ORIGINAL ARTICLE

Spinal cord ability ruler: an interval scale to measure volitional performance after spinal cord injury

R Reed^{1,13}, M Mehra^{2,13}, S Kirshblum³, D Maier⁴, D Lammertse⁵, A Blight⁶, R Rupp⁷, L Jones⁸, R Abel⁹, N Weidner⁷, EMSCI Study Group¹⁰, SCOPE¹¹, A Curt¹² and J Steeves^{1,11}

Study design: Retrospective statistical analysis of database.

Objective: Spinal cord injury (SCI) clinical trials are challenged to enroll participants, and early trial outcomes have often been equivocal. We hypothesized that a specifically designed novel true linear interval-scaled outcome measure targeted to simultaneously track a broad range of SCI will enable more inclusive enrollment of participants and valid comparisons of functional changes after SCI.

Methods: To define a single SCI measurement framework, we used items from existing measures. To evaluate linearity and validity of the measure, we used rigorous psychometric Rasch analysis on two data sets from over 2500 traumatic SCI participants (all levels and severities of SCI) within the EMSCI (European Multicenter study about SCI) database.

Results: Volitional performance was found to be the unidimensional construct that would detect and track a treatment effect from a central nervous system-directed therapeutic. Along with early evidence for voluntary neurological control of upper-extremity muscle contractions, volitional performance is best described by goal-directed activities of daily living that are increasingly difficult to re-acquire when activity within more caudal spinal segments is required. Validity of the Spinal Cord Ability Ruler (SCAR) as a linear interval construct was confirmed with Rasch analysis. All measurement items were properly ordered, as well as being precise and stable across clinically relevant groups. Only 5/24 items had some misfit. Targeting was excellent over time after SCI, with few gaps and only modest floor and ceiling effects (3% each).

Conclusions: SCAR is a quantitative linear measure of volitional performance across an inclusive range of tetraplegic and paraplegic SCI.

Spinal Cord (2017) **55**, 730–738; doi:10.1038/sc.2017.1; published online 21 March 2017

INTRODUCTION

Spinal cord injury (SCI) is a rare disorder (12 500 injuries per year in the US) that represents a significant unmet medical need.¹ Despite the conduct of clinical trials with different promising candidate treatments for SCI, based on well-defined animal models, it has been difficult to convincingly demonstrate a treatment effect in humans. Part of the uncertainty surrounding SCI clinical trial outcomes might be removed if there was an outcome tool that created a more sensitive and linear interval-level measure, which could be utilized for measuring changes in volitional performance in most potential SCI trial participants, regardless of level or severity of injury.

Rating scales have been designed to describe or classify various aspects of impairment or disability after SCI. The International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI), the Functional Independence Measure (FIM) and the Spinal Cord Independence Measure (SCIM) represent commonly used ordinal scales to assess outcomes after SCI.^{2–5} The total score of these multidimensional rating scales is computed by summation of the scores from individual items in the scale. However, due to the ordinal

nature of each item in these scales, it is not clear to what extent a change in the total score accurately represents the overall functional capacity of an individual living with SCI.³ The interval between any two values within an item may be unknown or unequal (that is, nonlinear). Furthermore, a 1-point difference in an easy item is likely not the same as a 1-point difference in another more difficult item.

For example, ISNCSCI classifies patients by upper- and lower-extremity muscle strength and sensory perception. The motor score is based on the voluntary strength of key muscles (myotomes) and is similar to the well-described 6-point Medical Research Council classification of muscle strength.⁶ Scoring methods such as changes in upper-/lower-extremity motor scores (UEMS/LEMS) are frequently applied as endpoints in interventional clinical studies. However, it is common understanding that this is an ordinal scoring method, and any change along the motor scores ranging from 0 to 5 is non-interval in nature and limits quantitative comparisons across different levels of motor function. People living with SCI gain little functional ability when improving from 0 to 1, whereas a similar magnitude of change (that is, a 1-point motor score change) with a transition from 3 to 4

¹ICORD, University of British Columbia and Vancouver General Hospital, Blusson Spinal Cord Centre, Vancouver, British Columbia, Canada; ²Tigermid-MacroStat LLC, Gaithersburg, MD, USA; ³Kessler Institute for Rehabilitation and Rutgers New Jersey Medical School, West Orange, NJ, USA; ⁴Trauma Center Murnau, Murnau, Germany; ⁵Craig Hospital, Englewood, CO, USA; ⁶Acorda Therapeutics Inc., Hawthorne, NY, USA; ⁷Spinal Cord Injury Center, Heidelberg University Hospital, Heidelberg, Germany; ⁸Craig H. Neilsen Foundation, Encino, CA, USA; ⁹Trauma Center Bayreuth, Bayreuth, Germany; ¹⁰EMSCI Study Group, C/O Balgrist University Hospital, Zurich, Switzerland; ¹¹SCOPE, C/O ASIA Central Office, Richmond, VA, USA and ¹²Spinal Cord Injury Center, University Hospital Balgrist, University of Zurich, Zurich, Switzerland

¹³These authors contributed equally to this work.

Correspondence: Dr J Steeves, ICORD, University of British Columbia and Vancouver General Hospital, Blusson Spinal Cord Centre, 818 West 10th Avenue, Vancouver, British Columbia, Canada V5Z 1M9.

E-mail: steeves@icord.org

Received 22 November 2015; revised 16 December 2016; accepted 21 December 2016; published online 21 March 2017

may actually enable functionally effective movements.⁷ Recent evaluations of motor scores in SCI suggest that the UEMS and LEMS consist of distinctly different domains, and should be reported separately.⁸ Finally, total UEMS and LEMS scores are difficult to interpret, as the same total score can be obtained with many distinct combinations of individual item scores across the key muscles, and therefore they lead to differing functional abilities.⁸

Functional activity assessments (that is, any version of FIM or SCIM) describe independence in activities of daily living (ADL). The SCIM was developed specifically to assess multiple dimensions (domains) after SCI and includes commonly accepted voluntary ADL tasks, as well as involuntary (autonomically influenced) bladder, bowel and respiratory functions.² Despite its ordinal, multidimensional nature, a sum of weighted scores is typically calculated. Some favorable psychometric properties (that is, test/re-test and inter-rater reliability) have been reported on SCIM.⁹ Interestingly, a Rasch analysis study of SCIM reported distortion or misfit within several items, although no changes were made to the SCIM assessment scores to improve the validity of the assessment.¹⁰ Despite concern about the non-interval nature of SCIM and ISNCSCI, these scales are being applied in clinical settings to evaluate recovery patterns across different types of human SCI,^{11,12} although it is likely that each of these measurement tools may not be equally effective in detecting and tracking a subtle or meaningful change across all levels and severities of SCI.

Current incremental and sequential enrollment strategies have led to expensive and operationally challenging multi-year trials and/or truncated studies with insufficient number of subjects (power) to detect a potential benefit of a therapeutic intervention. However, in an already rare disorder, what is needed for an inclusive enrollment strategy is a linear (interval level) outcome measurement tool that can be accurately and reliably utilized to track functional changes across a broad range of SCI levels and severities. Achieving such an outcome measurement tool does not remove the need to compare similar (homogeneous) cohorts within an inclusive trial protocol with stratification as appropriate.^{13,14}

There is nothing inherently wrong with categorizing motor functions or independence in ADLs to describe clinical conditions. It should be remembered that ISNCSCI, FIM, SCIM and many other classification scales were not designed as trial outcome measures. According to well-developed statistical knowledge using Rasch measurement theory, the current ordinal ISNCSCI and SCIM scoring have limitations in measuring treatment benefit accurately.

As in the physical sciences, attempts to make meaningful measurement estimations using categorical or ordinal rating scales should focus on one dimension at a time and not attempt to conflate 2 or more domains into a single score. In addition, measurements should also be based on hierarchical 'more than/less than' decisions that are invariant across important participant characteristics such as SCI level and severity. Finally, any proposed measure should be quantitatively tested and validated using rigorous psychometric methods to determine the following: (1) a legitimate interval score can be generated; (2) the selected scoring options are reliable and free from excessive variability (for example, assessment error); and (3) the scale maintains adequate construct validity and measures the attributes it purports to measure.^{15,16}

The SCAR project was intended to determine whether a subset of already existing assessment items could be combined and modified with good measurement principles to create and validate an interval measure with precision across a broad range of SCI severities and injury levels. SCAR provides an accurate scale for detecting potential

treatment effects, allowing for inclusive SCI protocols and thereby improving clinical trial efficiency.

MATERIALS AND METHODS

Data selection

De-identified data from the European Multicenter Study about Spinal Cord Injury (EMSCI), collected from July, 2001 to December, 2015, were extracted for analysis. The Ethical Committee of the Canton of Zurich, Switzerland, has previously approved the European Multicenter study about Spinal Cord Injury (EMSCI) project, upon which this project is based, and this approval is valid for any statistical data analysis presented here. EMSCI contains among others ISNCSCI and SCIM assessments from a broad range of individuals with traumatic and ischemic SCI. We accessed traumatic SCI records only for this analysis. Records consisted of tetraplegic and paraplegic injury (45% and 55%, respectively), with all degrees of severity (AIS A (45%), B (13%), C (17%), D (24%) and E (1%)). Baseline neurological level of injury (NLI) from C1–S2 were included. Male participants comprised 79% of records, and age at injury ranged from 13 to 94.

Two mutually exclusive and randomly selected sets of 3000 observations were extracted from this EMSCI traumatic injury data set. Participants had assessments at one or more of the following time points: 0 to 15 days, 16 to 40 days, 70 to 98 days, 150 to 186 days and 300 to 400 days after injury. The first data set (model development data set) was used to evaluate the psychometric properties of SCAR and contained data from 1305 participants. Results were reviewed, and changes to the analysis were proposed where clinically appropriate. The second data set (validation data set) contained data from 1280 subjects and was used to independently confirm the initial Rasch analysis results. After this confirmation, the entire EMSCI traumatic data set (7518 records from 2777 participants) with dates of assessment up to December 2015 was used to produce the final scale metrics.

Development of hypotheses

A combined working group (organized by SCOPE and EMSCI) of measurement experts, statisticians and SCI clinical investigators met on an ongoing basis for the following reasons: (1) to identify a single fundamental concept to measure clinical outcome targets; (2) to generate testable hypotheses around a conceptual framework; (3) to select conceptually appropriate items from within ISNCSCI and SCIM that are clinically relevant; and (4) to rigorously test the psychometric properties of the assessment tool using Rasch analysis.

Conceptually, the working group proposed volitional performance as the unidimensional metric to measure CNS-directed treatment outcomes for all types of SCI studied in acute and sub-acute SCI trials. Volitional performance is defined as voluntary task-specific physical actions contributing to independence in ADLs (for example, specific SCIM items associated with voluntary movement). Furthermore, it was hypothesized that volitional performance items that involve the use of muscle groups increasingly caudal (distant) to the NLI are more challenging to re-acquire and would require recovery of CNS function at more caudal spinal levels to perform. Finally, underlying and often preceding the performance of any volitional performance items are testable upper-extremity muscle contractions (for example, voluntary movement against some or normal resistance). Thus, the UEMS items for each cervical cord segment can be a useful metric when SCIM items cannot be performed, such as very early after SCI.

There is, however, at least one type of cervical SCI that will not follow these hypotheses and that is central cord syndrome (CCS). CCS has not been rigorously defined, but is characterized by greater sustained impairment in the arms and hands relative to that in the legs (without necessarily a gradient from rostral to caudal cervical levels). As individuals with CCS will have persistent functional deficits within the upper extremity, they will not fit the hypothesized rostral-caudal segmental continuum for the reacquisition of more difficult tasks. Uniform and globally accepted diagnostic criteria of CCS are lacking.¹⁷ Through the evaluation of motor-evoked potentials, it has been shown that central and peripheral motor pathways, devoted to the motor control of the hands, are more affected after cervical SCI than those devoted to the lower limb.¹⁸ However, CCS is a problematic diagnosis without extensive electrophysiological measurement, and a syndrome that is still controversial as

Figure 1 Model of volitional performance after traumatic cervical SCI. Study participants with greater volitional motor performance would move toward the right in the conceptual model. At very acute time points after SCI, improvements in volitional motor activity within the cervical cord can be assessed by manual muscle testing (UEMS). As recovery progresses, ADLs are recovered in a rostrocaudal direction along the spinal cord, with the more difficult ADLs requiring functional neural circuits within increasingly more caudal spinal segments.

to its incidence and prevalence. A definition based on a difference of at least 10 points in the total UEMS and LEMS has been proposed, but total UEMS and LEMS scores are also difficult to interpret, and most expert advisors felt that applying this single criterion to the diagnosis of CCS is insufficient for research purposes.¹⁷

As CCS has been characterized by greater impairment in the hands relative to that in the legs and our model indicates that, for cervical injuries, items assessing more distal muscle groups responsible for ambulation tasks should have poorer outcomes than more proximal muscle groups responsible for hand-related tasks (for example, self-care). CCS was defined functionally for our purposes as a mild cervical SCI where SCIM items assessing voluntary ambulation (for example, walking) were functionally equal to or better than those SCIM items assessing hand function. There can be many reasons contributing to such a characteristic pattern, and CCS is but one explanation. Nevertheless, all participant data with this characteristic recovery pattern were removed from the analysis. This resulted in approximately 9.8% of records being removed, which is very likely a greater percentage of people than the number who actually have CCS.

Statistical analyses

Rasch methods as they relate to statistical analysis and the methodology used to validate outcome tools in clinical trials are explained in greater detail elsewhere.¹⁹ Application of the Rasch measurement model is a useful tool to construct measurement scales with good psychometric properties (for example, unidimensional, invariant, responsive to measure a change with an interval-level scale).^{15,16,19,20}

Rasch analysis fits a special type of logistic regression model that evaluates whether the selected items measure a single concept on a linear continuum. The central aspect of Rasch measurement is that a person's completion of an item/task is governed only by his or her ability, and how difficult each item is to perform. On the basis of this construct, Rasch analysis first models the probability of responses on each item assessed and then, using maximum likelihood estimation, establishes a total score for each participant by how he or she performed on each item in the scale. It evaluates the legitimacy of this total score by assessing how well the sample data performed as predicted by the Rasch model along the linear continuum. The difference between observed and modeled results is described as item fit and indicates the degree to which a reliable, valid measure is achieved.^{15,16,19,20}

Determination of item fit to the Rasch model was based on the combined assessment of several indicators as follows: the ordering of response options (that is, scoring choices for each item), two statistical indicators (fit residual and chi-square) and the evaluation of item characteristic curves or ICCs.^{15,16,19} Description of these indicators may be found in conjunction with the data

presented in the Results section. Items that do not fit within the acceptable range were further investigated for cause. Additional statistical characteristics of the Rasch model were also tested, including the following: scale-to-sample targeting (that is, how well the constructed scale measured the participants and how well the data from participants was able to construct the scale), local independence of items and scale reliability (Person Separation Index, or PSI, comparable to Cronbach's alpha).^{15,16,19,20} A Bonferroni adjustment was used to account for multiple statistical comparisons. Rasch analysis was performed using RUMM 2030.²¹ Statistical details of the Rasch methodology used and psychometric properties of SCAR will be included in a future publication.

RESULTS

Measurement concept and hypotheses

There is common agreement that recovery of motor function after SCI is of high clinical priority as it is fundamental for improved ADL outcomes. We chose to focus on volitional performance as the underlying single dimension for tracking recovery over time. Volitional performance was defined as task-specific physical actions that are voluntary in nature and contribute to independence in daily living. These physical actions focused on volitional ADLs as outlined in SCIM and include self-care, transfer and mobility items. Underlying the performance of these volitional ADLs, and testable at an early time point after SCI, are meaningful and controlled muscle contractions (defined as voluntary movement against some resistance). These voluntarily initiated and guided muscle contractions are the 'building blocks' of volitional performance. More importantly, at early time points after cervical SCI, SCIM items cannot often be measured, but cervical cord motor scores can be measured, and recovery of functional upper-extremity muscle contractions often precede the reacquisition of ADLs.

For a measure to be applicable for the assessment of all individuals with SCI, it must evaluate a wide spectrum of voluntary actions in a graded fashion, from complete tetraplegia with severe motor impairment of upper and lower limbs to less affected performance of volitional ADLs in people with more caudal incomplete paraplegic lesions. To ensure comprehensiveness of the measure, a modeled continuum of traumatic SCI was developed that identified the range of volitional performance deficits following SCI. In the model, some degree of paralysis or paresis occurs at all anatomical levels caudal to the NLI, whereas above the lesion there is normal function. Function below the lesion depends on several factors as follows: NLI, severity of SCI (relative degree of complete versus incomplete SCI), extent of any partially preserved functions after SCI, and the fact that lower- and upper-extremity muscles are innervated by more than one spinal segment.⁵ Depending on these factors, paralysis at all spinal levels below the lesion may be complete or partial, resulting in an inability or a limited ability to independently perform functional voluntary muscle contractions or volitional ADLs using muscle groups dependent on innervation from nearby spinal levels.

Thus, we hypothesized that residual voluntary control of muscle groups increasingly caudal to the NLI of SCI is decreased (not necessarily linearly), and therefore ADLs requiring the use of these muscle groups are increasingly more difficult to perform (that is, re-acquire). When no volitional ADLs can be performed without complete assistance (for example, at early time points after SCI), voluntarily controlled muscle contractions within the upper extremity still represent a measurable component of volitional performance, and extend the measurement of volitional performance to include and subsequently track participants from time points immediately after SCI (Figure 1).

Figure 2 Conceptual framework of SCAR measure. The conceptual framework for volitional performance (Figure 1) creates an objective interval measure after SCI. It has a scoring scale focussed on meaningful transitions extending from complete dependence to independence in task-specific physical actions without the need for adaptive devices. Volitional cervical cord motor activity (as assessed by UEMS) increases the linearity of the measure when the most basic volitional ADLs, as assessed by voluntary SCIM activities, cannot be performed independently (e.g., during the very acute period after cervical SCI). For these cervical motor scores, the desired recovery goal would be active movement against moderate to full resistance.

The SCAR conceptual framework

Broad consensus exists across the clinical field that fundamental ADLs consist of the following: feeding, grooming, bathing, dressing, transfers, bowel/bladder function and ambulation.²² Because of the dependence of respiration on high cervical cord function, respiratory function is also often monitored as an ADL item and included in SCIM.² Regardless of their importance after SCI, respiratory and bowel/bladder function have only some degree of voluntary control and are ultimately activated by involuntary mechanisms mediated by the autonomic nervous system.^{23,24} Therefore, these ADLs were not considered as being completely within a volitional performance measurement framework. As SCIM was developed specifically for SCI, all items from SCIM representing measurement of volitional ADLs were adopted for the framework, except most items from the respiration and sphincter management section of SCIM (SCIM sub-score 2). More importantly, the inclusion of respiration, bladder and bowel function automatically means the original SCIM scale is multidimensional, and thus it can never be developed into a continuous linear, interval measure; it remains an ordinal scale that describes a person on the basis of weighted categories across multiple domains. We were seeking to develop a continuous quantitative interval-level measure for SCI.

Where possible, level of independence in volitional SCIM items was confined to four options as follows: complete dependence on assistance from others, partial dependence on assistance from others, independence with devices and independence without devices. These response options or scores for each SCIM item were based on literature reviews of clinically meaningful activity changes after SCI^{2,3,22} and input from the working group. Thus the necessary re-scoring preserves the range found within conventional scores for individual SCIM items, but removes any misfit or disorder (i.e. variability in the selection or assignment of a score; for further explanation see below).

Likewise, the scoring options for each key upper-extremity (C5–C8) muscle contraction in INSCSCI were simplified into the following three choices that represent meaningful gradations in voluntary muscle activity: none, non-functional and functional muscle

contraction (that is, movement against moderate or full resistance, MRC ≥ 4).^{5–8} UEMS for T1 and LEMS for L2–S1 cord segments were not included in SCAR, as SCIM items adequately characterize volitional performance in this range (that is, functional paraplegia) and LEMS was redundant.

In brief, a total of 24 volitional performance items from INSCSCI and SCIM using the above defined response options were included in the conceptual framework outlined in Figure 2. Justification for the re-scoring is provided below.

Measurement properties

It is often believed that more scoring options provide more useful information and allow researchers to better discriminate meaningful transitions for an item. Although more scale points have the potential to increase precision, there is a limit to a person's ability to reliably differentiate between adjacent scores. It is equally possible that beyond a certain point, numerical scoring options will not enable researchers to reliably distinguish a meaningful transition, but instead describe a change that has little or no utility and result in inconsistencies across examiners.²⁵ This loss of clarity is demonstrated quantitatively by Rasch analysis when scores for an item have disordered probabilities for being selected and do not follow a logical sequence of increasing intensity in function or ability.^{15,16,19,20} This indicates that the item is not working as intended to measure the concept, and results in increased variability in item scoring and a reduction in the scale's reliability.

An example of an ordered SCAR item (mobility indoors) compared to the same item using the original scoring options defined by the SCIM, is graphically displayed in Figure 3 (a and b, respectively). The probability of a response on the y-axis is plotted against the ability of a population to achieve distinct scores along the x-axis. For the item to be reliably ordered, each response option should have the highest probability of occurring at some point along the x-axis, as identified by the peak of each bell-shaped curve being above all others. This order would be expected if the response options represent an increasing difficulty to perform or achieve the item. The scoring options in Figure 3a (original SCIM scoring) are disordered because the curve for scores 4 to 7 never provide a distinct peak that is above all others at some point along the x-axis. This indicates a source of inconsistency in scoring, as it is possible to score a participant with very similar volitional performance anywhere from 3–8 for this item. Conversely, when the response options were limited to only denote meaningful transitions (Figure 3b), the scoring options become ordered, so each response option (score) is most likely to be selected until the next ordered score becomes the more likely occurrence, based on the person's volitional performance.

In accordance with our findings, Rasch analysis of the original SCIM III scoring options by Catz and colleagues indicated that 8 of 19 identifiable items (50%) had disordered thresholds.^{2,10} Ignoring this disorder was not an option if we wanted to generate a quantitative continuous interval-level linear scale such as the SCAR. Overcoming the inconsistent item scoring on the SCIM II/III by the identification and introduction of meaningful response options (Figure 3a) both improved scale precision and generated a legitimate total SCAR score that fit the Rasch model for each EMSCI participant, regardless of level or severity of SCI.

Part of the activity of the working group was to evaluate interim results of the Rasch analysis during the development phase with the model development data set to understand how items and response options (scores) were performing as a measure of independence in volitional performance. After review, the 'Grooming' item was

re-scored by the working group to combine the response options of 'independence with devices' and 'independence without devices'.

These review processes were a fundamental step to improve the scoring response options and required a careful discussion between clinicians and statisticians to find the most clinical meaningful and productive scoring scale. After these adjustments were made, all scoring options displayed proper ordering in the model development data set. Analysis of the ICC curves indicated an overall good concurrence between observed and expected values. Scale reliability

was supported by a high Pearson Separation Index (PSI; a measure analogous to Cronbach's alpha) and was calculated at 0.97 out of 1. The psychometric characteristics of the validation (second) data set were very similar to that of the model development data set, including the following: targeting, ordering, residual correlations and ICC curves with the two statistical tests of fit being similar. The PSI was also 0.97 within the second data set.

On the basis of the similarities between the test and validation data set, the entire EMSCI data set with 8486 observations across 2923 participants was then used to construct SCAR. Reliability of the SCAR remained high and consistent across repeated EMSCI data sets at 0.97. A description of the psychometric properties of the SCAR, using the entire data set, is described below.

Scale-to-sample targeting

SCAR was designed to measure the full range of volitional performance deficits regardless of level or severity of SCI. Assessment time points up to 1 year after SCI were available for inclusion in our analysis. For a linear interval-level scale to target this population well, items and the scoring options should measure different and meaningful aspects of volitional performance and, based on the difficulty in performing the items, be distributed broadly and regularly (like notches on a ruler) along the entire range of capability for the sample population, ensuring comprehensive measurement precision along the entire range of SCAR. As SCAR was determined to be an interval-level scale and Rasch analysis was used to create a total score for each participant, it is legitimate to combine each participant's score and the difficulty rating of the item on the same scale to inspect scale-to-sample targeting, as shown in Figure 4.

The upper histogram in Figure 4 represents the distribution of participant assessments over the first year after SCI. His/her volitional performance score distributes each participant relative to others, with greater ability located toward the right. The lower histogram represents the relative location of the various scores for each ISNCSCI and SCIM item. The location of each item score is determined by how all the SCI participants performed on that item, with scoring options from items that fewer participants could successfully perform (that is, those with increased difficulty) appearing toward the right.

Response options or scores for SCAR are distributed widely across the entire range of the measurement scale, and the majority of participants are situated widely, but predominantly in the middle

Figure 3 Disordered and ordered scoring of SCIM activity tasks—SCIM mobility indoors item before (a) and after (b) re-scoring. The probability of a scoring option being used to describe a person's capability for each designated SCIM activity task is plotted against the likelihood of that score within the item's rating scale being distinctly selected (>50%). The example here is the mobility indoors item from the SCIM using the original scoring options defined by the SCIM II/III (a) compared to the same item using re-scored options according to Rasch analysis (b). If each score within the item's scale (see tables) represents a meaningful change in the performance of the item, then each score should have the highest probability of distinctly occurring along the x-axis. (a) Scoring options (0–8) are disordered because the probability curves for score 1, 5, 6 and 7 never achieve a distinct peak that is above all others at some point along the x-axis. This indicates it is possible to score participants with very similar volitional performance with different and variable scores. (b) Conversely, when the response options were limited to only denote meaningful transitions for a SCIM item, each score (0–3) has the highest probability of being selected until the next meaningful improvement in volitional performance is attained and becomes more likely.

Figure 4 Person-item distribution of SCAR across tetraplegic and paraplegic AIS A–D participants at acute through chronic assessment time points. The upper histogram represents the frequency distribution for volitional motor performance of 2777 people living with traumatic SCI (tetraplegic and paraplegic, without suspected central cord syndrome) against increasingly difficult activity tasks (from left to right in lower histogram) over the first year after SCI. Participants with greater volitional motor performance are located toward the right in the upper histogram. The lower histogram represents the relative location of the various scores along the SCAR item scale with more difficult scoring options placed toward the right. The person-item distribution provides visual confirmation of the most responsive part of the scale.

range of the scale (Figure 4). Scale-to-sample targeting of the SCAR indicates good targeting with only some small gaps at the left and right margins of the scale. Ceiling effects of only about 3% and floor effects of around 2.4% were observed in the data set (Figure 4) and are to be expected because of the inclusion or deterioration of some participants graded as having no volitional segmental motor activity at C5 (located on the left side and interpreted as a floor effect) and those assessed to have a perfect score on the SCAR and have subsequently moved to a location at the far right-side edge of the scale (that is, a ceiling effect).

Ordering of scale items and response options

Beyond targeting, Rasch analysis highlights the strengths and limitations of a scale by demonstrating that items and response options map out along a proper hierarchical 'more than/less than' structure so that intensity of the attribute can always be estimated along a linear continuum. Figure 5 provides an alternative display of SCAR items with increasing difficulty arranged vertically by the average difficulty of their respective scoring option thresholds (listed in the legend), with more difficult items toward the bottom. The relative ordering vertically is as would be expected based on both the volitional performance model and clinical judgment of the working group. Response options (scoring) for each item are also located with proper ordering in a logical sequence of increasing intensity horizontally, with more difficult categories toward the right. The transition points between the response options for each item form a diagonal line of increasing intensity from top left to bottom right across the figure, indicating each item is working as a series of meaningful transitions intended to measure a particular range of volitional performance. When these transition points are represented horizontally along the volitional performance continuum, they become the notches on the SCAR scale, as shown by the lower histogram in Figure 4.

To represent an individual's location along the SCAR depicted in Figure 5, you would extend a vertical line upwards from the base that best describes a person's SCAR score in terms of the items they can accomplish at that assessment time. Any subsequent change in the person's ability to successfully achieve more or less ISNCSCI or SCIM

items during recovery would move the individual's location to the right or left, respectively. It should be noted that a logit score can be easily transformed mathematically into a score on a range from 0 to 100 using the following formula:

$$\text{Value on 0 to 100 scale} = \frac{(\text{value on logit scale} - \text{lowest value on logit scale})}{(\text{highest value on logit scale} - \text{lowest value on logit scale})} \times 100$$

Statistical measures of fit

The residual test of fit examines the response of each participant to each item, and provides evidence that an item might discriminate between levels of transition for the concept (in this case volitional performance) either more or less than the model predicts. In Rasch analysis, item over-discrimination (negative values) and under-discrimination (positive values) are equally undesirable.¹⁹ For the entire data set, the majority of items (19 of 24 items) had residuals that fit within the acceptable range.¹⁹ Of the five items that had fit residuals outside the acceptable range, two items 'Use of Toilet' and 'Bathing Upper Body' had larger values, whereas the remainder 'Transfer from bed to Wheelchair', 'Mobility Over Moderate Distances' and 'Bathing Lower Body' were just slightly outside the acceptable range.

The χ^2 test of fit examines the item-scale relationship. The mean score for groups of participants with similar total scores (classes) are estimated for each item, and the magnitude of divergence of this observed mean score from the value predicted by the model is estimated. If the observed to expected difference is statistically significant, the item should be further examined for cause. In the entire data set, one item 'Bathing Upper Body' had χ^2 probabilities that were statistically significant after adjustment for multiple comparisons.

ICC curves and item stability across clinically meaningful subgroups

Another indicator of item fit to the Rasch model is the ICC curve. This indicator provides a tool to assist in the analysis of discrepancies in the two statistical measures of fit through visualization and statistical

Figure 5 SCAR items by difficulty and response options. SCAR items are arranged vertically by their relative difficulty to perform after SCI, with more difficult items situated toward the bottom. Each item's scoring options are indicated by the horizontal scale (at bottom of figure), with greater independence in each item increasing toward the right. Note there are only three transitions for UEMS (no movement, non-functional movement and normal movement), but often four transitions for selected volitional SCIM items (dependence, partial dependence, independence with assisting devices and independence). The transition points between the scoring options for each item form a diagonal line of increasing demands from top left to bottom right, indicating each item is working, as intended to measure a particular range of volitional motor performance without significant item disorder.

analysis of how clinically relevant subgroups within the above mentioned classes perform for each item. An important premise of Rasch analysis is that participants located at the same position on the SCAR should have the same expected value for any single item at that location, regardless of whatever clinically relevant group to which they might belong (for example AIS A or AIS D). Differential item functioning (DIF) represents the extent to which observed data violates this ideal premise across these clinically relevant groups. For the SCAR, DIF was evaluated by gender, age (<70 versus ≥70), AIS (ASIA Impairment Scale) classification (AIS A–E) and injury level (cervical, thoracic, lumbosacral). No statistically significant DIF was observed in any subgroup identified after adjustment for multiple comparisons.

Local independence of items

In well-functioning scales the error in response to one item in a scale should be independent of the error in response to any other item. Local dependence, analyzed by looking for correlations among item residuals (observed–expected item values), should be low. Only 1 item

pair out of 276 possible combinations ('mobility indoors' paired with 'mobility over moderate distances') had residuals that were modestly correlated (>0.6) indicating a probable overlap between the ability required to perform these two tasks.

DISCUSSION

SCAR focuses on a single underlying measurement construct (volitional performance) by combining select items from two common clinical assessment tools (ISNCSCI and SCIM) where the item response options (scoring) were adjusted to track meaningful transitions in muscle function and ADLs. The results from the model development revealed good measurement properties and generated a validated linear interval-level measure with repeatable precision across a broad range of SCI levels and SCI severities with the exception of those individuals suspected of having a CCS. On the basis of the stringent Rasch criteria that the SCAR satisfies, inclusive SCI clinical trials will benefit from applying this methodology as an outcome measurement tool that will include and track changes in volitional performance including ADLs across the majority of SCI study participants. Fortunately, clinical investigators do not have to adjust their scoring of ISNCSCI or SCIM items, any adjustment can be easily made as part of the statistical analysis according to the criteria validated by Rasch modeling.

The international SCI community established nearly a decade ago that no CNS-directed therapy should be considered effective unless it improves the ability of people to independently complete common ADLs.^{11,12,26} SCI causes deficits in multiple domains of functioning²⁶ that occur to different extent and degree in the overall patient population, as well as recover to different degrees and at different rates. In general, to accurately assess beneficial treatment interventions in experimental participants when compared to appropriate controls requires a thoughtful identification and rationalization of the specific health measurement construct to be utilized, along with its range of impact and rate of spontaneous recovery.²⁷

The Spinal Cord Outcomes Partnership Endeavor (SCOPE) recommends that scales used in SCI should, apart from having adequate documentation of scale validity and reliability, be capable of providing accurate and quantifiable information regarding therapeutic benefit.²⁶ Accordingly, a higher score should always indicate a more favorable outcome with interpretable differences across scores. However, with current ordinal assessment scoring scales (for example, ISNCSCI and SCIM), the same sum score can be obtained with a variety of item score combinations. Thus, because of the inherent variability within these ordinal scales, two patients with similar scores can have markedly different functional abilities.^{3,8} Therefore, if scale sum scores (or sub-scores) do not consistently measure meaningful transitions of scale items, they cannot be expected to accurately measure a treatment benefit.

The large number of records from the entire EMSCI data set calibrates the scale locations for the SCAR items, augmenting the precision of item location estimates and enabling the precise scoring and tracking of an individual in future acute and sub-acute clinical trials where the SCAR can be utilized. The SCAR has been shown to be precise in measuring a participant's volitional performance at time points up to 1 year post injury. More importantly, it accomplishes this with fewer response options (scores) for most items than used in the original assessments. This may seem counter-intuitive, but for ordinal scales, more scoring choices can just as likely contribute to rating indecision and thus reduce the accuracy of the assessed outcome and the reliability of the item score.^{15,16,19,22,25}

It is important to note that even though a broad range of tetraplegic and paraplegic SCI were included over an assessment timeframe of 1

Table 1 Sample size calculations for hypothetical acute or sub-acute SCI clinical trials with SCAR versus SCIM scale as clinical endpoint

Study Duration (start – stop)	Cervical SCI Study Criteria (Inclusion / Exclusion for both SCAR and SCIM)			
	AIS Grade	NLI	Age	Baseline SCAR
< 2 weeks – 6 months	A, B, C	C4 – C7	18 - 70	10 - 70
1 month – 6 months	A, B, C	C4 – C7	18 - 70	10 - 70

A

SCAR Study Duration	Statistics: Change from Baseline to Study Endpoint in Historical Controls					
	N	Mean	SD	SEM	MDD	SDD
< 2 weeks – 6 months	165	18.10	12.74	2.22	6.15	3.93
1 month – 6 months	294	13.46	9.12	1.59	4.40	2.11

Difference between Treatment and Historical Controls for Hypothetical SCAR Score Endpoints			
Disparity in SCAR Score Endpoint	10 points	8 points	5 points
Historical Controls (baseline to study endpoint)	20	20	20
Treatment Group (baseline to study endpoint)	30	28	25
Common SD (standard deviation)	15	15	15
Study Power	80%	80%	80%
Number of participants per study arm	37	57	143

B

SCIM Study Duration	Statistics: Change from Baseline to Study Endpoint in Historical Controls					
	N	Mean	SD	SEM	MDD	SDD
< 2 weeks – 6 months	46	37.33	26.03	4.53	12.56	15.20
1 month – 6 months	110	26.48	17.47	3.04	8.43	6.60

Difference between Treatment and Historical Controls for Hypothetical SCIM Score Endpoints			
Disparity in SCIM Score Endpoint	10 points	8 points	5 points
Historical Controls (baseline to study endpoint)	35	35	35
Treatment Group (baseline to study endpoint)	45	43	40
Common SD (standard deviation)	30	30	30
Study Power	80%	80%	80%
Number of participants per study arm	143	222	567

Abbreviations: N, participant sample size; mean, average score; MDD, minimal detectable difference; s.d., standard deviation; s.e.m., standard error of the mean; SDD, smallest detectable difference. Table 1 highlights potential differences in the necessary sample size for an inclusive acute or sub-acute cervical SCI clinical trial when using SCAR (spinal cord ability ruler) or SCIM (spinal cord independence measure) III as a clinical outcome tool. In both examples, the same study participants have been selected (upper panel). Hypothetical participants are between the ages of 18–70 and have either a complete or incomplete C4–C7 cervical SCI. At baseline (acute ≤ 2 weeks after SCI; sub-acute = 1 month after SCI), each participant must have a minimum SCAR score of 10/100 and no more than a maximum of 70/100. The hypothetical trial end time is 6 months after SCI. As can be seen in the subsequent tables (A, B), the mean change and standard deviation over the study duration is greater (~double) when using the SCIM as a functional outcome measurement tool (panels B). This contributes to an almost fourfold decrease in the required number of participants per study arm for a statistically valid trial when using the SCAR tool as an outcome measurement (panels A).

year post injury, floor/ceiling effects were low and the vast majority of assessments were spread out broadly, but predominantly within the most sensitive (middle) range of the SCAR (Figure 4). This gives confidence that the SCAR will precisely score a broad range of SCI participants and track any changes that may occur during the first year after SCI for all participants, except those with CCS. It is possible that SCAR could be used to track study participants over a longer time period, but EMSCI data does not provide assessment data at longer time intervals.

The Rasch-transformed SCAR scores generate numerical scores on an interval scale, which permit analysis of continuous data using more powerful parametric statistics. Equally important is the original scores, normally used in the assessment of ISNCSCI and SCIM, that do not have to be altered during assessments. ISNCSCI and SCIM II/III assessments can continue to be administered in the normal manner, and the Rasch-validated rules will allow a skilled statistician to easily convert those original scores into ordered SCAR scoring options. The SCAR scores can then be utilized as the primary clinical endpoint,

providing a more precise methodology to evaluate therapeutic efficacy based on improvements in volitional performance for each trial participant relative to an appropriate control group.

Limitations and future studies

Although use of the Rasch-transformed SCAR scores provides a more efficient and precise measure of treatment interventions, some questions/limitations remain to be addressed in future studies.^{13,14} Perhaps the most obvious limitation is the presence of CCS.^{17,18} As we can broadly identify these individuals at later time points after SCI, we can now determine whether there is a modification to the scale that would allow us to predict their natural history of recovery and accurately locate and track such individuals along the SCAR.

Use of the SCAR does not obviate the need for treatment group stratification based upon baseline prognostic factors likely to impact subsequent outcomes (for example, derived from unbiased recursive partitioning^{13,14}). Relevant stratification techniques should also be used to ensure that the treatment arms in a study population are well

balanced, and SCAR with SCI natural history data (for example, EMSCI) can assist in this regard. Finally, the natural history of recovery for identifiable SCI sub-populations (for example, cervical AIS C participants, thoracolumbar AIS B and so on) should be assessed for the changes in SCAR scores across various recovery times after SCI. Such information is useful when determining a minimal detectable difference or estimating expected trial endpoints for different types of SCI.²⁸

Like with any other scoring instrument, the definition of a minimal clinically important difference (MCID) requires intensive clinical and clinical participant judgment, as it cannot be determined by statistics alone.²⁸ However, hypothetically assuming a difference between treated and control groups of either 5, 8, or 10 points (out of 100) to demonstrate a treatment effect on the SCAR results in a fourfold decrease in clinical sample size requirements as shown in Table 1. These sample size calculations provide an adequately powered study, while maintaining operationally feasible pragmatic enrollment requirements. Finally, although we have repeatedly validated multiple randomly selected EMSCI data sets with the Rasch model, it is important to confirm our findings using other equivalent databases and SCAR in an ongoing SCI clinical trial.

CONCLUSIONS

Use of the Rasch-transformed SCAR in future clinical trials will allow the simultaneous inclusion of tetraplegic and paraplegic, complete and incomplete SCI participants (for example, AIS A–C) while keeping sample sizes at operationally feasible levels. It is important to note that clinical investigators do not have to change their ISNCSCI and SCIM assessment procedures or conventional scoring of individual items on these scales to use the SCAR in clinical trials. When prospectively defined in the protocol and detailed in the statistical analysis plan, Rasch-transformed results from the SCAR can be used to generate legitimate total scores for each participant and subsequently analyze using conventional parametric statistical analyses. Our approach of clearly defining the target health construct of volitional performance, conceptualizing and mapping out the key clinical features to be measured, including meaningful scoring options for each item, and rigorously testing these assumptions using Rasch analysis has shown that it is possible to improve SCI clinical assessment tools. The same concept of measuring the performance of volitional tasks along with assessment of functional muscle contractions may be useful in improving scale performance in other neurological conditions and may represent an important advance in the measurement of motor function in neurological interventional trials.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are grateful to Asubio Pharma for their support, which aided some of the statistical analysis. We also appreciate the efforts of René Koller, EMSCI database manager. We acknowledge support provided by EMSCI (funded by IRP Switzerland) and SCOPE for facilitating meetings and guiding discussions to formulate the SCAR concept.

- 1 Spinal Cord Injury (SCI) Facts and Figures at a Glance. *J Spinal Cord Med* 2015; **38**: 249–250.
- 2 Catz A, Itzkovich M, Tesio L, Biering-Sorensen F, Weeks C, Laramie MT *et al*. A multicenter international study on the Spinal Cord Independence Measure, version III: Rasch psychometric validation. *Spinal Cord* 2007; **45**: 275–291.
- 3 Bluvstein V, Front L, Itzkovich M, Benjamini Y, Galili T, Gelernter I *et al*. A new grading for easy and concise description of functional status after spinal cord lesions. *Spinal Cord* 2012; **50**: 42–50.
- 4 Fromovich-Amir Y, Biering-Sorensen F, Baskov V, Juocevicius A, Hansen HV, Gelernter I *et al*. Properties and outcomes of spinal rehabilitation units in four countries. *Spinal Cord* 2009; **47**: 597–603.
- 5 Kirshblum SC, Waring W, Biering-Sorensen F, Burns SP, Johansen M, Schmidt-Read M *et al*. Reference for the 2011 revision of the International Standards for Neurological Classification of Spinal Cord Injury. *J Spinal Cord Med* 2011; **34**: 547–554.
- 6 Medical Research Council. *Aids to the Examination of the Peripheral Nervous System, Memorandum No. 45*. Her Majesty's Stationery Office: London, UK, 1981.
- 7 Nesathurai S. Steroids and spinal cord injury: revisiting the NASCIS 2 and NASCIS 3 trials. *J Trauma* 1998; **45**: 1088–1093.
- 8 Marino RJ, Graves DE. Metric properties of the ASIA motor score: subscales improve correlation with functional activities. *Arch Phys Med Rehabil* 2004; **85**: 1804–1810.
- 9 Anderson KD, Acuff ME, Arp BG, Backus D, Chun S, Fisher K *et al*. United States (US) multi-center study to assess the validity and reliability of the Spinal Cord Independence Measure (SCIM III). *Spinal Cord* 2011; **49**: 880–805.
- 10 Itzkovich M, Tripolski M, Zeilig G, Ring H, Rosentul N, Ronen J *et al*. Rasch analysis of the Catz-Itzkovich spinal cord independence measure. *Spinal Cord* 2002; **40**: 396–407.
- 11 Fawcett JW, Curt A, Steeves JD, Coleman WP, Tuszynski MH, Lammertse D *et al*. Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: spontaneous recovery after spinal cord injury and statistical power needed for therapeutic clinical trials. *Spinal Cord* 2007; **45**: 190–205.
- 12 Steeves JD, Lammertse D, Curt A, Fawcett JW, Tuszynski MH, Ditunno JF *et al*. Guidelines for the conduct of clinical trials for spinal cord injury (SCI) as developed by the ICCP panel: clinical trial outcome measures. *Spinal Cord* 2007; **45**: 206–221.
- 13 Tanadini LG, Steeves JD, Hothorn T, Abel R, Maier D, Schubert M *et al*. Identifying homogeneous subgroups in neurological disorders: unbiased recursive partitioning in cervical complete spinal cord injury. *Neurorehabil Neural Repair* 2014a; **28**: 507–515.
- 14 Tanadini L, Hothorn T, Jones L, Lammertse DP, Abel R, Maier D *et al*. Toward inclusive trial protocols in heterogeneous neurological disorders: prediction-based stratification of participants with incomplete cervical spinal cord injury. *Neurorehabil Neural Repair* 2014b; **29**: 867–877.
- 15 Hobart JC, Cano SJ, Thompson AJ. Effect size statistics can be misleading: is it time to change the way we measure change? *J Neurol Neurosurg Psychiatry* 2010; **81**: 1044–1048.
- 16 Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet* 2007; **6**: 1094–1105.
- 17 Pouw MH, Van Middendorp JJ, Van Kampen A, Curt A, Van De Meent H, Hosman AJ. Diagnostic criteria of traumatic central cord syndrome. Part 3: descriptive analyses of neurological and functional outcomes in a prospective cohort of traumatic motor incomplete tetraplegics. *Spinal Cord* 2011; **49**: 614–622.
- 18 Curt A, Ellaway PH. Clinical neurophysiology in the prognosis and monitoring of traumatic spinal cord injury. *Handb Clin Neurol* 2012; **109**: 63–75.
- 19 Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 2009; **13**: iii, ix-x, 1–177.
- 20 Belvedere SL, De Morton NA. Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *J Clin Epidemiol* 2010; **63**: 1287–1297.
- 21 Andrich D, Sheridan B, Luo G. Perth: RUMM Laboratory; 2030. 2008. Rasch Unidimensional Measurement Model. Computer software program.
- 22 Lindeboom R, Vermeulen M, Holman R, De Haan RJ. Activities of daily living instruments: Optimizing scales for neurologic assessments. *Neurology* 2003; **60**: 738–742.
- 23 Guyenet PG. Regulation of breathing and autonomic outflows by chemoreceptors. *Compr Physiol* 2014; **4**: 1511–1562.
- 24 Drake MJ. Management and rehabilitation of neurologic patients with lower urinary tract dysfunction. *Handb Clin Neurol* 2015; **130**: 451–468.
- 25 Weng LJ. Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educ Psych Measure* 2004; **64**: 956–972.
- 26 Alexander MS, Anderson K, Biering-Sorensen F, Blight AR, Brannon R, Bryce TN *et al*. Outcome measures in spinal cord injury: recent assessments and recommendations for future directions. *Spinal Cord* 2009; **47**: 582–591.
- 27 Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E *et al*. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1—eliciting concepts for a new PRO instrument. *Value Health* 2011; **14**: 967–977.
- 28 Wu X, Liu J, Tanadini LG, Lammertse DP, Blight AR, Kramer JL *et al*. Challenges for defining minimal clinically important difference (MCID) after spinal cord injury. *Spinal Cord* 2015; **53**: 84–91.