: a Metric of Cumulative Degenerative Disease Predicts Time Spent on Interpretation and Reporting

Michael Travis Caton Jr.[1] · Walter F. Wiggins[2] · Stuart R. Pomerantz[3] · Katherine P. Andriole[4]

Received: 18 July 2020 / Revised: 3 March 2021 / Accepted: 6 May 2021 / Published online: 23 May 2021

Abstract

Conventional measures of radiologist efficiency, such as the *relative value unit*, fail to account for variations in the complexity and difficulty of a given study. For lumbar spine MRI (LMRI), an ideal performance metric should account for the global severity of lumbar degenerative disease (LSDD) which may influence reporting time (RT), thereby affecting clinical productivity. This study aims to derive a global LSDD metric and estimate its effect on RT. A 10-year archive of LMRI reports comprising 13,388 exams was reviewed. Objective reporting timestamps were used to calculate RT. A natural language processing (NLP) tool was used to extract radiologist-assigned stenosis severity using a 6-point scale (0 = "normal" to 5 = "severe") at each lumbar level. The composite severity score (CSS) was calculated as the sum of each of 18 stenosis grades. The predictive values of CSS, sex, age, radiologist identity, and referring service on RT were examined with multiple regression models. The NLP tool accurately classified LSDD in 94.8% of cases in a validation set. The CSS increased with patient age and differed between men and women. In a univariable model, CSS was a significant predictor of mean RT ($R^2 = 0.38$, $p < 0.001$) and independent predictor of mean RT ($p < 0.001$) controlling for patient sex, patient age, service location, and interpreting radiologist. The predictive strength of CSS was stronger for the low CSS range (CSS = 0–25, $R^2 = 0.83$, $p < 0.001$) compared to higher CSS values (CSS > 25, $R^2 = 0.15$, $p = 0.05$). Individual radiologist study volume was negatively correlated with mean RT (Pearson's $R = -0.35$, $p < 0.001$). The composite severity score predicts radiologist reporting efficiency in LMRI, providing a quantitative measure of case complexity which may be useful for workflow planning and performance evaluation.

Introduction

Efficient interpretation and reporting are central principles of delivering value in diagnostic radiology [1]. Common performance metrics such as the relative value unit (RVU) are used to quantify the value of physician work as a function of time [2]. However, the RVU does not account for variability in procedural complexity which may significantly alter the duration of a medical procedure [3]. This discrepancy can generate systemic bias and misaligned incentives for physicians and healthcare organizations [4]. In diagnostic radiology, we hypothesize that studies with more complex pathology would require more cognitive energy, resulting in longer reporting time (RT) that is not captured by the RVU [5–7] .

To address this question, we studied degenerative disease of the lumbar spine (LSDD) on MRI (LMRI). Compared with other imaging reports, LMRI reporting is typically structured by anatomic level and each level is assigned a grade of relative severity of stenosis [8]. Prior work has shown that inter-radiologist agreement for spinal canal stenosis (SCS) and neural foraminal stenosis (NFS) is moderate to strong [9–13]. Moreover, the descriptive terminology for LSDD is standardized by an interdisciplinary consortium of national societies allowing comparison between radiologists with different training and levels of experience [14]. These

✉ Michael Travis Caton Jr.
travis.caton@gmail.com

[1] Department of Radiology and Biomedical Imaging, University of California San Francisco, 505 Parnassus Ave L305, San Francisco, CA, USA

[2] Department of Radiology, Duke University, Durham, NC 27708, USA

[3] Department of Radiology, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114, USA

[4] Partners Center for Clinical Data Science, 100 Cambridge Street, Boston, MA 02114, USA

features enable analysis of reporting text by natural language processing (NLP) algorithms, which allow a large corpus of text to be analyzed rapidly.

We hypothesized that a measurable relationship exists between disease severity (complexity) and RT owing to increased cognitive energy required to render a diagnosis and assign a severity value [15, 16]. The primary aim of this study was to quantify the relationship of radiologist-assigned LSDD severity and reporting time, leveraging NLP to analyze a large volume of reports from many individual radiologists. To this end, we present and validate an NLP tool and derive a composite "severity score" reflecting the cumulative severity of LSDD at 18 discrete sites in the lumbar spine. We then evaluate the predictive value of this composite severity score on objective reporting times.

Methods and Materials

Data Selection

We performed an institutional review board-approved review of LMRI reporting data in compliance with the Health Insurance Portability and Accountability Act for which informed consent was waived. The 10-year archive of report text and reporting metadata comprised 43,255 LMRI studies from a single, multi-site institution (BLINDED FOR REVIEW), representing 29 different MR scanners. Studies which were co-interpreted with trainee assistance ($n = 23{,}607$) or reported on the weekend ($n = 3202$) were excluded because the workflow for these studies was presumably different. We also excluded studies in which the calculated RT was over 60 min ($n = 3058$) because this was felt to be beyond the conventional RT for an attending radiologist without significant interruption (exclusion criteria summarized in supplemental Fig. 1) [17]. Reports were signed by 152 unique attending radiologists including generalists, fellowship-trained musculoskeletal radiologists, and neuroradiologists. Patient age, sex, and the requesting service (documented as "inpatient", "outpatient", "emergency", or "unknown") were recorded. Reporting documentation timestamps were extracted from the institution's reporting software API (Powerscribe, Nuance Communications, Burlington VT, USA). The timestamps marked the initialization and finalization of the radiology report to the 1/100 of a second; the difference between these time points was recorded as the RT (min).

Natural Language Processing

We applied a customized NLP algorithm to raw radiology report text. Using a rule-based approach employing a dictionary of customized regular expressions (RegEx), the algorithm was designed to extract a severity rating using a 6 point scale (0 = "normal", 1 = "mild", 2 = "mild to moderate", 3 = "moderate", 4 = "moderate to severe", and 5 = "severe"). A severity score was extracted for spinal canal stenosis (SCS) and left and right neural foraminal stenosis (LNFS, RNFS) for each of six spinal levels: T12-L1 through L5-S1. The NLP was designed using empirically, iteratively developed syntactic and semantic rules including common radiology terminology and phraseology to elicit a severity score (0–5) for the 18 locations (e.g., T12-L1 SCS, T12-L1 LNFS), resulting in a 6 × 3 matrix for each study (Fig. 1). When the model failed to assign a score, a default value of 0 ("normal") was applied. To test the accuracy of the model, we randomly selected 100 studies out of the full dataset ($n = 43{,}255$) and manually reviewed the radiology reporting text to assess for discrepancy or error. For each case, the reporting text was manually reviewed and assigned a 0–5 value by a radiologist. Scores were considered concordant if NLP and manual review matched exactly and any degree of discordance was considered unsuccessful.

Calculation of Severity Scores

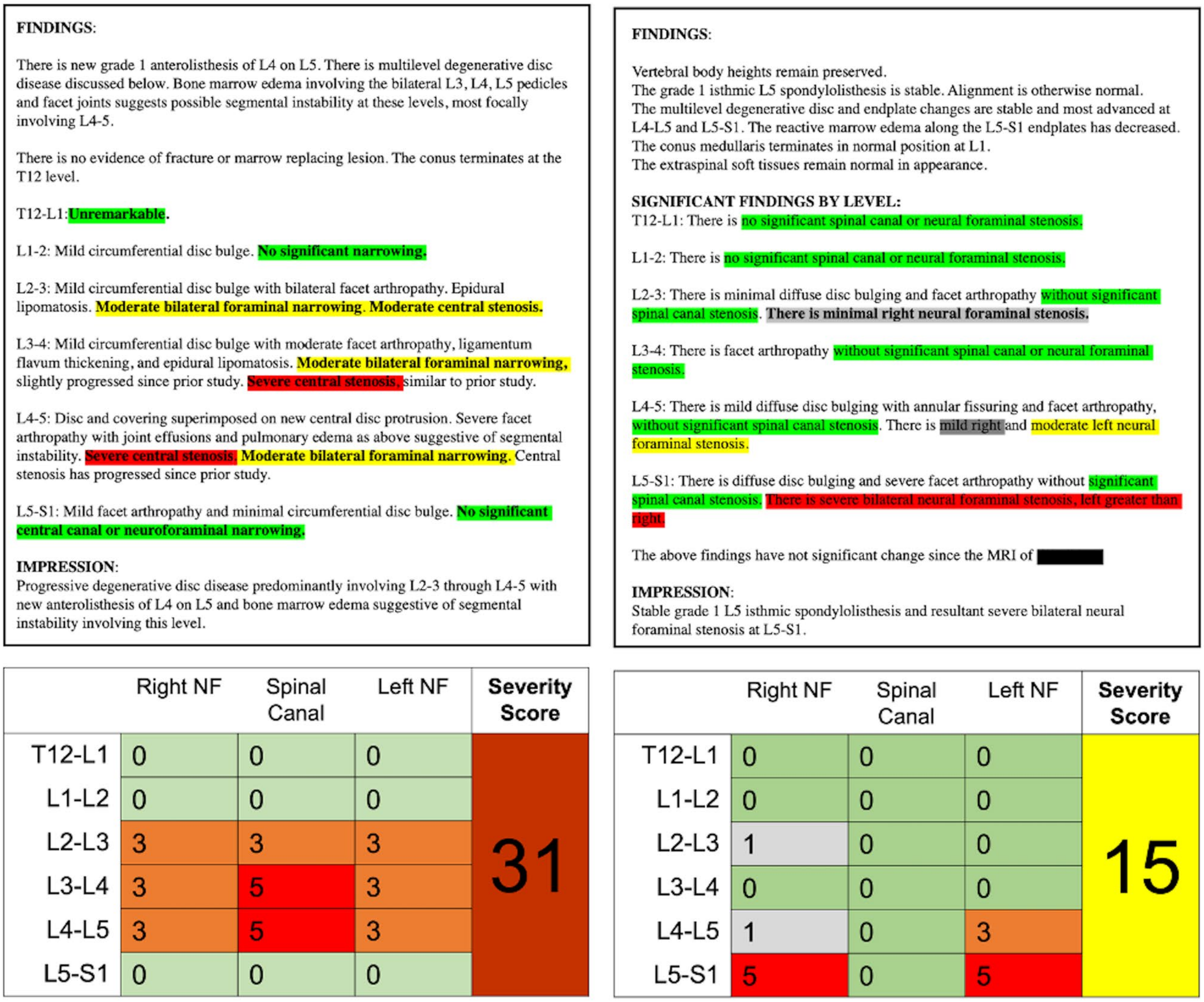
The composite severity score (CSS) was calculated as the sum of the constituent severity scores ($n = 18$) resulting in scale ranging from 0 ("normal" at all locations) to a theoretical maximum of 90 ("severe" at all 18 locations).

The distribution of CSS was assessed for the study population ($n = 13{,}388$) and subdivided by age and sex. Age groups were defined as < 40, 40–49, 50–59, 60–69, 70–79, and ≥ 80. The sex-based distributions and age group distributions were compared to exponential theoretical distributions using quantile–quantile plots. The Kolmogorov–Smirnov (K-S) test was used to test difference in distribution between subgroups.

Quantifying Relationships Among CSS, Age, Sex, and RT

CSS and RT were grouped by sex and age and inter-group differences were calculated using ANOVA and pairwise *t* tests with the Benjamini–Hochberg method correcting for multiple comparisons. CSS were clustered into deciles and compared to RT using pairwise *t* tests. The results are visualized in box-and-whisker diagrams which show the median (central bar), interquartile range (box margins).

Simple, univariable linear least squares regression was used to assess the predictive relationship of CSS, age, sex, service location, and interpreting radiologist on RT. The distribution of residuals indicated heteroscedasticity (Supplemental Fig. 2); therefore, the regression was performed using log–log transformation. Significant predictors were then grouped, and a multiple least squares regression was performed to assess for independent effects of each variable



FINDINGS:

There is new grade 1 anterolisthesis of L4 on L5. There is multilevel degenerative disc disease discussed below. Bone marrow edema involving the bilateral L3, L4, L5 pedicles and facet joints suggests possible segmental instability at these levels, most focally involving L4-5.

There is no evidence of fracture or marrow replacing lesion. The conus terminates at the T12 level.

T12-L1:**Unremarkable**.

L1-2: Mild circumferential disc bulge. **No significant narrowing.**

L2-3: Mild circumferential disc bulge with bilateral facet arthropathy. Epidural lipomatosis. **Moderate bilateral foraminal narrowing. Moderate central stenosis.**

L3-4: Mild circumferential disc bulge with moderate facet arthropathy, ligamentum flavum thickening, and epidural lipomatosis. **Moderate bilateral foraminal narrowing,** slightly progressed since prior study. **Severe central stenosis,** similar to prior study.

L4-5: Disc and covering superimposed on new central disc protrusion. Severe facet arthropathy with joint effusions and pulmonary edema as above suggestive of segmental instability. **Severe central stenosis. Moderate bilateral foraminal narrowing.** Central stenosis has progressed since prior study.

L5-S1: Mild facet arthropathy and minimal circumferential disc bulge. **No significant central canal or neuroforaminal narrowing.**

IMPRESSION:
Progressive degenerative disc disease predominantly involving L2-3 through L4-5 with new anterolisthesis of L4 on L5 and bone marrow edema suggestive of segmental instability involving this level.

	Right NF	Spinal Canal	Left NF	**Severity Score**
T12-L1	0	0	0	31
L1-L2	0	0	0	
L2-L3	3	3	3	
L3-L4	3	5	3	
L4-L5	3	5	3	
L5-S1	0	0	0	

FINDINGS:

Vertebral body heights remain preserved.
The grade 1 isthmic L5 spondylolisthesis is stable. Alignment is otherwise normal.
The multilevel degenerative disc and endplate changes are stable and most advanced at L4-L5 and L5-S1. The reactive marrow edema along the L5-S1 endplates has decreased.
The conus medullaris terminates in normal position at L1.
The extraspinal soft tissues remain normal in appearance.

SIGNIFICANT FINDINGS BY LEVEL:
T12-L1: There is no significant spinal canal or neural foraminal stenosis.

L1-2: There is no significant spinal canal or neural foraminal stenosis.

L2-3: There is minimal diffuse disc bulging and facet arthropathy without significant spinal canal stenosis. **There is minimal right neural foraminal stenosis.**

L3-4: There is facet arthropathy without significant spinal canal or neural foraminal stenosis.

L4-5: There is mild diffuse disc bulging with annular fissuring and facet arthropathy, without significant spinal canal stenosis. There is mild right and moderate left neural foraminal stenosis.

L5-S1: There is diffuse disc bulging and severe facet arthropathy without significant spinal canal stenosis. There is severe bilateral neural foraminal stenosis, left greater than right.

The above findings have not significant change since the MRI of [redacted]

IMPRESSION:
Stable grade 1 L5 isthmic spondylolisthesis and resultant severe bilateral neural foraminal stenosis at L5-S1.

	Right NF	Spinal Canal	Left NF	**Severity Score**
T12-L1	0	0	0	15
L1-L2	0	0	0	
L2-L3	1	0	0	
L3-L4	0	0	0	
L4-L5	1	0	3	
L5-S1	5	0	5	

Fig. 1 Two examples of NLP analysis of archived radiology report text (**A**). Using regular expression methods and the structured or semi-structured nature of LMRI spine reporting, the NLP algorithm extracted severity scores for the lumbar spine, resulting in a 6×3 matrix of values ranging 0–5 (normal–severe). The composite severity score is the sum of these values

on RT. For each test, p value < 0.05 was considered significant. A curvilinear regression was then fitted to model predictive value of CSS on mean RT. This was repeated for subgroups of low-CSS (0–25) and high-CSS (> 25) groups. Unless otherwise stated, values are reported with ± standard error of the mean (SEM). All statistical analysis and data visualizations were performed using R statistical computing software (v 3.6.2, The R Corporation).

Results

Study Characteristics

From the initial dataset of 43,255 studies, we included 13,388 after applying exclusion criteria (Supplemental Fig. 1) representing 12,326 unique patients. The mean patient age was 54.8 ± 0.1 years and the sex distribution was 54.5% women. The mean RT for the study population was 14.41 ± 0.1 min (median = 10.7), with no significant difference in mean RT by patient sex (14.56 ± 0.15 min for men vs. 14.29 ± 0.14 min for women, $p = 0.17$). These differences are summarized in Table 1.

Natural Language Processing

The NLP accuracy at the level of CSS was 94.8% (93 misclassifications out of 1800 test values in random sample of 100 cases). The NLP was 100% accurate in 5/18 level instances (27.8%) and was least accurate at right L5-S1, correctly classifying the severity in 86% of cases.

Table 1 The mean composite severity score (CSS) and interpretation and reporting time (RT) for the full study population and subgroups by sex and age

	CSS mean ± SE	*p* Value		RT (min)	*p* Value	
Total population ($n = 13{,}388$)	6.04 ± 0.07	---		14.41 ± 0.1	---	
Sex		Pairwise *t* test	ANOVA		Pairwise *t* test	ANOVA
Men ($n = 6088$)	7.0 ± 0.11	$< .001$	---	14.56 ± 0.15	0.17	---
Women ($n = 7300$)	5.23 ± 0.09			14.29 ± 0.14		
Age groups						
< 40 ($n = 2885$)	1.63 ± 0.06	---	$< .001$	11.9 ± 0.2	---	$< .001$
40–50 ($n = 2440$)	3.20 ± 0.10	$< .001$		13.23 ± 0.22	$< .001$	
50–60 ($n = 2844$)	5.38 ± 0.12	$< .001$		14.38 ± 0.21	$< .001$	
60–70 ($n = 2639$)	8.35 ± 0.18	$< .001$		15.7 ± 0.23	$< .001$	
70–80 ($n = 1817$)	11.47 ± 0.25	$< .001$		16.32 ± 0.28	0.07	
80 + ($n = 763$)	13.29 ± 0.44	$< .001$		18.83 ± 0.47	$< .001$	

*Interquartile range

Relationship of CSS, Age, and Sex

Mean CSS for the full study population was 6.04 ± 0.07. The relationship of CSS and RT to patient age is shown in Fig. 2A and showed a moderate positive correlation with Pearson correlation coefficient $R = 0.44$ ($p < 0.001$) and a weak positive correlation between age and RT (Pearson coefficient $R = 0.16$, $p < 0.001$). The distribution of CSS was exponential and differed by different age group with increasing flattening of the distribution at higher age groups, indicating a shift toward higher CSS (Fig. 2B). The distribution of CSS differed between men and women (K-S test, $p < 0.001$, Fig. 3A). Mean CSS was higher for men than women overall (7.0 ± 0.11 vs. 5.23 ± 0.09, $p < 0.001$) and the relative difference in proportion by sex is shown for each CSS value in Fig. 3B. The relationships of age and sex to CSS are summarized in Table 1.

Predictive Value of CSS on RT

There was a modest, positive correlation between CSS and RT (Pearson's $R = 0.23$, $p < 0.001$). In a log–log linear regression model, CSS was a significant predictor of RT (adjusted $R^2 = 0.063$) according to the function $\ln(\mathrm{RT}) = 2.14 + 0.17(\ln(\mathrm{CSS}))$ ($F_{1,13386)} = 905.6$, $p < 0.001$) meaning that for every 1% increase in CSS, we predict a 0.17% increase in RT. The predictive relationship of mean RT to CSS was better modeled using a 3rd order polynomial regression which shows an adjusted $R^2 = 0.38$ ($p < 0.001$) (Fig. 4). Polynomial regression for the low CSS range (0–25) showed a stronger predictive value ($R^2 = 0.83$, $p < 0.001$) compared to high CSS range (> 25) ($R^2 = 0.15$, $p = 0.05$) (Supplemental Fig. 3).

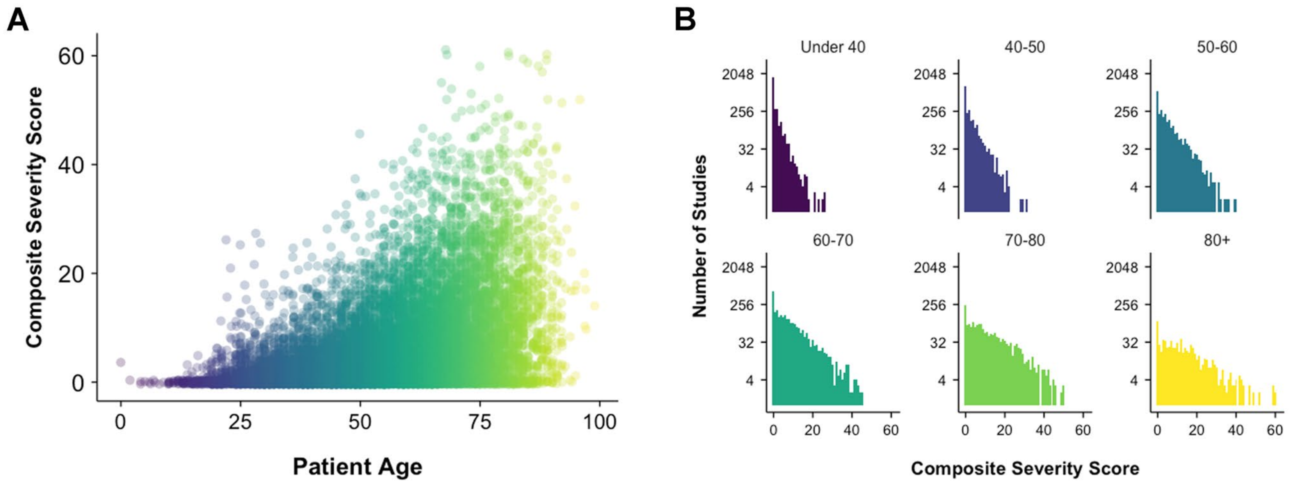


Fig. 2 Distribution of composite severity score (CSS) showed a moderate positive correlation between age and disease severity (Pearson's $R = 0.44$, $p < 0.001$) **A**. The distribution of CSS by age group **B** showed greater proportion of higher CSS in older patients

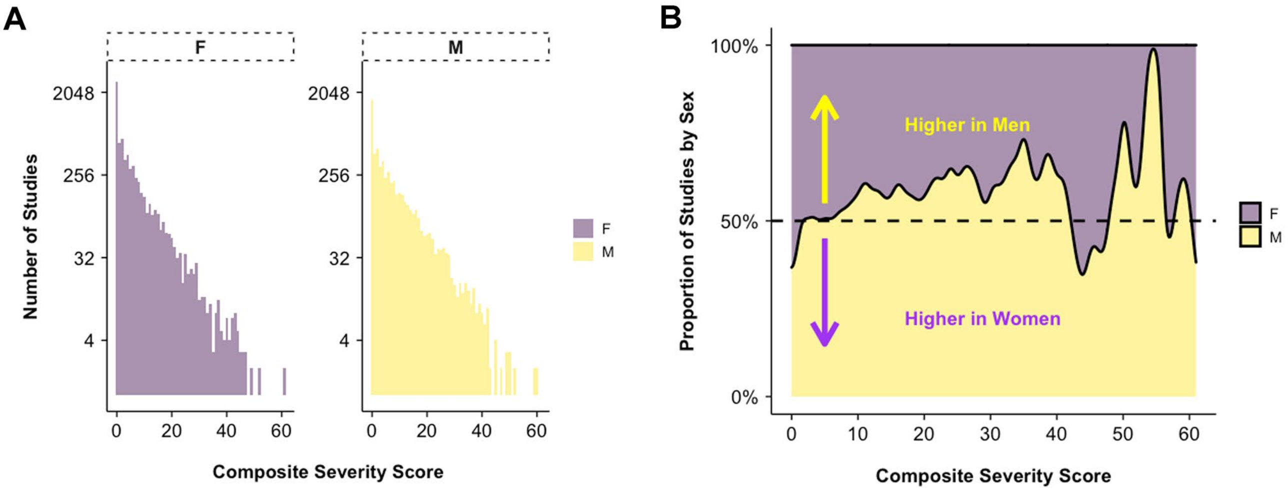


Fig. 3 The distribution of composite severity scores (CSS), stratified by sex (M = male, F = female, **A**) showing an exponential pattern. Distribution of CSS for men and women differed (K-S test $p < .001$). A large portion of both groups including 1797/6088 (27.87%) men, 2679/7300 (36.69%) women had CSS = 0 (normal). The proportions of the sex distributions are compared in **B** by CSS value indicating relative greater proportion of normal and low CSS for women (0–5) and relative greater proportion of higher CSS (10–40) in men

The interpreting radiologist identity was also a statistically significant predictor for RT, accounting for a larger portion of variance than CSS (adjusted $R^2 = 0.45$, $F_{(151,13236)} = 73.97$, $p < 0.001$). Simple regression models also identified smaller, significant predictive value of age (adjusted $R^2 = 0.04$, F$(_{1,\ 13386}) = 528.4$, $p < 0.001$), sex (adjusted $R^2 = < 0.01$, $F_{(1,\ 13386)} = 4.47$, $p = 0.03$), and requesting service (adjusted $R^2 = 0.026$, $F_{(3,\ 13384)} = 118.7$, $p < 0.001$). In the multiple regression model incorporating all 5 predictors (CSS, interpreting radiologist, patient age, patient sex, and requesting service), each variable was a significant and independent predictor of RT ($p < 0.001$ for each). In this 5-variable model, the adjusted R^2 was 0.52 ($F_{(216,\ 13171)} = 67.56$, $p < 0.001$) and the interpreting radiologist was the most important predictor (partitioned $R^2 = 0.43$) followed by CSS (partitioned $R^2 = 0.043$), age (partitioned $R^2 = 0.025$, requesting service (partitioned $R^2 = 0.19$), and sex (partitioned $R^2 = 0.00024$).

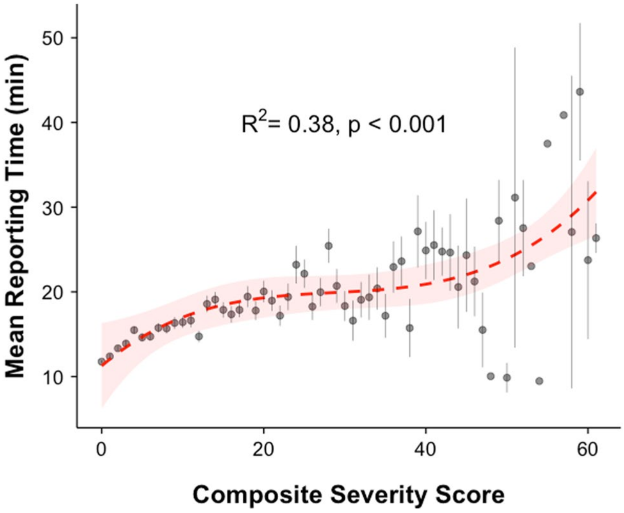


Fig. 4 CSS was a significant predictor of RT in least squares polynomial regression ($R^2 = 0.38$, $p < 0.001$). Mean values for each CSS are plotted with vertical lines indicating standard error. Red dashed line shows the regression model with shading indicating 95% confidence interval of the model

We next examined the CSS in 10 groups, based on relative sample size distribution deciles. The relationship of CSS groups and RT is shown in Fig. 5. There were significant stepwise increases between groups (ANOVA, $p < 0.001$) and pairwise analysis showed a significant, stepwise for each sequential group except the transition from CSS 21–30 to CSS 31–40 ($p = 0.74$) and CSS 31–40 to CSS 41–50 ($p = 0.13$) (Table 2).

Because the effect size of individual radiologist identity was large, the mean RT as a function of mean CSS was plotted for individual radiologists, excluding the 5% upper and lower tail outliers (Fig. 6). There was a significant positive relationship between mean CSS and RT by radiologist (Pearson's $R = 0.26$, $p = 0.002$). There was also a significant negative correlation between the radiologist's study volume (number of studies interpreted by each radiologist over the 10-year period) and mean RT (Fig. 7A, Pearson's $R = -0.35$, $p < 0.001$). Pairwise t tests between radiologist volume quintiles showed significant decrease in mean RT between the 3rd and 4th quintiles (Fig. 7B) (25.02 ± 5.13 min vs. 20.76 ± 5.30 min, $p = 0.02$) and the 4th and 5th quintiles (20.76 ± 5.30 min vs. 15.00 ± 3.67 min, $p = 0.004$). There were no significant pairwise differences in the 1st–3rd quintiles ($p > 0.05$). The correlation between

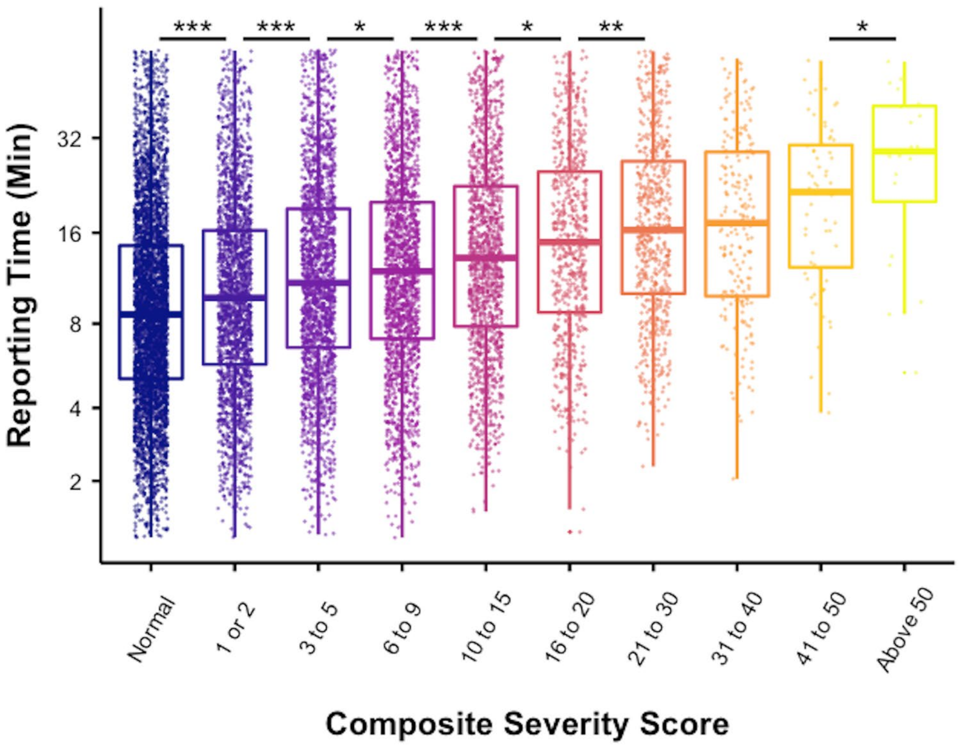


Fig. 5 The relationship of CSS group to RT shown as box-and-whisker plots. The box delineates the median and interquartile range. Individual CSS-RT relationship is superimposed as a color-matched dot within each CSS group. Significant pairwise increase in RT is indicated by *$p<0.05$, **$p<0.01$, ***$p<0.001$

individual radiologist volume and CSS was nonsignificant ($p=0.75$).

Discussion

The CSS is a tool for quantifying radiologic complexity in lumbar spine MRI and is a significant, independent predictor of interpretation and reporting time (RT). The relationship of CSS and RT supports the anecdotal observation that severe LSDD is more challenging to interpret than normal anatomy, measured in terms of time spent generating a final radiology report. The CSS may be useful as a 'correction factor', allowing studies to be weighted based on overall complexity. The CSS could then be applied to quality improvement by defining efficiency benchmarks, ensuring equitable distribution of work, and improving existing measures of radiologist productivity (i.e., RVU).

The principle finding of this study is that the CSS is a reasonable predictor of mean RT ($R^2=38\%$). Notably, the CSS performed better at low-range CSS than high-range CSS ($R^2=83\%$ vs. 15%), suggesting a de-coupling of this

Table 2 The relationship of CSS and RT subdivided by 10 CSS groups. Mean RT increased in a stepwise fashion (pairwise *t* test) at each increment except for CSS 21–30 to 31–40 and CSS 3140 and 41–50

Interpretation and reporting time (RT) as a function of composite severity score (CSS) group

CSS group	Mean RT (min) (± SE)	Median RT (min) (interquartile range)	*p* value	
			Pairwise *t* test	ANOVA
Normal (*n* = 4376)	11.76 (±0.15)	8.60 (9.42)		
1 or 2 (*n* = 2154)	12.89 (±0.22)	9.78 (10.50)	< .001	< .001
3–5 (*n* = 2012)	14.72 (±0.26)	10.97 (12.48)	< .001	
6–9 *n* = 1801)	15.49 (±0.28)	11.99 (12.95)	0.04	
10–15 (*n* = 1438)	17.07 (±0.33)	13.26 (14.72)	< .001	
16–20 (*n* = 680)	18.36 (±0.48)	14.92 (16.32)	0.02	
21–30 (*n* = 637)	20.22 (±0.53)	16.33 (16.92)	.003	
31–40 (*n* = 204)	20.52 (±0.91)	17.21 (19.01)	0.74	
41–50 (*n* = 66)	22.96 (±1.53)	21.60 (18.06)	0.13	
Above 50 (*n* = 20)	29.61 (±3.25)	29.05 (19.63)	.02	

*Pairwise *t* test with Benjamini–Hochberg correction

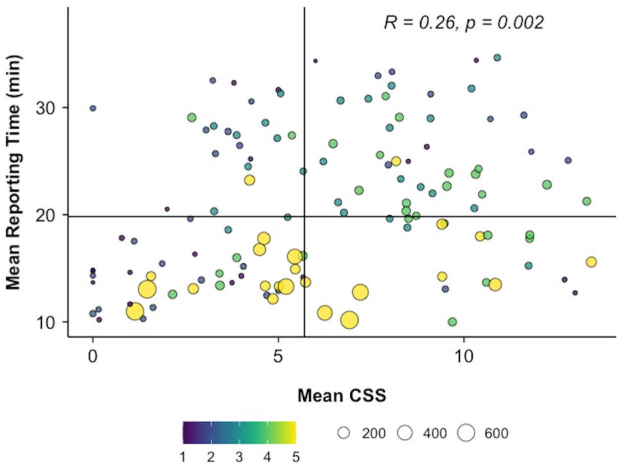


Fig. 6 Individual radiologist practice patterns: the relationship of mean CSS and mean RT for each radiologist included in this study ($n = 152$), excluding 5% upper and lower tail outliers. Each point represents an individual radiologist and point size is proportional to cumulative volume of LMRI for that individual. The volume quintiles are indicated by the color scale. Vertical and horizontal lines represent median RT and CSS, respectively. There was a significant positive correlation between mean CSS and mean RT ($R = 0.26$, $p = 0.002$)

relationship for LMRI studies with large burden of LSDD. In multiple regression, CSS was relatively less important than the identity of the interpreting radiologist. Taken together, these findings indicate that LSDD severity is only one of many factors which influence real-time radiologist efficiency. Nonetheless, we can extract meaningful benchmarks from this large dataset which estimates an increase in median RT of 50% for LMRI studies with a CSS of ~ 12 relative to a normal study (CSS = 0). The Current Procedural Terminology (CPT) Code for non-contrast LRMI (72,148) in the Washington, DC area, is assigned a global RVU of 6.32, of which 1.48 RVU is allocated for physician work (estimated payment ranging from $220.87 to $273.47) [4, 18, 19]. LMRI can therefore account for a considerable portion of revenue for both academic and private groups. Within a radiology practice, a CSS-weighted RVU could result in more equitable recognition of work, accounting for procedural complexity, a process which has been used to improve performance and quality in pediatrics and surgery[20, 21]. Moreover, the CSS was a stronger relative predictor of RT than conventional demographic information (sex, age) and referring service (inpatient, outpatient, emergency). The CSS also remained a robust predictor despite significant variation between radiologists with different training backgrounds and experience level.

Differences observed in CSS by sex and age group are largely consistent with prior cohort studies of LSDD in cohort populations, supporting the feasibility of the model [22–25]. The distribution of CSS by sex indicates a higher prevalence of normal-low CSS for women and conversely, higher prevalence of high CSS for men, consistent with prior work showing sex differences in lumbar spine degeneration patterns, attributed to occupational and hormonal differences, gestation, among others [26]. In our population, the difference in CSS was significant (7.00 vs. 5.23 for men and women, respectively, $p < 0.001$), and the distribution of CSS also differed by sex. However, despite these anatomic differences, there was no sex difference in RT. It is unclear if this asymmetry is due more to sex-specific differences in natural history and disease prevalence, other unaccounted

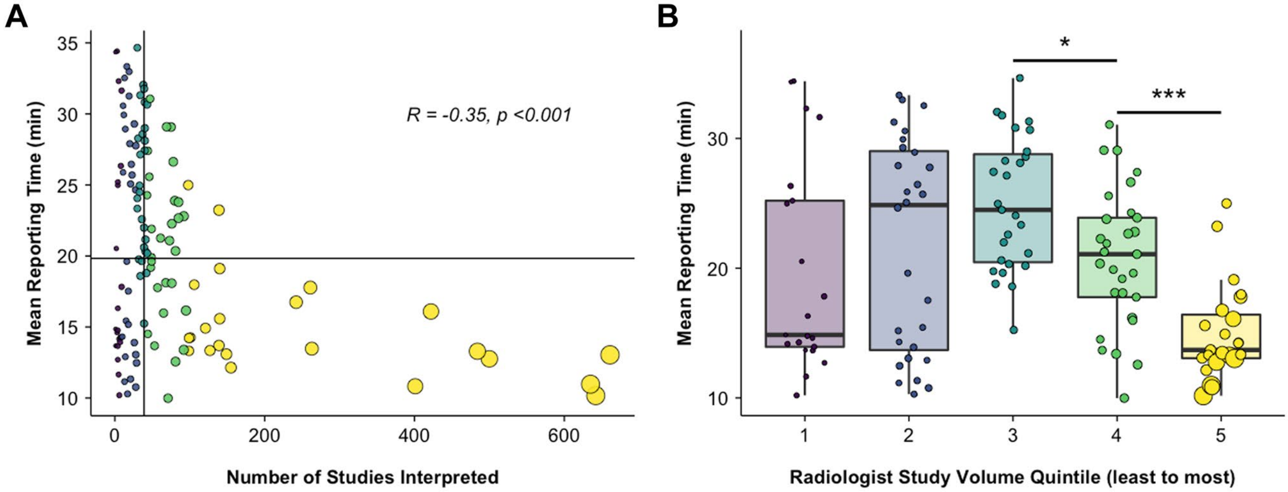


Fig. 7 The relationship of individual radiologist LMRI volume and RT, color-coding represents LMRI volume by quintile (purple = 1st… yellow = 5th). **A** Mean RT as a function of radiologist volume, indicating a significant negative relationship ($R = -0.35$, $p < 0.001$). Vertical and horizontal lines indicate median LMRI volume (by radiologist) and median RT, respectively. Between-group differences for quintiles are shown as box and whisker plots **B**; between-group differences are indicated by *$p < 0.05$, ***$p < 0.001$

sex differences in anatomy, or potentially due to implicit bias. While the CSS is not sufficient to directly inform surgical decision-making, the utility of CSS in predicting surgical candidacy as well as its correlation with functional status could be evaluated in future work.

Lastly, while CSS was an important predictor of RT, the effect size attributable to different radiologists was tenfold higher in the multiple regression model (43% vs. 4.3%). In a population of 152 radiologists, significant variation in training, experience, and clinical role could all contribute to this effect. Still, the relationship of CSS and RT showed a significant correlation across radiologists despite substantial differences in individual practice patterns ($R = 0.26$, $p = 0.002$). There was a highly significant negative relationship between cumulative study volume and RT indicating that radiologists who read more LMRI over the 10-year period tended to report studies more efficiently. This effect was driven largely by the top quintile of radiologists, who read a disproportionately high number of studies. This may reflect experience; radiologists with higher cumulative volume have likely been in practice longer and may need less time to interpret LMRI; alternatively, radiologists who interpret studies quickly (regardless of experience) are likely to have higher volume overall. The impact of CSS on individual reader practice could be studied in future, prospective work.

The major strength of this study is the large and heterogenous dataset from which the input and outcome variables are derived. By modeling the LSDD pattern of over 12,000 unique patients, the CSS reflects large sample distribution which is likely representative of the general population. The CSS also incorporates the interpretation styles and practice behaviors of over 150 radiologists. The use of timestamp-derived performance metrics is also a strength because these values represent "real-world" practice and more likely to capture realistic practice patterns than an artificial experimental setting.

The principal limitations of this study lie in the assumptions used to build our measurement variable (CSS) and outcome variable (RT). Our CSS model does not account for NLP model error, which, although uncommon, could bias the CSS from ground truth report text. Moreover, the NLP tool was validated on the complete dataset rather than the study population ($n = 13{,}388$). It is plausible that the exclusion criteria for this study may introduce additional unmeasured bias in NLP accuracy. Further, we opted to use reporting text rather than the underlying radiographic images that were not referenced as a standard. Instead, our CSS presumes high fidelity translation of disease severity in the report and does not account for inter-reader variability. Nonetheless, the simplicity of the CSS provides an intuitive understanding of the interpretation process. Lastly, the generalizability of our results outside of this single academic institution is uncertain. Significant variations in reporting style, case mix, and patient demographics all likely contribute to variance in RT in other organizations and may not translate to private practice workflows. Extrinsic factors such as time of day, day of week, and more granular clinical factors such as chief complaint are all likely important predictors of RT not included in this model. Future work should strive to integrate intrinsic (i.e., disease severity, sex, age) and extrinsic (i.e., time, date, scanner) factors to predict RT toward a more comprehensive and equitable model of radiologist efficiency.

Conclusion

The CSS is an NLP-based method for analyzing lumbar MRI reports which allows *quantitative* characterization of degenerative disease severity in a large population of imaged patients. The CSS summates the cumulative severity of 18 lumbar spine components providing a global marker of LMRI study complexity which is a significant and independent predictor of radiologist reporting time. The CSS may improve existing quality metrics by allowing one to weight metrics such as the RVU by study complexity.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-021-00462-1.

References

1. Enzmann DR. Radiology's value chain. *Radiology* 2012;263:243–52.
2. Leslie-Mazwi TM, Bello JA, Tu R, et al. Current procedural terminology: history, structure, and relationship to valuation for the neuroradiologist. *Am J Neuroradiol* 2016;37:1972–6.
3. Shah DR, Bold RJ, Yang AD, et al. Relative value units poorly correlate with measures of surgical effort and complexity. *J Surg Res* 2014;190:465–70.
4. Weiner SL, Tu R, Javan R, et al. Health care economics: a study guide for neuroradiology fellows, part 1. *Am J Neuroradiol* https://doi.org/10.3174/ajnr.A5381.
5. Rathnayake S, Nautsch F, Goodman TR, et al. Effect of radiology study flow on report turnaround time. *Am J Roentgenol* 2017;209:1308–11.

6. Dang PA, Freer PE, Humphrey KL, et al. Addition of tomosynthesis to conventional digital mammography: effect on image interpretation time of screening examinations. *Radiology* 2014;270:49–56.
7. Weinberg BD, Richter MD, Champine JG, et al. Radiology resident preliminary reporting in an independent call environment: multiyear assessment of volume, timeliness, and accuracy. *J Am Coll Radiol* 2015;12:95–100.
8. 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;24:3224–32.
9. Winklhofer S, Held U, Burgstaller JM, et al. Degenerative lumbar spinal canal stenosis: intra- and inter-reader agreement for magnetic resonance imaging parameters. *Eur Spine J* 2017;26:353–61.
10. Carrino JA, Lurie JD, Tosteson ANA, et al. Lumbar spine: reliability of MR imaging findings. *Radiology* 2009;250:161–70.
11. Drew B, Bhandari M, Kulkarni AV, et al. Reliability in grading the severity of lumbar spinal stenosis. *Clin Spine Surg* 2000;13:253–8.
12. Lurie JD, Tosteson AN, Tosteson TD, et al. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. *Spine* 2008;33:1605–10.
13. Miskin N, Gaviola GC, Huang RY, et al. Intra- and intersubspecialty variability in lumbar spine mri interpretation: a multireader study comparing musculoskeletal radiologists and neuroradiologists. *Curr Probl Diagn Radiol* 2020;49:182–7.
14. Fardon DF, Williams AL, Dohring EJ, et al. Lumbar disc nomenclature: version 2.0: Recommendations of the combined task forces of the North American Spine Society, the American Society of Spine Radiology and the American Society of Neuroradiology. *Spine J* 2014;14:2525–45.
15. Nodine CF, Kundel HL. The cognitive side of visual search in radiology. In: O'regan JK, Levy-schoen A, eds. *Eye Movements from Physiology to Cognition*. Amsterdam: Elsevier; 1987:573–82.
16. Nodine CF, Krupinski EA. Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. *Acad Radiol* 1998;5:603–12.
17. Cowan IA, MacDonald SL, Floyd RA. Measuring and managing radiologist workload: measuring radiologist reporting times using data from a Radiology Information System. *J Med Imaging Radiat Oncol* 2013;57:558–66.
18. ACR Prepares 2020 Medicare Physician Fee Schedule Final Rule Impact Tables.
19. Weiner SL, Tu R, Javan R, et al. Health care economics: a study guide for neuroradiology fellows, part 2. *Am J Neuroradiol* 2018;39:10–7.
20. Davenport DL, Henderson WG, Khuri SF, et al. Preoperative risk factors and surgical complexity are more predictive of costs than postoperative complications. *Ann Surg* 2005;242:463–71.
21. Bergersen L, Gauvreau K, McElhinney D, et al. Capture of complexity of specialty care in pediatric cardiology by work RVU measures. *Pediatrics* 2013;131:258–67.
22. Ishimoto Y, Yoshimura N, Muraki S, et al. Associations between radiographic lumbar spinal stenosis and clinical symptoms in the general population: the Wakayama Spine Study. *Osteoarthritis Cartilage* 2013;21:783–8.
23. Kalichman L, Cole R, Kim DH, et al. Spinal stenosis prevalence and association with symptoms: the Framingham Study. *Spine J* 2009;9:545–50.
24. Fu K-MG, Rhagavan P, Shaffrey CI, et al. Prevalence, severity, and impact of foraminal and canal stenosis among adults with degenerative scoliosis. *Neurosurgery* 2011;69:1181–7.
25. Cheung KMC, Karppinen J, Chan D, et al. Prevalence and pattern of lumbar magnetic resonance imaging changes in a population study of one thousand forty-three individuals. *Spine* 2009;34:934–40.
26. Manson NA, Goldberg EJ, Andersson GBJ. Sexual dimorphism in degenerative disorders of the spine. *Orthop Clin* 2006;37:549–53.