RADAR

www.brookes.ac.uk/go/radar

OXFORD BROOKES UNIVERSITY

Directorate of **Learning Resources**

A STRUCTURED REVIEW OF OUTCOME MEASURES USED FOR THE ASSESSMENT OF REHABILITATION INTERVENTIONS FOR SPINAL CORD INJURY.

Dawson, J.[1] Shamley, D.[2] Jamous, M.A.[3]

From: the School of Health and Social Care, Oxford Brookes University; Department of Public Health, University of Oxford and the National Spinal Injuries Unit, Stoke Mandeville Hospital, UK.,

[1]Jill Dawson DPhil, Senior Research Scientist, Department of Public Health, University of Oxford, Old Road Campus, Headington, Oxford OX37LF & Visiting Professor, Oxford Brookes University, Oxford.

[2]Delva Shamley PhD, Senior Lecturer in Physiotherapy, School of Health and Social Care, Oxford Brookes University, Marston Road Campus, Jack Straws lane, Oxford OX3 0FL.

[3]Ali Jamous MD, Associate Specialist in Spinal Injuries, National Spinal Injuries Centre, Stoke Mandeville Hospital, Buckinghamshire Hospitals NHS Trust, HP21 8AL.

Corresponding author:
Dr Jill Dawson,
Senior Research Scientist,
Department of Public Health,
University of Oxford,
Old Road Campus,
Oxford OX37LF
Telephone: 01865 289423
FAX: 01865 289436

A STRUCTURED REVIEW OF OUTCOME MEASURES USED FOR THE ASSESSMENT OF REHABILITATION INTERVENTIONS FOR SPINAL CORD INJURY.

ABSTRACT (216 words)

Objectives. To review the measurement properties of outcome measures of function or mobility currently used in the context of spinal cord injury (SCI).

Methods. A keyword search of multiple databases to identify original papers. Papers were reviewed where they had included an evaluation, with people with SCI, of the psychometric properties of an outcome measure, which included functional or mobility domains.

Information was extracted concerning item generation, ease and intended method of use, and scale properties, in particular: reliability, validity and responsiveness. Conclusions were reached concerning the psychometric properties of each instrument.

Results. Eight outcome measures were identified (plus adapted versions). Five had originally been developed specifically for evaluating patients with SCI (chiefly reflecting clinicians' perspectives), the remaining 3 had not.

The psychometric properties of the instruments varied, with some mixed/contradictory evidence likely relating to differing study sample sizes, characteristics and variable quality. Instruments also varied in stated purpose or emphasis.

Conclusion. In addition to weighing evidence concerning measurement properties, users need to consider the stated purpose and item content of instruments in relation to their specific aims. With regard to the former, while reviewed instruments had some flaws, the Spinal cord Independence Measure (revised version III), Quadriplegia Index of Function (Short-form), Needs Assessment Checklist and SIP68 appeared the best, despite limited evidence of their responsiveness.

A STRUCTURED REVIEW OF OUTCOME MEASURES USED FOR THE ASSESSMENT OF REHABILITATION INTERVENTIONS FOR SPINAL CORD INJURY.

INTRODUCTION

The annual incidence of spinal cord injury (SCI) is estimated to lie between 10 and 83 per million inhabitants per year and often results in catastrophic dysfunction and disability.[1] Progress is gradually being made in the treatment of SCI to limit damage, prevent or treat complications prolonging survival, improve function, and enhance recovery. Any new interventions will likely include pharmacological, surgical and rehabilitation approaches, and all will require evaluations of their efficacy using appropriate outcome measures.[2]

SCI is initially diagnosed in terms of the level at which the injury has occurred, which tends to equate with the observed degree of neurological and functional deficit. Nearly half of all spinal cord injuries are functionally incomplete, with some function preserved below the level of the lesion (although there is much variation between groups[3]). In such cases, the majority of people will likely experience useful recovery (to ASIA grade C or D[4]), including the ability to walk.[5] Rehabilitation interventions and outcomes of SCI have thus tended to particularly focus on functional status.

Since the mid-twentieth century, health status questionnaires and rating forms have been used to assess patients in a range of clinical settings, in order to document outcomes of care. These have usually been completed by health-care staff and primarily represented their perspectives. However, during the past two decades, health care has become more patient-centred, with measures emerging which assess the impact of a wide range of health care interventions, from the patient's perspective[6]. Such 'patient-reported outcome measures (PROMs)' have gradually been introduced as an important outcome (or 'end-point') in randomised clinical trials and observational studies[7].

For results to be meaningful, it is imperative that any measures used to assess outcomes in any health care context cover domains (e.g. pain, physical function, perceived independence) that are relatively specific and appropriate to a particular context or study aim. Evidence also needs to have been presented demonstrating that the questionnaire (and any associated scales) has acceptable measurement properties, including: reliability, validity, responsiveness, acceptability and feasibility.[7] Another property that overlaps with reliability, validity and responsiveness, and which is particularly pertinent to measurement scales, is that of precision. In prospective outcome studies, such as a trial, the responsiveness of an outcome measure, that is, its ability accurately to detect change when it has occurred, is a particularly important aspect.[8] These stipulations apply to the SCI context no less than they do for any other condition.

The purpose of this paper is to provide a structured review of instruments, that are widely used for the assessment of function or mobility in the context of SCI where they have also received any form of psychometric evaluation in that context. Evidence of their measurement properties is presented and non-scientific practical considerations are also highlighted to further facilitate clinical decisions.

METHODS

Search Strategy

The following databases were searched for the years 1969-2006: National Library of medicine (Pub Med), Cochrane, CINAHL, and AMED. The search was limited to the English language. The term "spinal cord injury" was combined with the terms "classification or assessment, Index, Scale, outcomes measure or measurement, functional outcomes, mobility and functional assessment". Papers were selected by reviewing their titles and abstracts with additional references identified from the reference lists of selected papers. General search engines were used to access non-

peer reviewed professional and specialist guidelines and workshops on Spinal Cord

Injury websites such as the International Campaign for Cures of Spinal Cord Injury

Paralysis, Spinal Cord Medicine, American Spinal Injury Association, the National

Institution of Neurological Disorder and Stroke(NINDS Spinal Cord Injury), American and

Canadian Spinal Research Organisation, International Spinal Injuries & Rehabilitation

Centre (UK).

Inclusion Criteria

Reports of any studies evaluating the use of an outcome measure to assess

function or mobility in spinal cord injured patients were initially identified. Abstracts of all

papers and titles were independently assessed by two reviewers (HA and DS) and

agreement confirmed by a third (JD). Full copies of the selected papers were then

obtained. Details of a measure were only included in the final review where some

evidence of its psychometric properties had been published, which had been evaluated

in the context of SCI.

Data extraction

Using criteria for evaluating outcome measures described by Fitzpatrick et al

(1998)[7], data were independently extracted by three reviewers [DS, HA,JD]. Evaluation

of measures gave particular consideration to the following criteria:

Reliability - concerned with reproducibility and internal consistency. It assesses the

extent to which an instrument is free from random error or the amount of a score that

represents signal rather than noise. Test-retest reliability is designed to take account of

variation over time in stable patients. The results of tests of internal consistency (eg.

Cronbach's alpha (α)) and test-retest reliability (eg. intra-class correlation, Bland-Altman

methods) are presented. Reliability estimates of $\alpha \geq 0.7$ are needed to claim internal consistency and are recommended for instruments intended for use at the group level[7]. Estimates need to be higher ($\alpha \geq 0.9$) where instruments are to be applied to individuals.[9]

Validity - addresses whether an instrument measures what it is intended to measure.[7] The best evidence for validity involves assessing an instrument against a true value for the measure: a 'gold standard'.[10] In the SCI context, the American Spinal Injury Association (ASIA) Standards for Neurological Classification of Spinal Cord Injury[4] is the most widely used and accepted method to evaluate and classify the level and degree of impairment of patients' SCI.[11] This system represents a 'gold standard' for assessing neurological (motor and sensory) impairment in SCI, but was not designed to assess functional ability or locomotion and does not therefore represent a 'gold standard' for assessing criterion validity of instruments focused on these domains. Thus, any evidence presented for the validity of instruments that involved comparisons with ASIA scores has been cited in this review as evidence of convergent or concurrent, rather than criterion, validity[a].

In the absence of a 'gold standard' for direct comparison, evidence for validity can take many forms. The source of instrument items and evidence for content and face validity may be presented, which can include qualitative examination of instrument content. Quantitative evidence derived from factor analysis or principal components that support dimensionality, or internal construct validity, is commonly presented. External construct validation generally includes comparisons with other instruments which may include standard clinical assessments.[7] Frequently, this involves demonstrating that a measure is closely correlated with different measures of the same trait ('convergent

validity'), or that a measure correlates little with measures intended to indicate a different trait ('discriminant validity').[9]

More recently, Rasch analysis: a more stringent assessment of underlying scale structure and dimensionality, is increasingly being undertaken.[12] Rasch models test how well instruments conform to uni-dimensionality, hierarchy and interval location of items by examining patterns of individuals' performance on the range of items in a scale and patterns of items' difficulty or severity.[7]

Responsiveness/Sensitivity - concerned with the extent to which an instrument is sensitive to meaningful changes in health status. This property is particularly important for instruments applied in clinical trials.[9] Responsiveness preferably needs to be assessed in a prospective study, where change in health status is likely to occur for the majority. Here, effect sizes are commonly employed (other methods include using paired t-test comparisons or the responsiveness statistic[8]), which is a method of calculating the magnitude of change measured by an instrument in a standardised way which allows direct comparisons to be made between different instruments and scales.[13]

RESULTS

Instruments

Table 1 lists details of any instruments identified as being widely used for the assessment of function or mobility in the context of SCI which had received *any* form of psychometric evaluation in that context, together with brief details of published studies contributing to that evidence.

A total of eight instruments (plus modified versions) were identified, namely: the Barthel Index of Disability (BI)[14] (Modified BI[15]); the Functional Independence Measure (FIM)[16] (Adapted Turkish version[17] and shortened version - the Fone FIM[18]); the

---

[a] This is irrespective of the term used in any cited articles.

Quadriplegia Index of Function (QIF)[19] (Short-form QIF[20]); the Spinal Cord Independence Measure (SCIM)[21] (revised SCIM[22]); the Walking Index for Spinal Cord Injury (WISCI)[23]; the Needs Assessment Checklist (NAC)[24]; the Spinal Cord Injury Functional Ambulation Inventory (SCI-FAI)[25] and the Short Sickness Impact Profile (SIP68)[26].

In contrast to other measures cited in table 1, the BI, MBI, FIM (and adapted versions) and the SIP68, were not originally designed to assess patients with SCI specifically, (although the SIP68 developmental study included 5% SCI patients), but were instead designed for application in a range of rehabilitation settings. While the psychometric properties of the BI and MBI have been evaluated within a number of contexts (eg. older people, stroke patients), few details of any such evaluations could be found involving SCI patients, apart from within a study involving a Turkish translation of the MBI.[27] This latter version had items altered to suit different cultural norms, which also risked altering their meaning from earlier English formats (although psychometric reassessment was formally conducted).

By contrast with the BI and MBI, the original FIM has been assessed in a number of studies with SCI patients[28] [29-31], as have the more recent adapted versions[17] [18;32;33] of the FIM. The SIP68 has also been evaluated with SCI patients in two studies with relatively large sample sizes[34;35].

Instruments that were designed specifically to assess the function or mobility of SCI patients are: the Quadriplegia Index of Function (QIF) [19;36] and Short-form QIF[20]; the Spinal Cord Independence Measure (SCIM)[21] and Catz-Itzkovich revised SCIM [22]; the Walking Index for Spinal cord injury (WISCI)[23]; the Needs Assessment Checklist (NAC)[24], and the Spinal Cord Injury Functional Ambulation Inventory (SCI-FAI)[25].

Item generation

While the precise formal method of item generation was rarely specified (exceptionally, originators of the WISCI specify using a modified Delphi technique[23]), the majority of instruments were devised by health care providers. Only the SIP68 appeared to have involved patients and their carers[37] in the initial process. Two instruments (WISCI[23], SCI-FAI[25]) used blinded ratings of videotaped footage of patients to aid consensus within a research team, after a list of candidate items had been produced. Computerised techniques (regression or principal components analysis) were employed to select/reduce items to produce the Short-form QIF[20] and the SIP68[37]. The former also involved interviews with patients at this stage.

As well as being chiefly designed by health care providers, the majority of instruments were designed to be used by clinical raters. While this generally did not preclude gaining input from patients and carers, many of the instruments also had quite complex scoring systems requiring raters to undergo training in their use and interpretation. The minority of instruments that could be self-rated were: the FIM[16] (particularly, the adapted version[38]) and shortened Fone-FIM[18], the NAC[24] and the SIP68[37].

Measurement properties of instruments within the SCI context (summarised in table 2)

Table 2 provides details of the measurement properties of instruments tested in the context of SCI, reported by studies cited in table 1.

The BI, FIM and SCIM appeared to have been used with SCI patients most often, largely reflecting the greater length of time that had elapsed since first publication. This did not necessarily indicate a greater degree of instrument evaluation having occurred in the SCI context.

Ceiling/floor effects

In many cases (BI, QIF, SCIM, SCI-FAI) little formal evidence had been presented concerning overall score floor or ceiling effects. In all other cases, clustering of extreme values (allowing for no further improvement or deterioration to be measurable on subsequent assessment) commonly occurred in some subscales eg. cognitive scale: FIM [30] [17]; communication scale: FIM [30] [17] and SIP68[35]; stairs: MBI[27]; mobility: SCIM-III[3] and SIP68[35]; walking: WISCI[39;40]. An item-level analysis of the BI noted the presence of floor or ceiling effects for some items, including feeding and grooming, at admission or discharge.[41]

Reliability

No evidence of the internal consistency of the instrument or subscales could be for the BI, QIF, WISCI or SCI-FAI. Evidence of adequate internal consistency (Cronbach's α >0.7) was reported for all other measures apart from the FIM locomotion (Cronbach's α 0.4)[42] and SIP68 emotional stability (Cronbach's α 0.68)[34] subscales. Evidence of optimal internal consistency (Cronbach's α range 0.80 to 0.90) was reported for the Turkish MBI[27], Short-form QIF[20], NAC (most subscales)[24;43] and SIP68 (all but one subscale)[34].

Test retest reliability (TRT) (assessing the same rater's responses on different occasions) had not been assessed in relation to the QIF, the short-form QIF, The WISCI, or the SCI-FAI. TRT assessment of the NAC, had used correlations alone (correlation is inadequate, as it is not a test of agreement), which produced high values r ≥ 6.9[24]. TRT

reliability (using Kappa or ICC) was good (K values 0.61 - 0.80, or ICC values >0.70) for the MBI[27], SCIM[44], SCI-FAI[25] and SIP68[35].

Evidence for inter-rater (or observer) reliability was not found for the BI, short-form QIF, NAC or SCI-FAI. Assessments of the QIF and FIM (patient versus clinician rating compared in the latter case) used correlations alone, which were moderate to high (r range 0.55 - 0.95)[19;29]. Using Kappa, ICC or Kendall coefficient of concordance[45], the FIM (Turkish version)[17] and SCIM-II produced some moderately low values (<0.5)[44]; the MBI[27] had moderate to good levels of inter-rater reliability (K values > 0.5)[27], with very high values (K values 0.81 - 1.00 or ICC values >0.90) presented for the SCIM-III[22].

Validity

Given the absence of any 'gold standard', correlation comparisons between other measures (purporting to measure a broadly similar construct) were taken to denote evidence, or otherwise, of *convergent, concurrent* or *discriminant validity*. Results of Principal (factor) components, or alternatively Rasch analysis, supporting underlying scale structure were taken as evidence of *construct validity*.

Evidence of acceptable *concurrent validity* was found with the FIM[46]; and for some sub-scales, some of the time, representing the FIM Turkish version[17], MBI (Turkish version)[27], short form QIF[20] and SIP68[34].

Evidence of construct validity could appear inconsistent or contradictory in some instances. Thus, Rasch analysis revealed problems concerning disordered thresholds (grooming and stairs items) and model misfit affecting bladder and bowel items in the FIM motor domain[17;31], (bladder and bowel items in the MBI Turkish version were also associated with considerable levels of misfit in Rasch analysis[27]), that appeared to be accentuated in cross-cultural comparisons[30]; while evidence from factor analysis,

produced an unproblematic two-factor solution (together using all 18 items), with each

scale further associated with a high level of internal consistency (itself a form of

convergent validity)[47].

Findings from Rasch analysis highlighted some flaws relating to *construct validity*

for the SCIM version III, with 'walking outdoors' and 'stair management' items within the

mobility subscale, and 'toilet use' within the respiration/sphincter subscale exhibiting

misfit. The latter subscale also contained some items with disordered category

thresholds eg. bowel management.[3 22] Factor analysis on the SIP68 represented data

from a heterogeneous (mainly non-SCI) population and was therefore considered largely

inappropriate, and no evidence of construct validity (note that absence of evidence is not

synonymous with evidence of absence) was found for the BI, the QIF (or short-form

QIF), the SCIM (early version), the WISCI, the NAC, or the SCI-FAI.

Evidence of *convergent validity* (correlations) was presented for most instruments

in relation to other instruments purporting to measure something similar. Thus the short-

form QIF score correlated with the Upper Extremity Motor Score (r> 0.8)[20]; the SCIM

with the FIM (r> 0.8)[21;48], the WISCI (r=0.97)[49] and with the NAC (r=0.47 to 0.85)[24].

Considerable evidence of convergent validity was presented for the WISCI: with the FIM

(r≥ 0.7)[23;49], the LEMS(r=0.47 to 0.91)[50], the Berg Balance Scale[39], the BI and

Rivermead mobility index (both r=0.67)[49], and the SCIM (r=0.97)[49]. The NAC correlated

with the WISCI and the Hospital Anxiety and Depression Scale (HADS) (r=0.47 to

0.85)[24]; the SCI-FAI with gait scores, walking speed (r≥ 0.7) and the Lower Extremity

Motor Score (LEMS) (r>0.6)[25]; and the SIP68 correlated with relevant domains of the

SF-36 (r≥ 0.57), ADL (somatic autonomy r=0.81)[35], and with the BI (r=0.54 to 0.91)[34].

Responsiveness

Evidence of good responsiveness was found for the BI and FIM motor scale (both had effect size 0.9, comparing scores between patients' admission and discharge dates)[19]. There was other (weaker) evidence suggesting that the FIM was less sensitive than the QIF[46;51]); that the FIM was similar to the BI[51], the QIF was better than the BI[19] and the FIM[46]. The original SCIM and SCIM version III were each found to be superior to the FIM[21;52;53] and the WISCI was possibly superior to the (Locomotor Functional Independence Measure) LFIM and SCIM[39] and had superior sensitivity to walking recovery than the BI, RMI, SCIM, LEMS or FIM[49]. Few studies used effect sizes and overall, evidence of responsiveness was generally quite weak. There was no evidence concerning the responsiveness of the MBI, the short-form QIF, the NAC, or the SIP68 in the SCI context.

DISCUSSION

This review focussed on instruments that are widely used to assess function or mobility in patients with SCI, which have also received some form of psychometric evaluation in that context, and complements and extends the scope of previous reviews in this area[54-57]. Eight instruments were identified, together with adapted or shortened versions. There were two main findings.

The first finding was that, with the exception of the SIP68, none of the measures identified had involved interviews with any patients at all, at the design stage, for the purpose of item generation. This finding naturally leads to the conclusion that current measures *may* not represent SCI patients' perspective, but more likely represent the perspective of clinicians.

Patients increasingly expect to be involved in decisions about their care and to receive accurate information to facilitate their involvement[58]. Thus, the use of

instruments that represent chiefly the clinician's perspective might be considered inappropriate by some, or only appropriate in certain circumstances or in relation to particular domains. Nonetheless, the extent to which patients are involved at all, even in rating the different instruments, remains generally quite limited.

Variation in the extent to which patients are involved in rating questionnaires could be influenced by a number of factors. For instance clinicians may (not unreasonably) believe that patients' and clinicians' ratings of their functioning will differ, but may also assume that clinical observers will provide more objective and accurate scores. Indeed it has been asserted by Itzkovich et al [44] that direct observation of individuals' functioning is more accurate and less subject to bias than patients' self-reports because patients may have unrealistic or uninformed expectations, particularly in relation to goal-setting and achievement. Their score ratings for the SCIM have therefore tended to rely entirely on observations rather than subjective reporting. However, results from a small-scale study, by the same authors, found that any differences between patients' ratings obtained by interview and ratings produced by observers actually appeared insignificant.[44] The extent to which patients are involved in rating questionnaires may also relate to the intended purpose of the instrument. Thus, by contrast with the SCIM, the NAC is more concerned with measuring individuals' rehabilitation success in achieving set goals and patients are invariably involved in rating questionnaire items.[59;60] However, item ratings on the NAC differ in another way from the SCIM (and other measures), in that no distinction is made between someone being able to carry out a task verbally (by asking someone to do it for them) versus carrying out the task themselves. Berry et al defend this, believing that a patient with a higher level of injury should be able to achieve independence, through others, by articulating their own needs. They also argue that, while the patient's perception of their independence might be at the cost of accuracy, their active involvement in the process

engenders compliance. [59;60] Others have also noted that perceived control has the strongest association with life satisfaction.[33] These arguments appear reasonably compelling where outcome measures are used for the purpose of individual goal-setting, but are problematic in other contexts eg. trials comparing outcomes of different interventions.

Clearly while current measures mainly represent clinicians' perspectives, this does not mean that all clinicians share the same perspective (many of these issues have been discussed elsewhere.[54;55]) This nonetheless still leaves the question of whether there is a need for a new self-reported measure to be developed for SCI, that fully represents patients' perspectives.

Our second main finding was that the quality of evidence for the psychometric properties of instruments reviewed was very variable, occasionally quite poor. Evidence for responsiveness, particularly evidence of instruments' ability to detect meaningful change was particularly lacking. Evidence of instruments' psychometric properties also sometimes appeared to be conflicting (eg. different studies' evidence for construct validity). There are a number of likely reasons for this. In the first instance, it is only since the early to mid 1990s that a well-described psychometric methodology has become established for developing and reporting health status questionnaires, that is applicable to clinical situations.[61] A number of the instruments here reviewed, were produced and assessed prior to the mid-1990s.

A possible reason to explain apparently conflicting findings - particularly regarding *construct validity* - is that different statistical procedures, such as factor analysis (representing Classical Test Theory) versus Rasch analysis, have fundamentally different requirements. For instance, instrument scales that arise through application of factor analysis are treated as interval scales, when they are generally based on ordinal

level item scoring; while the Rasch model - to which a scale is compared in Rasch analysis - is a more stringent test, as it is a statistically proven interval scale.[62]

While Rasch analysis may be regarded as more stringent than conventional psychometric analysis, the context of its application remains important, as is the case with classical psychometric methods, Thus another reason why evidence from different studies might differ relates to differing sample sizes and composition relating to age, gender; the range, extent and type of SCI; as well as the context in which studies have occurred - often varies from one study to the next. This is important because the measurement properties of an instrument are not just *of the instrument*: they are of the instrument *pertaining to the population and context in which they are developed, and tested*.[61] Thus, if a measure is designed and calibrated with one group of patients, its measurement properties may change if applied to a different group of patients, such as those representing different age-groups or different clinical characteristics. The use of a measure in a different context from the original developmental study (eg. a hospital pre-post-surgical context versus the context of community based rehabilitation) can also affect the measurement properties.

This leads to the issue of how health and social care providers are to choose between instruments designed for assessing outcomes of interventions for SCI, and whether this review can support particular recommendations.

There will never be a perfect questionnaire or measure of outcome and efforts to produce one risks a proliferation of imperfect examples from which assessors and trialists must then choose, which is to be discouraged. Choosing the right measure involves identifying the most appropriate measure for the chosen patient group, context and purpose, where evidence exists to show that the questionnaire has exhibited adequate measurement properties pertaining to a similar patient group and context.

Of those instruments reviewed, if a generic measure is considered to be appropriate for a particular purpose, then, of the BI/MBI, FIM, and SIP68 generic measures, the SIP68 has the best measurement properties. However, evidence for its responsiveness has not been evaluated within the SCI context, only once this has occurred can its use in clinical trials be sanctioned.

Regarding SCI condition-specific, multidimensional measures (ie. that comprise different dimensions represented by a number of sub-scales), which aim to cover the full range of SCI; of these the SCIM and the NAC had comparable measurement properties. These were mostly good, although a few shortcomings concerning some subscales of the SCIM-III, based on Rasch analysis, may indicate the need for further refinements.[3] The responsiveness of the NAC has also not been assessed. However, as the NAC and the SCIM each reflect somewhat different (although likely related) constructs, and are each applied in different ways, choosing between these instruments depends crucially on the purpose that any potential user has in mind.

Where highly-specific measures are required, for the assessment of mobility/ambulation alone, the WISCI and SCI-FAI are both supported by evidence for acceptable levels of reliability, validity and responsiveness. The only caution concerns their use with low level quadriplegic subjects with whom ceiling effects are likely.

For the assessment of patients with quadriplegia, the QIF has good measurement properties generally. The short form QIF has particularly high levels of internal reliability and could therefore be used to assess progress in individual patients, which is not the case for any of the other measures reviewed - excepting the Turkish versions of the FIM and MBI - no other measures had sufficiently high reliability to permit this application, and they are therefore only suitable for making group comparisons.[63]

While conducting this review, we considered whether SCI represents a particularly challenging area for outcomes measurement. For instance, as others have

noted[56], SCI is a heterogeneous disorder in terms of level and severity of injury, and it is unsurprising that most measures will exhibit floor or ceiling effects when applied to groups of patients that largely represent one or other extreme end of the spectrum of injury. If 'broad spectrum' measures are considered appropriate, then this particular 'flaw' may need to be accommodated. In addition, traumatic SCI may be accompanied by other injuries. These have the potential to produce considerable amounts of 'noise' where the measurement of change in function is specifically concerned with interventions directed towards the SCI. We suggest that there are no simple means of dealing with these substantial challenges.

Recent developments in psychometric theory offer the opportunity to develop item banks which can be retained on computer. Patients can then complete items on-line and, on the basis of their responses to certain items, computer-adaptive testing will select the most appropriate items for them to complete thereafter. This method can reduce patient burden as it leads to fewer items being asked.[64] Whether such methods offer other improvements in SCI assessment remains to be evaluated.

Generally, the use of condition-specific measures with adequate measurement properties is clearly necessary in the context of SCI; but even this is not sufficient if outcome studies are to produce meaningful data. This requires using such measures to be complemented by rigorous planning and conduct of data collection methods, with outcomes data obtained at appropriate points in time, relative to a meaningful date: defined and operationalised in the same manner for all subjects. Study sample sizes also need to be adequate.

The process of conducting this review revealed that, if applied appropriately, while never perfect, a number of outcome measures that are currently available are likely good enough.

Reference List

1. Wyndaele M,.Wyndaele J-J. Incidence, prevalence and epidemiology of spinal cord injury: what learns a worldwide literature survey. *Spinal Cord* 2006;**44**:523-9.

2. Lim PAC,.Tow AM. Recovery and regeneration after spinal cord injury: A review and summary of recent literature. *Ann Acad Med Singapore* 2007;**36**:49-57.

3. Catz A, Itzkovich M, Tesio L, Biering Sorensen F, Weeks C, Laramee MT *et al.* A multi-center international study on the spinal cord independence measure, version III: Rasch psychometric validation. *Spinal Cord* 2007;**45**:275-91.

4. American Spinal Injury Association and International Medical Society of Paraplegia (eds). International standards for neurological classification of spinal cord injury, rev 2002. Chicago, IL: American Spinal Injury Association, 2002.

5. Burns SP, Golding DG, Rolle JWA, Graziani V, Ditunno JF. Recovery of ambulation in motor-incomplete tetraplegia. *Arch Phys Med Rehabil* 1997;**78**:1169-72.

6. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ* 2002;**324**:1417.

7. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;**2**:i-74.

8. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987;**40**:171-8.

9. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res* 1993;**2**:441-9.

10. Fayers PM, Machin D. Quality of Life - assessment, Analysis and Interpretation. Chichester: John Wiley & Sons Ltd, 2000.

11. Consortium for Spinal Cord Medicine. Outcomes following traumatic spinal cord injury: clinical practice guidelines for health-care professionals. Paralyzed Veterans of America., 1999.

12. Andrich D. Rasch Models for Measurement. London: Sage Publications, 1988.

13. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;**27**:S178-S189.

14. Mahoney FI,.Barthel DW. Functional evaluation: The Barthel Index. *Maryland State Medical Journal* 1965;**14**:61-5.

15. Shah S, Vanclay F, Cooper B. Improving the sensitivity of the Barthel Index for Stroke rehabilitation. *J Clin Epidemiol* 1989;**42**:703-9.

16. Hamilton BB, Fuhrer MJ. Rehabilitation Outcomes: Analysis and Measurement. Baltimore: 137-147.: Brooks, 1987.

17. Kucukdeveci AA, Yazuver G, Elhan AH, Sonel B, Tennant A. Adaptation of the Functional Independence Measure for use in Turkey. *Clin Rehabil* 2001;**15**:311-9.

18. Smith P, Hamilton BB, Granger CV. FIM decision Tree: the Fone FIM. Buffalo, NY.: Research Foundation of State University of New York., 1990.

19. Gresham GE, Labi MLC, Dittmar SS, Hicks JT, Joyce SZ, Stehlik MA. The Quadriplegia Index of Function (QIF): sensitivity and reliability demonstrated in a study of thirty quadriplegic patients. *Paraplegia* 1986;**24**:38-44.

20. Marino RJ,.Goin JE. Development of a short-form Quadriplegia Index of Function scale. *Spinal Cord* 1999;**37**:289-96.

21. Catz A, Itzkovich M, Agranov E, Ring H, Tamir A. SCIM - Spinal Cord Independence Measure: A new disability scale for patients with spinal cord lesions. *Spinal Cord* 1997;**35**:850-6.

22. Itzkovich M, Gelernter I, Biering Sorensen F, Weeks C, Laramee MT, Craven BC *et al.* The spinal Cord Independence Measure (SCIM) version III: Reliability and validity in a multi-center international study. *Disabil Rehabil* 2007;**29**:1926-33.

23. Ditunno JFJr, Ditunno PL, Graziani V, Scivoletto G, Bernardi M, Castellano V. Walking index for spinal cord injury (WISCI): an international multicenter validity and reliability study. *Spinal Cord* 2000;**38**:234-43.

24. Kennedy P,.Hamilton LR. The needs assessment checklist: a clinical approach to measuring outcome. *Spinal Cord* 1999;**37**:136-9.

25. Field-Fote EC, Fluet GG, Scott D, Schafer SD, et-al. The spinal cord injury functional ambulation inventory (SCI-FAI). *J Rehabil Med* 2001;**33**:177-81.

26. de Bruin AF, Buys M, de Witte LP, Diedericks JPM. The Sickness Impact Profile: SIP68. a short generic version, first evaluation of the reliability and reproducibility. *J Clin Epidemiol* 1994;**47**:863-71.

27. Kucukdeveci AA, Yazuver G, Tennant A, Sulder N, Sonel B, Arasil T. Adaptation of the modified Barthel Index for use in physical medicine and rehabilitation in Turkey. *Scand J Rehabil Med* 2000;**32**:87-92.

28. Hamilton BB, Laughlin JA, Fiedler RC, Granger CV. Interrater reliability of the 7-level functional independence measure (FIM). *Scand J Rehabil Med* 1994;**26**:115-9.

29. Karamehmetoglu SS, Karacan I, Elbasi N, Demirel G, Koyuncu H, Dosoglu M. The functional independence measure in spinal cord injured patients: comparison of questioning with observational rating. *Spinal Cord* 1997;**35**:22-5.

30. Lawton G, Lundgren-Nilsson A, Biering Sorensen F, Tesio L, slade A, Penta M *et al.* Cross-cultural validity of FIM in spinal cord injury. *Spinal Cord* 2006;**44**:746-52.

31. Lundgren-Nilsson A, Tennant A, Grimby G, Sunnerhagen KS. Cross-diagnostic validity in a generic instrument: an example from the Functional Independence Measure in Scandinavia. *Health Qual Life Outcomes* 2006;**23**:55.

32. Grey N,.Kennedy P. The Functional Independence Measure: a comparative study of clinician and self-ratings. *Paraplegia* 1993;**31**:457-61.

33. Fuhrer MJ, Rintala DH, Hart KA, Clearman R, Young ME. Relationship to life satisfaction to impairment, disability, and handicap among persons with spinal cord injury living in the community. *Arch Phys Med Rehabil* 1992;**73**:552-7.

34. Post WM, de Bruin A, de Witte L, Schrijvers A. The SIP68: A measure of health-related functional status in rehabilitation medicine. *Arch Phys Med Rehabil* 1996;**77**:440-5.

35. Nanda U, McLendon PM, Andresen EM, Armbrecht E. the SIP68: An abbreviated sickness impact profile for disability outcomes research. *Qual Life Res* 2003;**12**:583-95.

36. Gresham, G. E., Labi, M. L., Dittmar, S. S., and et al. Quadriplegia index of function. Arch Phys Med Rehabil 61, 493. 1980.

37. de Bruin AF, Diedericks JPM, de Witte LP, Stevens FCJ, Philipsen H. The development of a short generic version of the Sickness Impact Profile. *J Clin Epidemiol* 1994;**47**:407-18.

38. Heinmann AW, Hamilton BB, Granger CV, Wright BD, Linacre JM, Betts HB *et al.* Rating scale analysis of functional assessment measures. Final Report. Chicago: Rehabilitation Institute of Chicago., 1991.

39. Ditunno JF, Barbeau H, Dobkin BH, Elashoff R, Harkema S, Marino RJ *et al.* Validity of the Walking Scale for Spinal Cord Injury and other domains of function in a multicenter clinical trial. [epub ahead of print]. *Neurorehabil Neural Repair* 2007.

40. Kim MO, Burns AS, Ditunno JF, Marino RJ. The assessment of walking capacity using the walking index for spinal cord injury: self-selected versus maximal levels. *Arch Phys Med Rehabil* 2007;**88**:762-7.

41. O'Connor RJ, Cano SJ, Thompson AJ, Hobart JC. Exploring rating scale responsiveness. Does the total score reflect the sum of its parts? *Neurology* 2004;**62**:1842-4.

42. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the Functional Independence Measurement and its performance among rehabilitation inpatients. *Arch Phys Med Rehabil* 1993;**74**:531-6.

43. Kennedy P, Evans MJ, Berry C, Mullin J. Comparative analysis of goal achievement during rehabilitation for older and younger adults with spinal cord injury. *Spinal Cord* 2003;**41**:44-52.

44. Itzkovich M, Tamir A, Philo O, et al. Reliability of the Catz-itzkovich spinal cord independence measure assessment by interview and comparison with observation. *Am J Phys Med Rehabil* 2003;**82**:267-72.

45. Bland JM,.Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**1**:307-10.

46. Yavuz N, Tezyurek M, Akyuz M. A comparison of two functional tests in quadriplegia: the quadriplegia index of function and the functional independence measure. *Spinal Cord* 1998;**36**:832-7.

47. Stineman MG, Shea JA, Jette A, Tassoni CJ, Ottenbacher KJ, Fiedler R *et al.* The Functional Independence Measure: Tests of Scaling Assumptions, Structure, and Reliability Across 20 Diverse Impairment Categories. *Arch Phys Med Rehabil* 1996;**77**:1101-8.

48. Catz A, Itzkovich M, Steinberg F, et al. The Catz-Itzkovich SCIM: A revised version of the Spinal Cord Independence Measure. *Disabil Rehabil* 2001;**23**:263-8.

49. Morganti B, Scivoletto G, Ditunno P, et-al. Walking Index for Spinal Cord Injury (WISCI): Criterion validation. *Spinal Cord* 2005;**43**:27-33.

50. Ditunno JF, Scivoletto G, Patrick M, Biering Sorensen F, Abel R, Marino R. Validation of the walking index for spinal cord injury in a US and European clinical population.[epub ahead of print]. *Spinal Cord* 2007.

51. Cano SJ, O'Connor RJ, Thompson AJ, Hobart JC. Exploring disability rating scale responsiveness II: Do more response options help? *Neurology* 2006;**67**:2056-9.

52. Catz A, Itzkovich M, Agranov E, Ring H, Tamir A. The spinal cord independence measure (SCIM): Sensitivity of functional changes in subgroups of spinal cord lesion patients. *Spinal Cord* 2001;**39**:97-100.

53. Catz A,.Itzkovich M. Spinal Cord Independence Measure:Comprehensive ability rating scale for the spinal cord lesion patient. *J Rehabil Res Dev* 2007;**44**:65-8.

54. Cohen ME,.Marino RJ. The tools of disability outcomes research functional status measures. *Arch Phys Med Rehabil* 2000;**81**:S21-S29.

55. Lam T, Noonan VK, Eng JJ, and SCIRE Research Team. A systematic review of functional ambulation outcome measures in spinal cord injury. *Spinal Cord* 2007; **Advance online publication**:1-9.

56. Steeves JD, Lammertse D, Curt A, Fawcett JW, Tuszynski MH, Ditunno JF *et al*. Guidelines for the conduct of clinical trials for spinal crd injury (SCI) as develop by the ICCP panel: clinical trial outcome measures. *Spinal Cord* 2007;**45**:206-21.

57. Anderson, K., Aito, S., Atkins, M., and et al. Functional recovery measures for spinal cord injury: comparison by a multi-national work group. Available on-line 18th Nov 2006 under 'ASIA Proceedings 2006 Pre-course' at: http://www.asia-spinalinjury.org/bulletinBoard/proceedings.php. 2006.

58. Siegrist J,.Junge A. Conceptual and methodological problems in research on quality of life in clinical medicine. *Soc Sci Med* 1989;**29**:463-8.

59. Berry C,.Kennedy P. A psychometric analysis of the Need Assessment Checklist (NAC). *Spinal Cord* 2003;**41**:490-501.

60. Berry C,.Kennedy P. Internal consistency and responsiveness of the Skin Management Needs Assessment Checklist post spinal cord injury. *J Spinal Cord Med* 2002;**27**:63-71.

61. Streiner DL, Norman GR. Health measurement scales - a practical guide to their development and use. New York: Oxford University Press, 1989.

62. Prieto L, Alonso J, Lamarca R. classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes* 2003;**1**:1-27.

63. McHorney CA,.Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res* 2005;**4**:293-307.

64. Haley SM, Gandek B, Siebens H, Black-Schaffer RM, Sinclair SJ, Tao W *et al*. Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation: II. Participation outcomes. *Arch Phys Med Rehabil* 2008;**89**:275-83.

65. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. *Int.Disabil.Stud.* 1988;**10**:61-3.

66. Ditunno PL,.Ditunno JF. Walking index for spinal cord injury (WISCI II): Scale revision. *Spinal Cord* 2001;**39**:654-6.

67. van Hedel HJ, Wirz M, Dietz V. Assessing walking ability in subjects with spinal cord injury: validity and reliability of 3 walking tests. *Arch Phys Med Rehabil* 2005;**86**:190-6.

68. Segal ME, Ditunno JFJr, Staas WE. Interinstitutional agreement of individual functional independence measure (FIM) items measured at two sites on one example of SCI patients. *Paraplegia* 1993;**31**:622-31.

69. Hosek, S., Kane, R., Carney, M., and et-al. Charges and Outcomes for Rehabilitative Care: Implications for the Prospective Payment System. RAND-3424-HCFA. 1986. Santa Monica, California, Rand Corporation.

70. Itzkovich M, Tripolski M, Zeilig G, et al. Rasch analysis of the Catz-Itzkovich Spinal Cord Independence measure. *Spinal Cord* 2002;**40**:396-407.

71. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981;**19**:787-805.

Table 1. Health status instruments used for the assessment of outcomes of care for spinal cord injury, that have been evaluated, in that context, for any of their psychometric properties.

Instrument/ Developmental study	Context in which instrument first designed and used	Domains, number of items (n), scoring system	Method of item generation	Intended method of use	Studies involving instrument evaluation with SCI patients	Setting(s) /study population(s) (n, sex, mean age) relating to studies involving instrument evaluation with SCI patients
Barthel Index of disability(BI) Mahoney F, Barthel D. 1965. [14]	Chronic disease hospitals - patients with a neuromuscular or musculoskeletal disorder i.e. not SCI-specific.	1 functional disability domain. 10 (individual ADLs - mobility, basic self care) 2 or 3 response options. Total score 0 to 100 points (0= highest dependency).	Devised by health care providers. Method unclear.	Administered (assessed) by nurse, physio or doctor providing care - (+/- input from patient, relatives etc)	Cano, 2006 [51] O'Connor et al 2004[41]	237/1396 patients had 'spinal syndromes'. Analyses combined diagnostic groups (results similar for all) 57% male, mean age 52 (SD 16) 237 (17%)/1418 patients at a neurorehabilitation unit had SCI, of whom 57% male, mean age 52 (range 16-85).
Modified Barthel Index (MBI-II)* Shah S. 1989[15] **MBI version I[65] not evaluated with SCI.*	Patients with stroke referred for in-patient rehabilitation i.e. not SCI-specific.	1 functional disability domain. 10 (individual ADLs -mobility, domestic & self care). 5 response options. Total score 0 to 100 points (0= highest dependency).	Same items. Response options expanded by occupational therapists.	Administered (rated) by nurse, physio or Dr providing care. - (+/- input from pt, relatives etc)	Kucukdeveci A. et al 2000[27] minor modifications to wording during translation process (into Turkish). Measurement properties re-evaluated.	Hospital in-patients: 50 with stroke and 50 with SCI. Stroke patients: 74% female, mean age 58. SCI patients: 56% female, mean age 32. Level of injury for SCI patients: cervical 22%, thoracic 46%, lumber 32%.
Functional Independence Measure (FIM) Hamilton BB et al.1987.[16]	Designed for disabled people in general i.e. not SCI-specific.	6 areas of functioning, 18 items. Initially 4, then 7 level scale score range 1 to 7, 1= total assistance (poorest score) score. Max score 126.	Task Force of clinical professionals - mainly associated with rehabilitation.	Assessment by clinical raters (especially when involving in-patients) or by self-report	Karamehmetoglu et al 1997[29] Lawton et al 2006 [30] Lundgren-Nilsson et al 2006[31] Yavus et al 1998[46]	50 in-patients with SCI: 22% tetraplegic, 78% paraplegic. 76% male, mean age 40. 647 SCI patients at 19 rehabilitation facilities. 69% male, mean age 46. 471 patients at 9 rehabilitation facilities: included 157 with SCI. 70% male, age range 11-90. 29 tetraplegic patients admitted to rehabilitation centre. 67% male, mean age 37 range 14 - 66. 18 complete, 11 incomplete SCI.
FIM Adapted version[38]/	Involves a patient classification system	Subscale scores combined into two motor and cognitive domains	Same items	Ditto	Stineman, 1996[47], Cano, 2006 [51] Heinmann, 1991 [38]	1831 SCI patients discharged from rehabilitation units/hospitals. Mean age43. 237/1396 patients had 'spinal syndromes'. Analyses combined diagnostic groups (results similar for all) 57% male, mean age 52 (SD 16)
Shortened form/ 'the Fone FIM' adapted for self-report[18]				Self-rated by patient	Grey and Kennedy 1993[32] Fuhrer et al 1992[33] (using shortened form 'antecedant of the Fone FIM'	40 hospital in-patients with traumatic SCI. 32% tetraplegic, 68% paraplegic. 85% male. Mean age 30. 140 community based people with SCI. 41% quadriplegic, 43% paraplegic, 16% incomplete. 71% male, mean age 37.

25

Instrument/ Developmental study	Context in which instrument first designed and used	Domains, number of items (n), scoring system	Method of item generation	Intended method of use	Studies involving instrument evaluation with SCI patients	Setting(s) /study population(s) (n, sex, mean age) relating to studies involving instrument evaluation with SCI patients
Turkish adapted version [17]	Ditto - translated/ adapted for use in Turkey.	2 major domains: 13 motor and 5 social-cognitive function items 7 level scale	Minor word changes /translation by health care professionals and translators.	Assessment by clinical raters	Kucukdeveci et al [17]	50 patients with stroke & 50 with SCI. The latter: 56% female mean age 31.5.
Quadriplegia Index of Function (QIF) Gresham GE et al. 1980 [36]; 1986. [19]	Patients with Quadriplegia (complete lesion)	45 items for 10 areas of self-care & mobility. 7 areas scored on 5 point scale. Final score produced from 37 items with weighting system.	Devised by experienced multidisciplinary SCI team.	Assessment by experienced clinical raters.	Gresham GE et al. 1986. [19] Yavuz et al 1998 [46]	30 quadriplegic patients, of whom 20 rated by independent raters. No other sample details provided. 29 tetraplegic patients admitted to rehabilitation centre. 18 complete, 11 incomplete SCI. 67% male, mean age 37 range 14 - 66.
Short-form QIF Marino & Goin 1999 [20]	Patients with tetraplegia.	6 items (no weighting)	Original items reduced, using interviews with patients and regression analysis.	Assessment by clinical raters	Marino & Goin 1999 [20]	95 tetraplegic patients. 89% male mean age 31.
Spinal Cord Independence Measure (SCIM) Catz A et al. 1997. [21]	Patients with tetraplegia or paraplegia.	4 domains/subscales, comprising 16 items. (scoring: self-care 0-20; respiration & sphincter management 0-40; mobility 0-40).	Devised by specialist health care providers. Method not stated.	Assessment by clinical raters	Catz A et al. 1997. [21]	30 patients: 9 tetraplegic, 21 paraplegic. 73% male, mean age 45.
Catz-Itzkovich revised SCIM Catz A et al. 2001. [52]	Patients with tetraplegia or paraplegia.	3 subscales/18 items: self-care 0-20; respiration & sphincter 0-40; mobility 0-40.	Revised wording to scoring system by specialist health care providers.	Assessment by clinical raters	Catz A et al. 2001. [52]	28 patients: 6 tetraplegic and 22 paraplegic. 64% male, mean age 46.
Catz-Itzkovich revised SCIM (SCIM III) [3]	Patients with tetraplegia or paraplegia.	3 subscales/19 items Scoring - as above.	Refinements to address intercultural biases.	Assessment by clinical raters	Catz A et al 2007 [3] Itzkovich M. et al. 2007. [22]	425 patients representing 6 countries. 188 tetraplegic, 237 paraplegic. 73% male, mean age 47.

Instrument/ Developmental study	Context in which instrument first designed and used	Domains, number of items (n), scoring system	Method of item generation	Intended method of use	Studies involving instrument evaluation with SCI patients	Setting(s) /study population(s) (n, sex, mean age) relating to studies involving instrument evaluation with SCI patients
Walking Index for Spinal cord injury (WISCI) Ditunno JF Jr et al. 2000. [23]/ **Walking Index for Spinal cord injury (WISCI II)**[66]	Patients with spinal cord injury in a rehabilitation setting.	1 hierarchical scale representing 19 levels of walking limitation (includes use of aids and assistance) over 10 metres. One extra level added to scale in WISCI II.	Modified Delphi technique (clinical experts). Consensus reached using blinded ratings of videotaped examples.	Assessment by observers, particularly for use in the context of clinical trials.	Ditunno JF Jr et al. 2000. [23] Morganti B et al. 2005.[49] Ditunno JF Jr et al. 2007{4106} Ditunno JF Jr et al. 2007{4105} Ditunno, PL et al. 2001 [66] Van Hedel, HJ et al 2005[67]	Video segments 'representative group' of 40 patients. 284 patients, 65% male, mean age 50. 38% traumatic aetiology. Wide range of SCI (using ASIA grading). 150 SCI patients USA & Europe - but full assessment on n=77. Few details of sample characteristics. 146 patients from 6 regional SCI rehabilitation centres. 78% male, mean age 32. Asia gradings B, C, & D. Retrospective analysis. 103 patients. Few details given. 60% male. 75 SCI patients able to walk selected from all SCI patients attending one hospital for rehabilitation or ambulation check-up.
Needs Assessment Checklist (NAC) Kennedy P and Hamilton LR. 1999.[24]	Patients with SCI in a rehabilitation setting.	9 core domains, with 216 behavioural indicators (since reduced to 199). 4 levels of scores (0-3) (0: totally dependent, 3: totally independent)	Devised by specialist health care providers. Methods not stated.	Self-rated or administered by key worker with the patient to facilitate individual assessment and goal setting.	Berry C & Kennedy P. 2002.[59] Kennedy P et al 2003[43]	Tertiary care SCI centre. 43 patients, 88% male, mean age 42. 14% complete tetraplegia, 37% incomplete tetraplegia, 23% complete paraplegia, 26% incomplete paraplegia. 192 newly injured SCI patients. 77% male, mean age 41. Wide range of extent of injuries.
The Spinal Cord Injury functional Ambulation Inventory (SCI-FAI) Field-Fote E et al. 2001.[25]	Patients with incomplete SCI in a rehabilitation setting.	Three key domains of walking function in individuals with SCI: gait parameters, assistive device use and timed walk.	Devised by 10 specialist physical therapists based on viewing & ranking videotapes.	Assessment by clinical rater.	Field-Fote E et al. 2001. [25]	Videotapes of walking from 10 patients with incomplete SCI. Instrument assessed further on 22 patients with incomplete SCI. 77% male, mean age 32.
The SIP68 De Bruin AF et al. 1994[37]	Patients representing 10 different diagnostic groups including 41/835 (5%) with SCI	Six domains, 68 items (statements). Scores produced through simple addition of ticked (yes/agree=1) items. (No weightings) High score=worse problems.	Original SIP items developed with input from patients, carers, clinicians. Item reduction for SIP68 using principal comp't analysis.	Self-completion or assisted by interview (including by telephone).	De Bruin AF et al. 1994[37] Post M et al 1996[34] Nanda U et al 2003[35]	N=835 subjects from 10 diagnostic groups. 41/835 5% with SCI. 315 patients interviewed at home 4-10 years following rehabilitation at one specialised centre for varying degrees of SCI. 75% male, mean age 39. 398 patients with various disabilities including 40 (10%) with SCI - all male.

Table 2. Summary of measurement properties of health status instruments cited in table 1, where (if) evaluated in the context of spinal cord injury.

Instrument	Number of published studies where instrument has been used with SCI patients	Ease of use 1-3[a]	Floor (F[b])/ Ceiling (C) effects	RELIABILITY			VALIDITY	RESPONSIVENESS/ SENSITIVITY
				Internal consistency (Cronbach's alpha(α))	*Test/re-test (ICC) or Kappa (K)*	*Inter-rater or inter-observer*	*Construct, convergent, concurrent or discriminant..*	
Barthel Index of disability(BI) Mahoney F, Barthel D. 1965.[14] */BI modified scoring system Shah, 1989*[15]	>10	3	Overall F/C effects likely[51] but limited evidence. Item level F & C effects noted eg. feeding[41]	No evidence in SCI context	No evidence in SCI context.	No evidence in SCI context.	No evidence in SCI context.	Limited comparative evidence in SCI context suggesting less sensitive than the QIF.[19] Similar effect size to FIMmotor.[51]
Modified BI (MBI) - Turkish version[27]	1	3	Floor effect - stairs item.	0.88 at admission, 0.90 at discharge[27]	0.77 (ICC) [27]	Kappa >0.5[27]	*Construct* Rasch model: misfit to construct for bladder and bowel items. *Convergent* Overall agreement (Kruskal Wallis test) with ASIA impairment scale p≤ 0.005 at admission and discharge. Items' correlation with ASIA scores: ASIA motor scale: range r 0.25 to 0.69, ASIA sensory scale: range r 0.17 to 0.63.[27]	No evidence in SCI context
Functional Independence Measure (FIM).[16] */ Adapted FIM (2 scales)/ Shortened form 'the Fone FIM'*[18] [33,]	>10	3	Substantial ceiling effect with cognitive scale.[30] Lesser F/C effects than BI[51]	Locomotion subscale low α 0.4[42] Motor subscale α 0.94 Cognitive subscale α 0.90[47]	Repeated measures on two sites of one sample.[68]	Patient interview versus observation by clinician Spearman r = 0.94.[29]	*Concurrent* Statistically significant (trend) score decrement between patients with varying ASIA levels.[46]. *Construct* Rasch model: misfit to construct for bladder and bowel items.[31] [17] Rasch analysis: scoring system invalid in cross-cultural comparison.[30] *Convergent* Low correlation between cognitive subscale and formal neuropsychological testing.[69] *Construct* Factor analysis supported motor & cognitive dimensions.[47]	Comparative evidence Suggesting FIM less sensitive than the QIF.[46] Motor subscale effect size 0.90 between admission and discharge=similar to BI.[51]
FIM Turkish version[17]	1	3	Ceiling effect with communication & cognitive subscales.[17]	Overall Motor: α 0.93 Overall Cognit.: α 0.98.[17]	Motor: 0.90 Cognit.:0.98.[17]	Kappa >0.48.[17]	*Convergent* Varied correlations (range0.02 soc cognit to 0.58 sphincter) between FIM motor scores & ASIA motor/sensory scores. No correlation between FIM cognit. scores & ASIA motor/sensory scores.[17]	Statistically significant score changes between admission and discharge.[17]

Instrument	Number of published studies where instrument used with SCI patients	Ease of use 1-3[a]	Floor (F[b])/ Ceiling (C) effects	RELIABILITY			VALIDITY	RESPONSIVENESS/ SENSITIVITY
				Internal consistency (Cronbach's alpha)	*Test/re-test (ICC) or Kappa*	*Inter-rater or inter-observer*	*Construct, convergent, concurrent or discriminant..*	
Quadriplegia Index of Function (QIF) [36]	At least 5	3	No evidence found.	No evidence found.	No evidence found.	Pearson (r) range 0.55 to 0.95 between 3 raters' scores (all p<0.001).[19]	*Convergent* Statistically significant correlations, range r=0.64 to 0.91, between QIF scores & ASIA motor, light touch & pinprick scores.[46]	Evidence of superior responsiveness to the BI[19] and the FIM.[46]
Short-form QIF [20]	1	1	Ceiling effect in subjects with low level tetraplegia.	α 0.89[20]	No evidence found.	No evidence found.	*Convergent:* correlation with Upper Extremity Motor Score r=0.82[20] *Concurrent/discriminant:* progressive ('dose response') mean scores by ASIA motor level group (all but one - C7 vs C8 - statistically significant from other group scores).[20]	No evidence found.
Spinal Cord Independence Measure (SCIM) [21]	~10 (all versions)	2	No evidence found.	No evidence found.	K range 0.66 - 0.73[44]	Total agreement 85%. K range 0.66 - 0.98[21]	*Convergent:* correlation between SCIM and FIM r =0.85 p<0.01.[21]	Superior responsiveness to FIM suggested, although some contradictory findings.[21;52]
Catz-Itzkovich revised SCIM (SCIM II) [52]		2				K range 0.4 - 0.6, r =0.78 - 0.86 p<0.0001[44]	*Convergent:* correlation between SCIM II and FIM r =0.84 p<0.001.[48] *Construct:* Rasch model identified flaws including: Wheelchair-car transfer showed misfit and multidimensionality. Respiration also showed misfit. A number of taskss' scores exhibited disordered thresholds.[70]	
Catz-Itzkovich revised SCIM (SCIM III) [3]		2	Floor effect of mobility subscale.[3]	α >0.7[22]	ICC 0.94 to 0.97; K values 0.63 to 0.82 (p<0.001) [22]		*Construct:* Rasch model: confirmed unidimensionality for 3 subscales. Misfit identified in 'walking outdoors' & 'stair management' items of mobility subscale, & 'toilet use' within respiration/sphincter (RS) subscale. RS subscale had disordered category thresholds with some items eg. bowel management.[3 22].	Superior responsiveness to FIM.[53]

Instrument	No. of published studies where instrument used with SCI patients	Ease of use 1-3[a]	Floor (F[b])/ Ceiling (C) effects	RELIABILITY			VALIDITY	RESPONSIVENESS/ SENSITIVITY
				Internal consistency (Cronbach's alpha(α))	*Test/re-test (ICC) or Kappa (K)*	*Inter-rater or inter-observer*	*Construct, convergent, concurrent or discriminant..*	
Walking Index for Spinal cord injury (WISCI) [23]***/ WISCI-II*** [66]	>5	1	Ceiling effect at 6 months from SCI onset chiefly in relation to. walking function.[39]	No evidence found.	No evidence found.	Rating videotaped patients. Kendall coefficient of concordance (W) =0.87 p<0.001.[23]	*Convergent:* Stat signif correlation r=0.77 (p<0.001) with FIM.[23] *Construct:* High/significant correlations range r= -0.88 or -0.89 with 3 timed measures of ambulation.[67] *Convergent* Stat signif correlations range r=0.48 to 0.88 in change scores with LEMS, LFIM, FIM & Berg Balance scale.[39;50] *Convergent:* Stat signif correlations (all p<0.001) with BI and Rivermead mobility index (RMI) both r=0.67, SCIM r=0.97, LEMS r=0.58 & FIM r=0.7.[49]	Superior responsive-ness to LFIM & SCIM may simply reflect measurement of different domains.[39] Limited evidence WISCI has superior sensitivity to walking recovery than the BI, RMI, SCIM, LEMS or FIM - comparing distribution of scores.[49]
Needs Assessment Checklist (NAC) [24]	~5	1	No evidence found.	Mean α 0.86 all scales >0.7.[24] α range 0.7 to 0.9 [43]	Correlations range r=0.69 bladder, to 0.90 for skin management p≤ 0.01.[24]	No evidence found.	*Convergent:* Statistically significant correlations range 0.47 to 0.85 (p≤ 0.01) with SCIM & Hospital Anxiety & Depression Scale (HADS).[24]	No evidence found.
The Spinal Cord Injury Functional Ambulation inventory (SCI-FAI) [25]	1	1 - for an experienced physio-therapist	No evidence found.	No evidence found.	ICC =0.70, 0.80 & 0.84 for live video sessions (VS) 1 and 2 ratings respectively.[25]	No evidence found.	*Convergent:* Correlations between gait score and walking speed: r= -0.74 and -0.70 for VS1 and VS2 respectively; & with self-report of walking ability: r=0.70. Pre- & post-training correlation with LEMS was r=0.74 & 0.64 respectively.[25]	Limited evidence of sensitivity: %change in gait score significantly related to change in LEMS.[25]
The SIP68 [37]	~2	1	Ceiling effects with 3 scales: psych autonomy & communication, emotional stability, mobility range.[35]	α range 0.72 (emotional stability) to 0.85 (mobility control).[37] α range 0.68 (emotional stability) to 0.91.[34]	ICC Scores range 0.61 to 0.90.[35]	ICC for index-proxy pairs range 0.26 (mental health) to 0.85 (somatic autonomy).[35]	*Construct:* Factor analysis (sample contained SCI patients) : 6 factor solution, where original SIP[71], had had 12 subscales.[37] *Convergent:* Stat signif correlations with relevant SF-36 scales (phys func r= -0.66, mental health r= - 0.57) & ADL (somatic autonomy r=0.81)[35]; & with BI (som aut r= -0.91, mobility range r= -0.54)[34] *Convergent:* Stat signif correlations with level of lesion for subscales som aut (r= -0.72), mobility control (r= -0.47). Remaining scales range r= -0.07 (emotional stability) to r= -0.39 (mobility range).[34]	No evidence in SCI context.

[a] 1=rated as 'very straightforward' to apply and score, 2='moderately straightforward', 3='complex' requiring considerable training.
[b] clusters of scores near the lower limit of the data in a study