This information is current as of October 11, 2024.

Validation of the National Institute of Neurological Disorders and Stroke Spinal Cord Injury MRI Common Data Elements Instrument

J. Fisher, L. Krisa, D.M. Middleton, B.E. Leiby, J.S. Harrop, L.M. Shah, E.D. Schwartz, A. Doshi, S.H. Faro, F.B. Mohamed and A.E. Flanders

AJNR Am J Neuroradiol 2021, 42 (4) 787-793
doi: https://doi.org/10.3174/ajnr.A7000
http://www.ajnr.org/content/42/4/787

Validation of the National Institute of Neurological Disorders and Stroke Spinal Cord Injury MRI Common Data Elements Instrument

J. Fisher, L. Krisa, D.M. Middleton, B.E. Leiby, J.S. Harrop, L.M. Shah, E.D. Schwartz, A. Doshi, S.H. Faro, F.B. Mohamed, and A.E. Flanders

ABSTRACT

BACKGROUND AND PURPOSE: The National Institute of Neurological Disorders and Stroke common data elements initiative was created to provide a consistent method for recording and reporting observations related to neurologic diseases in clinical trials. The purpose of this study is to validate the subset of common data elements related to MR imaging evaluation of acute spinal cord injury.

MATERIALS AND METHODS: Thirty-five cervical and thoracic MR imaging studies of patients with acute spinal cord injury were evaluated independently in 2 rounds by 5 expert reviewers. Intra- and interrater agreement were calculated for 17 distinct MR imaging observations related to spinal cord injury. These included ordinal, categoric, and continuous measures related to the length and location of spinal cord hemorrhage and edema as well as spinal canal and cord measurements. Level of agreement was calculated using the interclass correlation coefficient and kappa.

RESULTS: The ordinal common data elements spinal cord injury elements for lesion center and rostral or caudal extent of edema or hemorrhage demonstrated agreement ranging from interclass correlation coefficient 0.68 to 0.99. Reproducibility ranged from 0.95 to 1.00. Moderate agreement was observed for absolute length of hemorrhage and edema (0.54 to 0.60) with good reproducibility (0.78 to 0.83). Agreement for the Brain and Spinal Injury Center score showed the lowest interrater agreement with an overall kappa of 0.27 (0.20, 0.34). For 7 of the 8 variables related to spinal cord injury, agreement improved between the first and second evaluation. Continuous diameter measures of the spinal cord and spinal canal using interclass correlation coefficient varied substantially (0.23 to 0.83).

CONCLUSIONS: Agreement was more consistent for the ordinal measures of spinal cord injury than continuous measures. Good to excellent agreement on length and location of spinal cord hemorrhage and edema can be achieved with ordinal measures alone.

ABBREVIATIONS: BASIC = Brain and Spinal Injury Center; CDE = common data element; CRF = case report form; NINDS = National Institute of Neurological Disorders and Stroke; SCI = spinal cord injury; ICC = interclass correlation coefficient

In 2006, the National Institute of Neurological Disorders and Stroke (NINDS) began a process to develop common data elements (CDEs) to provide a standardized method for the collection of clinical data related to neurologic diseases.[1-3] Recognizing that there is a lack of clear and consistent terminology for spine disorders, particularly spinal cord injury (SCI), in 2014, the NINDS convened a workgroup comprising expert stakeholders for the development of SCI CDE instruments that included clinical care assessments and imaging.[3-8] This new set of SCI CDE instruments aimed to increase the efficiency and value of clinical research studies and treatment, increase data quality, facilitate data sharing, and help educate new clinical investigators.[3] Investigators are expected to incorporate the CDE modules in grant applications and National Institutes of Health–funded research.

The MR imaging SCI CDE subset was created to be a comprehensive and standardized terminology for describing MR imaging findings in patients with SCI. This collection consists of a case report form (CRF) containing 35 discrete measures and responses divided into 4 main categories: general imaging characteristics,

Received August 7, 2020; accepted after revision October 26.

From the Departments of Radiology (J.F., D.M.M., S.H.F., F.B.M., A.E.F.), Physical Therapy/Occupational Therapy (L.K.), Biostatistics (B.E.L.), and Neurosurgery (J.S.H.), Thomas Jefferson University Hospital, Philadelphia, Pennsylvania; Department of Radiology (L.M.S.), University of Utah, Salt Lake City, Utah; Department of Radiology (E.D.S.), Saint Elizabeth's Medical Center, Brighton, Massachusetts; and Department of Radiology (A.D.), Mount Sinai Medical Center, New York, New York.

Originating institution: Department of Radiology, Thomas Jefferson University Hospital, Suite 1080B Main Building, 132 S. Tenth St, Philadelphia, PA, 19107.

Funding: Craig H. Nielson Foundation (16G.565).

Please address correspondence to Adam E. Flanders, MD, Department of Radiology, Thomas Jefferson University Hospital, Suite 1080B Main Building, 132 S. Tenth St, Philadelphia, PA, 19107; e-mail: adam.flanders@jefferson.edu; @BFlansteak

Indicates article with online supplemental data.

http://dx.doi.org/10.3174/ajnr.A7000

spinal injury features, canal and cord measurements, and chronic SCI features. The responses are of 3 types: Boolean, categoric, and an ordinal range representing specific anatomic locations. These measures were chosen to represent both objective and subjective assessment derived from routine clinical MR images. The workgroup codified these features using existing CDEs that have proved value in the published literature, and when ones did not exist, the workgroup developed the feature and the response parameters.

As with the development of any CRF used for a clinical trial or research, the goal is to provide an instrument that provides useful data representations that are reproducible across trained observers and institutions, require minimal cognitive effort, minimize ambiguity, and are both accurate and precise. Reproducibility of the observations through rigorous testing by multiple observers is a needed step to validate the instrument before clinical or research use. However, the evaluation process may not entirely reproduce the clinical environment in which it is meant to be used such that datasets and observers are overly prepared or optimized. Therefore, the goal of this study is to determine the inter- and intrarater reliability of the NINDS MR imaging CDEs when assessed by MR imaging experts with familiarity with SCI. We hypothesize that there will be good to excellent agreement (kappa >0.4) among the expert raters after limited training.

MATERIALS AND METHODS

Collection of the MR Imaging Dataset

This study was given exemption status from the institutional review board. An MR imaging evaluation dataset was assembled to represent a range of subjective visual MR SCI features that would be used to validate the NINDS SCI MR imaging CRF feature set. Examinations were preselected from a research MR imaging archive of adult patients with spinal injuries (without SCI) and patients with SCIs assembled from 12 different institutions. All examinations were previously de-identified and anonymized. Thirty-five suitable patients were selected by the principal investigator (A.E.F.) to represent the range of features and responses that constitute the NINDS adult SCI CDE collection. This number of cases was determined based on estimates provided by a statistical power analysis assuming an interclass correlation coefficient (ICC) of 0.8 and a desired 95% confidence interval of ± 0.1 (ie, 0.7–0.9).

The selected representative examinations consisted of 31 acute cervical and 4 thoracic SCIs of varying levels and degrees of severity. Each preoperative MR imaging examination consisted of a localizer, sagittal T1, sagittal T2, T2 axial gradient-echo, and axial T2 sequence. The emphasis on cervical injuries reflected the prevalence of cervical injuries in the general population at large.

Import of Dataset into the Web-Based DICOM Viewer

Case selection, training, and scoring of the evaluation set were performed using a cloud-based zero-footprint DICOM web viewer on a suitable monitor. All 35 cases were uploaded to this platform for each rater to review and report on from any location. Five pilot or test examinations were also included for the purpose of training and familiarizing the examiners with the platform. These cases were not used in the statistical analysis.

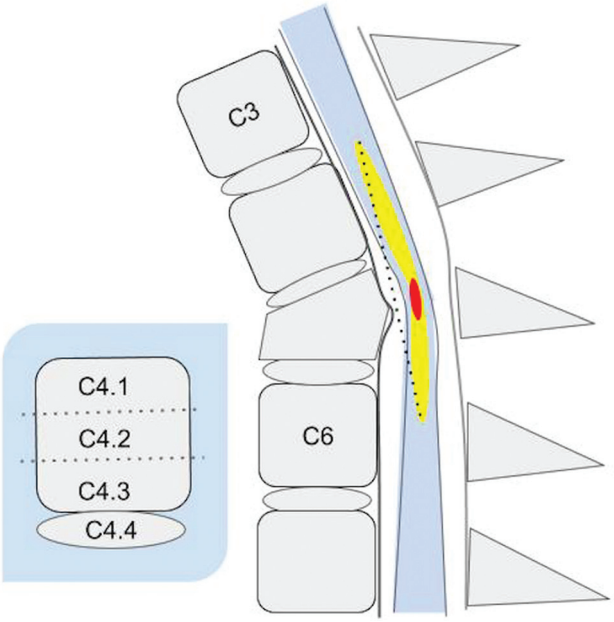

FIG 1. Graphic of a SCI on a sagittal T2-weighted image showing the anatomic location designations of the impact zone (*center*), rostral and caudal limits of spinal cord edema (*yellow*), and hemorrhage (*red*). By convention, each vertebral body is arbitrarily subdivided into 3 equal parts (designated as level.1, level.2, and level.3) with the intervertebral disc below the body as the fourth subpart (level.4). On this diagram, the rostral limit of edema is at C3.4, and the caudal extent is at C6.2. Hemorrhage (*red*) is demarcated by C5. 1 and C5.2. Lesion center is at C5.2. The *dotted line* represents actual continuous measurement of length of edema demarcated by the upper and lower boundaries on a T2-weighted image that a reviewer would create with electronic calipers.

Development of the CRF

The complete NINDS MR imaging CDE case report form contains 35 elements that encompass all facets of spinal trauma, including bony injury, surgical instrumentation, soft tissue injury, and SCI. Because the goal of this investigation was to evaluate features specific to SCI, a subset was selected for use. A limited-scope CRF was created using an extract of the most relevant NINDS SCI CDEs that could be used to describe SCI on clinical MR imaging. This included 8 specific imaging features focused on length and location of hemorrhage and edema in the spinal cord referenced by anatomic location. Each cervical segment is visually subdivided into 4 relatively equal subparts: upper, middle, and lower third of each vertebral body with the adjacent caudal intervertebral disc as the fourth part.[9] This anatomic reference is used to designate the location of the rostral or caudal extent of hemorrhage and edema as well as the injury center (Fig 1). The CRF also included an additional CDE that is not included in version 1 of the NINDS MR imaging CRF, called the Brain and Spinal Injury Center (BASIC) score,[10] which is a 5-part categoric assessment of spinal cord damage on axial T2-weighted images (Fig 2). The BASIC score is an ordinal scale that reflects the extent of hyperintensity on a select axial T2-weighted image.

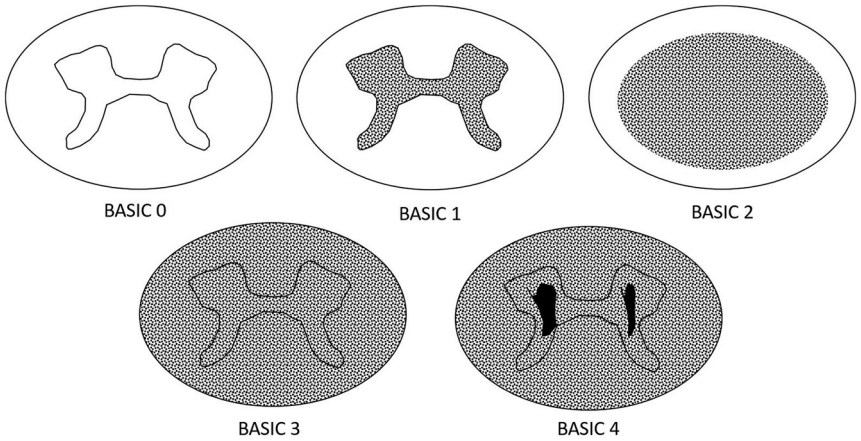

FIG 2. Graphic representation of the BASIC score CDE. The score is based on the extent of the cross-sectional T2-weighted abnormality. A score of 0 is normal. A score of 1 represents signal change in the central GM. A score of 2 represents signal change that extends beyond the central GM but does not involve the entire cross-sectional area. A score of 3 involves the entire cross-section of the spinal cord. A score of 4 features a grade III injury as well as hypointense foci in the central GM indicative of hemorrhage.

The second set of CDEs used in this study includes 9 continuous measures specified by the NINDS SCI CDE set that focus on discrete dimensions of the spinal canal and spinal cord that have been shown to be relevant to posttraumatic spinal cord dysfunction. These include measurements obtained directly at the level of injury compared with relatively normal canal or cord dimensions rostral and caudal to the injury (Figs 3–5).

The entire subset used for the study CRF consisted of 17 discrete elements derived from the original NINDS MR imaging SCI CDEs. Technical items related to scanner or software type, pre-existing hardware, injury subtype, and associated bony and soft tissue injury were not included in this assessment. The final CRF was published on a cloud-based form system with instructional material and visual references (Figs 1–5). Each rater was given access to the webform, and the data entries were automatically transferred to a spreadsheet (Google Forms and Google Sheets, Google Alphabet LLC) for analysis.

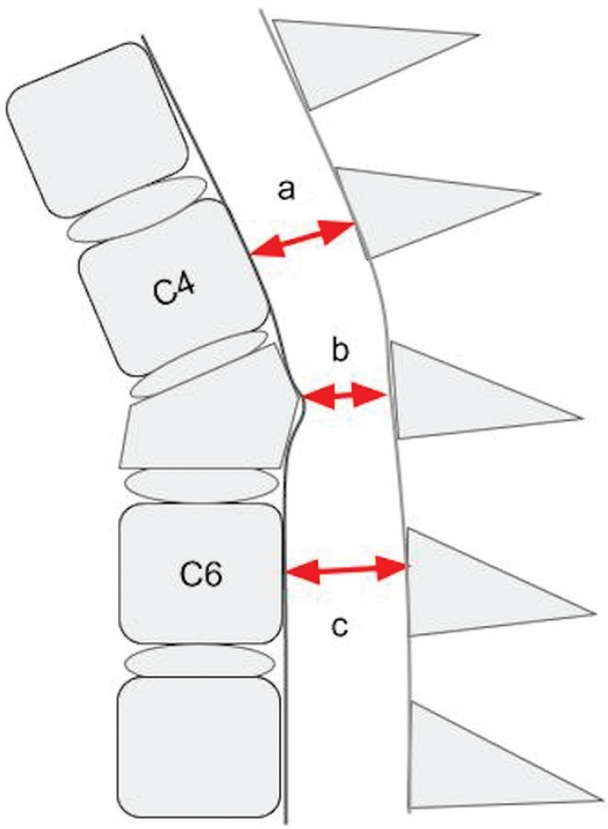

FIG 3. Graphic representation of a T2-weighted sagittal MR imaging illustrating an example of 3 key absolute measurements of the sagittal diameter of the spinal canal at the level of injury (*b*), above the level of injury (*a*), and below the level of injury (*c*). Reviewers were instructed to obtain the measurements from the dural boundaries instead of the cortical margins. Measurements obtained rostral and caudal to the injury level are made at the midbody level of the first normal-appearing body above and below the injury level, respectively.

Training of Expert Raters

Five distinct MR imaging examinations were selected to be used as a training set. A training guide was written and distributed based on the general and specific instructions given via the CRF module instructions (NINDS SCI CDEs version 1.0). This training guide outlined how to open the image viewer, view each case, make annotations, and record responses. This included visual examples of the controlled responses to each CDE (when applicable). Graphic references (Figs 1–5) were incorporated into the reviewer entry form to provide visual guidance for specific measurements.

Four Certificate of Added Qualification-certified neuroradiologists with expertise in spine imaging and 1 spine neurosurgeon with expertise in SCI were recruited and trained for correct use of the CRF using a combination of independent and virtual training. The incorporation of a clinician in the rater pool was intended to confirm that the CDEs could generalize to a nonradiology specialty. All raters were from 5 separate institutions, and all had greater than 10 years' experience in their specialty. The goal was to review the training manual, the terminology, the CDEs, and the response types. Examples were shown of each feature, criteria for assessment, and response. Training was kept to a single session with the expectation that the raters would apply their expertise to the criteria stipulated on the CRF using minimal supplemental instruction.

The 25 reviews on the 5 training cases were collected, outliers were identified based on consensus, and remediation was planned accordingly when necessary. This remediation included retraining of an individual reader(s) or modification

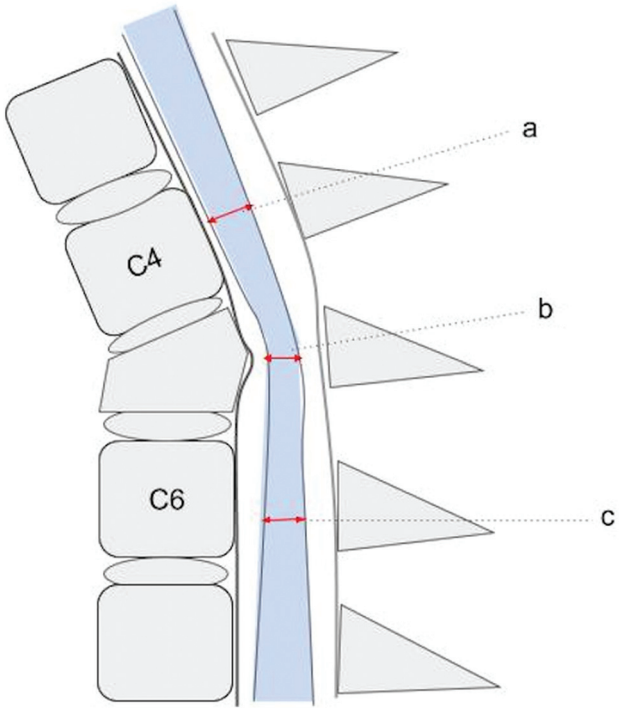

FIG 4. Graphic representation of a T2-weighted sagittal MR imaging illustrating an example of 3 key absolute measurements of the sagittal diameter of the spinal cord at the level of injury (*b*), above the level of injury (*a*), and below the level of injury (*c*). Measurements obtained rostral and caudal to the injury level are made at the midbody level of the first normal-appearing vertebral body above and below the injury level, respectively.

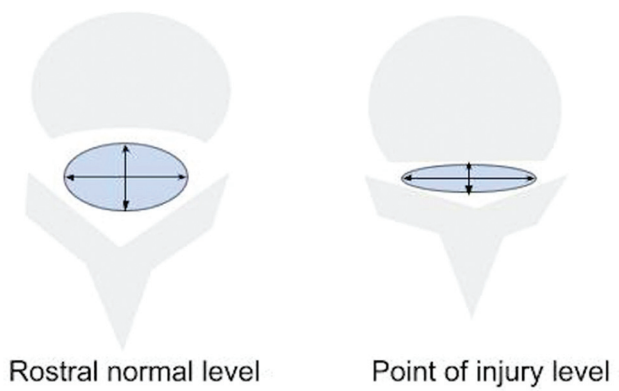

FIG 5. Graphic representation of an axial MR image of the spine illustrating the methodology for generating absolute measurements of the transverse and anteroposterior dimensions of the dural envelope (*arrows*). These were measured at 2 locations, the point of maximal compression on the axial dataset and at the first normal rostral midvertebral body level. Reviewers were instructed to estimate the margins from the dural envelope and not the bony cortical boundaries.

of a CDE description to mitigate ambiguity and revision of the training manual and the CRF. The NINDS CDEs and the BASIC scoring system were not modified in any way from their original form.

Data Collection

De-identified examinations were uploaded to the web viewer in batches of 5 numbered sequentially. The 5 expert raters reviewed each case independently and entered their responses into the web form accordingly. The readers were given the latitude to use the entire clinical imaging dataset to perform their evaluations. Data entries were checked in real time for missing entries, and evaluators were requested to complete entries as needed. No modifications to the values were allowed after the entries were recorded. One month after the completion of the first-round evaluation MR imaging dataset, the cases were randomized in order and reassessed by each of the readers to test both inter- and intrarater reliability. The second round of measurements and responses were performed de novo because the results from the first round of evaluation were not available to the readers.

Statistical Analysis

The CRF contained variables of 2 different data types: ordinal and continuous. Because the injury level convention for the entire spine (Fig 1) consists of more than 5 choices, the ordinal responses, which can range from C2.1 to C8.4, were treated as continuous variables. The BASIC score, with 5 ordinal choices, was assessed for agreement using the kappa statistic and a measure of percent agreement. For the continuous measures that included absolute measures of hemorrhage and edema length, as well as sagittal and transverse diameters of the spinal cord and spinal canal at multiple levels, each continuous element was rounded to the nearest integer (in millimeters). Assessing agreement of ratings among raters (ie, interrater reliability) and assessing the agreement of rating of a rater to the same patient's images over time (ie, test–retest reliability) was done with the ICC with confidence intervals.[11] Interrater ICCs were repeated, leaving out a single reviewer to assess for inconsistencies among the readers themselves.

RESULTS

Agreement results for the 7 SCI features are listed in Table 1. Reader agreement related to injury level, edema, and hemorrhage ranged from good to excellent (kappa >0.4). The ordinal CDEs that referenced the spinal cord feature of center and rostral or caudal extent of edema or hemorrhage to anatomic location (eg, Cx.1–Cx.4; Fig 1) demonstrated agreement among the 5 reviewers and showed an ICC ranging from 0.68 to 0.99. Moreover, reproducibility of these measures was excellent, ranging from 0.95 to 1.00. Agreement for absolute length of hemorrhage and edema was moderate, ranging from 0.54 to 0.60, with good reproducibility at 0.78 to 0.83. Agreement for the BASIC score was poor, showing the lowest interrater agreement, with an overall kappa of 0.27 (0.20, 0.34) with a percent agreement of 54.86 in round 1 and 0.42 (0.35, 0.50) and a percent agreement of 66% in round 2 with moderate intrarater agreement at 0.62 (0.53, 0.72). For 7 of the 8 variables related to SCI, agreement improved between the first and second evaluation with a modest reduction in agreement for absolute hemorrhage length (0.59 to 0.54).

Table 2 lists the agreement for morphologic continuous measurements of spinal canal and spinal cord diameter measured at, rostral to, and below the center of injury. Agreement on measures

Table 1: Inter- and intrarater agreement (ICC) and confidence intervals for CDE features related to SCI using 5 raters, 35 cases, and 2 rounds of evaluations[a]

	Interrater ICC: Round 1	Interrater ICC: Round 2	Intrarater ICC
Center of injury level	0.99 (0.99, 1.00)	0.99 (0.99, 1.00)	1.00 (1.00, 1.00)
Rostral extent of spinal cord edema	0.99 (0.98, 0.99)	0.98 (0.98, 0.99)	1.00 (1.00, 1.00)
Caudal extent of spinal cord edema	0.97 (0.95, 0.98)	0.97 (0.95, 0.98)	0.98 (0.98, 0.99)
Rostral extent of spinal cord hemorrhage	0.68 (0.56, 0.81)	0.70 (0.58, 0.82)	0.96 (0.95, 0.98)
Caudal extent of spinal cord hemorrhage	0.69 (0.56, 0.81)	0.69 (0.57, 0.82)	0.95 (0.93, 0.97)
Hemorrhage length (mm)[b]	0.59 (0.44, 0.74)	0.54 (0.39, 0.70)	0.78 (0.72, 0.84)
Edema length (mm)[b]	0.57 (0.42, 0.72)	0.60 (0.46, 0.75)	0.83 (0.78, 0.88)

[a] Quantitative and categoric features are specified per the methodology illustrated in Fig 1.
[b] Continuous data.

Table 2: Inter- and intrarater agreement (ICC) and confidence intervals for direct measures of spinal canal and spinal cord diameters derived from sagittal and axial MR images using the methodology featured in Figs 3–5 using 5 raters, 35 patients, and 2 rounds of evaluations

	Interrater ICC: Round 1	Interrater ICC: Round 2	Intrarater ICC
Sagittal cord diameter at level of injury (mm)[a]	0.83 (0.75, 0.91)	0.82 (0.74, 0.90)	0.84 (0.80, 0.89)
Sagittal cord diameter at first rostral segment above injury (mm)[a]	0.37 (0.20, 0.53)	0.27 (0.11, 0.43)	0.79 (0.73, 0.85)
Sagittal cord diameter at first caudal segment below injury (mm)[a]	0.43 (0.27, 0.60)	0.51 (0.36, 0.67)	0.69 (0.61, 0.77)
Sagittal canal measurement at level of injury (mm)[a]	0.80 (0.71, 0.89)	0.72 (0.60, 0.83)	0.85 (0.81, 0.90)
Sagittal canal measurement rostral to injury (mm)[a]	0.23 (0.07, 0.38)	0.31 (0.15, 0.48)	0.86 (0.82, 0.91)
Sagittal canal measurement caudal to injury (mm)[a]	0.56 (0.41, 0.71)	0.38 (0.21, 0.54)	0.90 (0.86, 0.93)
Anteroposterior diameter (mm) of the spinal cord at the point of maximal compression or epicenter of injury from axial image[a]	0.76 (0.65, 0.86)	0.69 (0.57, 0.82)	0.86 (0.82, 0.90)
Transverse diameter (mm) of the spinal cord at the point of maximal compression or epicenter of injury from axial image[a]	0.41 (0.25, 0.58)	0.33 (0.17, 0.50)	0.78 (0.72, 0.85)
Anteroposterior diameter (mm) of the spinal cord on axial image at the nearest adjacent rostral normal level[a]	0.37 (0.20, 0.53)	0.27 (0.11, 0.43)	0.82 (0.77, 0.87)

[a] Continuous data.

of the spinal cord and spinal canal using ICC varied substantially, ranging from poor to good categories (0.23 to 0.83). The most consistent and reproducible measure was sagittal cord and canal diameter at the level of injury in which interrater agreement ranged from 0.72 to 0.83. Intrarater agreement was more consistent at 0.84 and 0.85 for cord and canal, respectively. There was more of a disparity in agreement for measures that theoretically are less cognitively challenging, above and below the lesion center. Sagittal canal measurements obtained at the closest rostral segment showed consistently poor agreement at 0.23 and 0.31 with consistent performance with repeat measures at 0.86. Similarly, sagittal cord measurements at the closest normal rostral level also demonstrated poor overall agreement at 0.37 and diminished performance in the second review at 0.27. Intrarater performance was good at 0.79. Sagittal canal and cord diameter agreements obtained caudal to the level of injury also showed poor to moderate overall agreement (ICC, 0.38–0.56). Anteroposterior and transverse dimensions of the spinal cord measured in the axial plane also varied widely among the reviewers (ICC range, 0.27–0.76) with lowest agreement paradoxically at the levels that were traditionally cognitively easier to measure compared with the level of maximum compression or injury epicenter, where agreement was moderate to good. Intrarater performance for all 9 of these continuous measures fell in the moderate to excellent range (ICC, 0.69–0.90).

The Online Supplemental Data show the change of the aggregate ICC when a single observer is removed from the calculations for each of the continuous measures. This provides a general understanding of outlier observers as the source of variability of agreement. A substantial increase in ICC value for any 1 feature compared with the aggregate mean suggests outlier behavior. Although there was some source of variation attributed to 1 reader, in no case did the change in ICC value shift the degree of agreement into the next category. ICC for the BASIC score was also included for comparison, which remains in a similar performance category to the kappa. There was no consistent pattern for change from overall in inter- or intrarater agreement when ICCs were recalculated with a single observer removed.

DISCUSSION

The NINDS MR imaging CDE instrument for SCI was devised by a consensus expert panel based on existing literature and experience. The published instrument is endorsed by the NINDS as a recommended means for conducting clinical research and for evaluating specific features of MR imaging as they relate to SCI. No validation or reproducibility study was conducted after the publication by the NINDS. Therefore, the goal of this study was to validate the NINDS imaging CDE feature set for potential use in SCI clinical trials. To adhere to this objective, the experimental

design included several key elements: 1) the evaluation panel included a clinical (nonradiologic) domain expert, 2) we provided only a minimal amount of training to reduce training bias, and 3) we provided a complete clinical imaging dataset such that each reviewer expert was given the latitude to select the optimal image (s) to address each of the 17 CDE features without being directed to a specific image or series to make each assessment.

Several interesting patterns arise in examining the agreement among the groups of features. Absolute or continuous measures were less reliable or reproducible than ordinal or categoric features. That is, absolute measures requiring use of an electronic caliper were less reliable than semiquantitative ordinal methodologies overall, and this was reproducible on repeated measures. Selection of section and window or level also introduces measurement error caused by varying partial volume effects, resolution differences, and changes in cord angle between slices. Alternatively, the ordinal methodology for assessing edema or hemorrhage length and location based upon anatomic reference to the nearest adjacent vertebral segment[9] demonstrated consistent agreement in moderate, good, and excellent categories by ICC. This methodology has been successfully adapted for use in a series of SCI and MR imaging studies and was adopted by the NINDS as the preferred methodology for assessing extent and location of SCI on sagittal T2-weighted MR imaging because of the limited cognitive effort required. These results suggest that the method is sound and reproducible for characterizing lesion size and location on clinical MR imaging studies and more robust to real-world confounds than continuous distance metrics.

The BASIC score was developed as an axial adjunct to the sagittal method. It is an analogous visual measure of SCI by assessing the relative cross-sectional involvement of the spinal cord parenchyma in a single axial image. Using 7 evaluators with varied clinical backgrounds, Talbott et al[10] demonstrated mean and median kappa scores of 0.83 and 0.81 for BASIC assessment on preselected MR images of 20 patients with SCI. Kappa assessment for the BASIC CDE in our cohort performed substantially lower than reported by Talbott et al's original work (ICC 0.27 and 0.42) with only moderate intrarater agreement. The reason for this large disparity is not entirely clear, but the results do point to variations in how the studies were conducted. This includes more heterogeneity in our dataset with respect to scanner type and imaging protocol compared with Talbott et al's work. Image selection bias could have played a role in our lower performance. Moreover, in our analysis, there was no significant difference in agreement when removing the results of any single observer. One additional factor relates to reader inattention: the BASIC score was inserted as the last and final CDE in the CRF, and it is possible that this produced an inadvertent attention bias. Fortunately, this CDE is supplemental to the other SCI features.

Two related reliability studies were conducted by Fehlings et al[12] and Furlan et al.[13] These studies aimed to validate quantitative measures related to SCI. Fehlings et al[12] focused on determining reliability of 2 specific quantitative measures: maximum canal compromise and spinal cord compression in acute cervical SCI. Using 10 acute SCI MR imaging and CT cases rated by 28 spine surgeons, Fehlings et al[12] were able to detect an interobserver reliability range of 0.35–0.58 and intrarater reliability of 0.95–0.97 using the MR imaging.[12] These ranges of ICC for continuous responses are comparable to our own. Similarly, Furlan et al[13] conducted a study that measured cord compression and canal stenosis on 5 cases rated by 13 raters on 10 occasions. They reported an interrater reliability range of 0.55–0.61 and intrarater reliability range of 0.68–0.70 for their continuous variables.[13]

A number of spine-based reliability studies have been conducted to validate specific grading scales or systems for specific features or findings in spine imaging. These CDEs were created as a means to standardize reporting grading system for the evaluation of disc herniation and lumbar spinal stenosis.[14-16] A panel of 3 or 4 readers was trained to read a dataset of MR images on 2 separate occasions. The Spine Patient Outcomes Research Trial (SPORT) trained 4 clinical experts in spine MR imaging to use a grading system for determining nerve root compression after a lumbar intervertebral disc herniation.[14] They found overall intrarater kappa coefficients of 0.90, 0.84, and 0.63, respectively. Interrater reliability was found to be 0.81, 0.54, and 0.47.

Two other studies by Pfirrmann et al[15] and Schizas et al[16] used similar methods. Pfirrmann et al's[15] grading system determined an intraobserver kappa range of 0.72–0.77 and an interobserver kappa range 0.62–0.67. Schizas et al[16] found an intraobserver stenosis morphologic grading of 0.65 and interobserver kappa of 0.44.[16] These studies demonstrate that there is a broad range of agreement and reproducibility of spine MR imaging features and that agreements in the moderate or good range may still be adequate both to drive clinical decision making and to categorize patients for research and clinical trials. The magnitude of our agreement figures falls in line with other MR spine imaging reliability assessments and shows greatest value in characterizing SCI on sagittal T2-weighted images.

Limitations of this study included uncontrolled technical variations such as the browser type, screen resolution, and luminance of the monitors and environment used for the evaluations at the 5 different locations, which may have had an effect on perceptual abilities. Because the order of the responses on the CRF may have played a role in diminishing agreement for the items that appeared near the end of the CRF such as the BASIC score, one mitigation strategy might have been to reorder the 17 items for the second round of responses. Availability of the PACS presentation state of the annotated images from each observer might have provided direct comparison of the image selection and annotations, which may have provided some insight into the individual variation of absolute linear measures, choice of images, optimal window or width, and level of the spine and spinal cord dimensions on sagittal and axial images. Because the presentation state was not readily accessible as discrete data, this additional analysis was not pursued.

In summary, this investigation has demonstrated that the NINDS SCI MR imaging CDE set provides a valid and reproducible instrument for documenting the MR imaging features of SCI for clinical trials research, radiology reporting, and ultimately clinical decision making. The levels of agreement for the spinal cord features exhibited good to excellent agreement with multiple independent observers. Absolute measures of injury and dimensions of the residual spinal canal and spinal cord showed lower

reliability with repeat measures than ordinal semiquantitative measures overall. However, the magnitude of agreement was shown to be equivalent to prior multireader assessments of similar features. The recommended NINDS CDE system can be relied on for obtaining consistent results from domain experts with the minimum requisite amount of training.

Disclosures: Joshua Fisher—*RELATED: Grant:* *; *UNRELATED: Employment:* . Scott Faro—*RELATED: Grant:* See information from first author, Dr. Adam Flanders*; *Consulting Fee or Honorarium:* See information from first author, Dr. Adam Flanders, *Comments:* Reimbursed to provided quantitative analysis of the imaging data. Adam Flanders—*RELATED: Grant:* Nielson Foundation, *Comments:* Grant support for the research*. Laura Krisa—*RELATED: Grant:* Craig H. Neilsen Foundation*; *UNRELATED: Grants/Grants Pending:* NIH. James Harrop—*UNRELATED: Consultancy:* Depuy, Globus, Stryker Spine, *Comments:* Consulting and education honorariums not related to paper. Eric Schwartz—*RELATED: Consulting Fee or Honorarium:* Craig H. Nielson Foundation; *UNRELATED: Royalties:* LWW, *Comments:* Book royalties. Feroze Mohamed—*UNRELATED: Board Membership:* NIH. *Money paid to institution.

REFERENCES

1. Grinnon ST, Miller K, Marler JR, et al. **National Institute of Neurological Disorders and Stroke Common Data Element Project—approach and methods.** *Clin Trials* 2012;9:322–29 CrossRef Medline
2. Sheehan J, Hirschfeld S, Foster E, et al. **Improving the value of clinical research through the use of Common Data Elements.** *Clin Trials* 2016;13:671–76 CrossRef Medline
3. National Institute of Neurological Disorders and Stroke. **NINDS Common Data Elements Database 2006**, n.d. https://commondataelements.ninds.nih.gov
4. Mulcahey MJ, Vogel LC, Sheikh M, et al. **Recommendations for the National Institute for Neurologic Disorders and Stroke spinal cord injury common data elements for children and youth with SCI.** *Spinal Cord* 2017;55:331–40 CrossRef Medline
5. Bauman WA, Wecht JM, Biering-Sørensen F. **International spinal cord injury endocrine and metabolic extended data set.** *Spinal Cord* 2017;55:466–77 CrossRef Medline
6. Biering-Sorensen F, Charlifue S, Devivo MJ, et al. **Incorporation of the International Spinal Cord Injury Data Set elements into the National Institute of Neurological Disorders and Stroke Common Data Elements.** *Spinal Cord* 2011;49:60–64 CrossRef Medline
7. Biering-Sorensen F, Alai S, Anderson K, et al. **Common data elements for spinal cord injury clinical research: a National Institute for Neurological Disorders and Stroke project.** *Spinal Cord* 2015;53:265–77 CrossRef Medline
8. Biering-Sorensen F, Charlifue S, Devivo MJ, et al. **Using the spinal cord injury common data elements.** *Top Spinal Cord Inj Rehabil* 2012;18:23–27 CrossRef Medline
9. Flanders AE, Spettell CM, Tartaglino LM, et al. **Forecasting motor recovery following cervical spinal cord injury: value of MR imaging.** *Radiology* 1996;201:649–55 CrossRef Medline
10. Talbott JF, Whetstone WD, Readdy WJ, et al. **The Brain and Spinal Injury Center score: a novel, simple, and reproducible method for assessing the severity of acute cervical spinal cord injury with axial T2-weighted MRI findings.** *J Neurosurg Spine* 2015;23:495–504 CrossRef Medline
11. Koo TK, Li MY. **A guideline of selecting and reporting intraclass correlation coefficients for reliability research.** *J Chiropr Med* 2016;15:155–63 CrossRef Medline
12. Fehlings MG, Furlan JC, Massicotte EM, et al. **Interobserver and intraobserver reliability of maximum canal compromise and spinal cord compression for evaluation of acute traumatic cervical spinal cord injury.** *Spine (Phila Pa 1976)* 2006;31:1719–25 CrossRef Medline
13. Furlan JC, Fehlings MG, Massicotte EM, et al. **A quantitative and reproducible method to assess cord compression and canal stenosis after cervical spine trauma: a study of interrater and intrarater reliability.** *Spine (Phila Pa 1976)* 2007;32:2083–91 CrossRef Medline
14. Lurie JD, Tosteson ANA, Tosteson TD, et al. **Reliability of magnetic resonance imaging readings for lumbar disc herniation in the Spine Patient Outcomes Research Trial (SPORT).** *Spine (Phila. Pa. 1976)* 2008;33:991–98 CrossRef Medline
15. Pfirrmann CWA, Dora C, Schmid MR, et al. **MR image-based grading of lumbar nerve root compromise due to disk herniation: reliability study with surgical correlation.** *Radiology* 2004;230:583–88 CrossRef Medline
16. 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;35:1919–24 CrossRef Medline