MONASH University

ISCRR
Institute for Safety, Compensation and Recovery Research

Best Evidence for Measuring Spinal Surgical Outcomes:

A Snapshot Evidence Review

Mary Alice O'Hare

(**FINAL REPORT**)

5 May 2015

Research report number: 131.2-0515-R01

A joint initiative of

WorkSafe VICTORIA | TAC TRANSPORT ACCIDENT COMMISSION | MONASH University

This research report was prepared by:

Dr Mary Alice O'Hare, Institute for Safety, Compensation & Recovery Research (ISCRR), Monash University

For:

WorkSafe Victoria and the Health and Disability Strategy Group (HDSG)

Acknowledgements:

Samantha Barker (ISCRR)

Verna Smith (ISCRR)

Dr Clarissa Martin (ISCRR)

TABLE OF CONTENTS

Tables

Figures

ABBREVIATIONS

BPI	Brief Pain Inventory
ePPOC	Electronic Persistent Pain Outcome Collaboration
HR-PRO	Health Related Patient Reported Outcomes
IPT	Implantable Pain Therapy
MCS	Mental Component Scale- subscale of the SF-36
MOS	Medical Outcome Study
NASS	North American Spine Society
NRS	Numeric Rating Scale
NSFHS	National Survey of Functional Health Status
ODI	Oswestry Disability Index
ORQ	Occupational Role Questionnaire
PCS	Physical Component Scale- subscale of the SF-36
PGIC	Patient Global Impression of Change
PRO	Patient Reported Outcomes
PS	Prolo Scale
RDQ	Roland-Morris Disability Questionnaire
ROC	Receiver Operating Characteristic
SCOPE	Spinal Cord Outcomes Partnership Endeavour
SF-12	Medical Outcomes Study 12-item short-form health survey
SF-36	Medical Outcomes Study 36-item short-form health survey
SIP	Sickness Impact Profile
TAC	Transport Accident Commission
VAS	Visual Analogue Scale
VRS	Verbal Rating Scale
WHO	World Health Organisation
WL-26	Work-Limitations 26-item survey

EXECUTIVE SUMMARY

What the project involved

WorkSafe Victoria needs information on the best-practice methods for measuring outcomes post–spinal surgery. The surgeries considered are spinal fusions, and implantable pain therapy (IPT) or neurostimulation. This evidence review considered firstly clinical outcomes and secondly, patient-reported outcomes.

The Health and Disability Strategy Group (HDSG) is also interested in having surgeons use the electronic Persistent Pain Outcome Collaboration (ePPOC), so that surgical outcomes can be benchmarked nationally. Therefore, the Brief Pain Inventory (BPI), a patient Global Impression of Change Instrument (PGIC) and the SF-12 were assessed.

How the project was conducted?

A *'snapshot evidence review'* was conducted to address the research questions. This type of evidence review provides an overview of the evidence on a given topic. Of the 3,790 identified records, 46 studies were included in the review. As the HDSG were particularly interested in three specific instruments, a supplementary search was undertaken and results are briefly presented in the addendum to the present review.

What the project discovered

After spinal surgery: clinical outcomes

- The reported clinical outcomes are spinal fusion status (i.e. how well the vertebrae are fusing at the point of surgery) and complications.
- Basic radiography is the best way to report spinal fusion. However, it has limitations. Therefore, spinal fusion should be reported in one of three categories: solid fusion rate, non-union rate and levels fused.
- Complications are difficult to measure and, as a result, there are few tools to comprehensively measure them. Therefore, complications should be reported as percentage of patients with complications, and breakdown of complications by number.

After spinal surgery: patient-reported outcomes

- The most common patient-reported outcomes after spinal surgery are pain, function and impact on returning to work.
- The SF-36 and Oswestry Disability Index (ODI) measure pain-related interference with activities, and can therefore be used to assess back-specific disabilities.
- The SF-36 measures pain intensity, but it is generic and does not link to pain to a body part. Therefore, it is useful for determining overall health status and quality of life.
- The Visual Analog Scale (VAS), the Numeric Rating Scales and the Verbal Rating Scales measure pain intensity and are easy to use.
- There are no best-practice tools for measuring a client's ability to return to work, but the Prolo Economic Scale has been adapted for this purpose for more than 20 years.

After implantable pain therapy or neurostimulation

- There is no best-practice for assessing outcomes after IPT or neurostimulation. There are, however, some methods that are commonly used, but have not been rigorously tested.
- Commonly used methods to assess surgical outcomes include the SF-36, the ODI, Roland–Morris Disability Questionnaire and work status.

Multidimensional assessment

- Due to the complexity of spinal surgery outcomes, some suggest that a multidimensional approach is more suitable compared with measuring separate patient-reported outcomes.
- The literature often reports improvements at a group level, not an individual level. This may make some tools irrelevant.
- Measuring the 'minimum acceptable outcomes' for patients is a method that asks patients what the minimum outcome they would accept before having surgery, which could be a useful tool.
- A multidimensional assessment would need to include a core set of measurement tools. Table B outlines the core set of instruments for multi-dimensional measurement of spinal surgery outcomes.

The electronic Persistent Pain Outcome Collaboration

- The BPI and the SF-12 are both generic pain measurements tools that are reliable for spinal disorders; however, the PGIC has not been validated.
- The BPI considers two pain measurements: intensity and interference with daily function.
- The SF-12 is not as valid as the SF-36 (the latter has 36 questions, the former 12), but is suitable if time and money are constraints.

What the recommendations are

Table A outlines the recommendations for measuring patient-reported and clinical surgical outcomes.

Table B outline the core set of instruments for multi-dimensional measurement of spinal surgery outcomes.

Table A Recommendations for best-practice methods for measuring post-surgical outcomes.

INTERVENTION	REPORTED BY	OUTCOME	MEASUREMENT
Spinal fusion	Patients	Pain-related interference with activities	Generic: SF-36 Specific: Oswestry Disability Index, Roland-Morris Disability Questionnaire
		Pain intensity	Visual Analog Scale, Numeric Rating Scales, Verbal Rating Scales
		Functional status	SF-36[a]
	Clinicians	Fusion status	Radiography, measuring solid fusion rate, non-union rate and levels fused
		Complications	General assessment that includes the percentage of patients with a complication, and the breakdown of complications (by number)
Implantable pain therapy or neurostimulation[b]	Patients	Back-specific disability	Oswestry Disability Index, Roland–Morris Disability Questionnaire
		Pain level	Visual Analog Scale

a In addition, the Sickness Impact Profile may be useful for severely ill patients.
b Note that the review did not find any best-practice methods for measuring outcomes after implantable pain therapy or neurostimulation; therefore, the measurement tools listed here are commonly used tools.

To benchmark outcomes for the ePPOC, the BPI and SF-12 are suitable for use. If time and money are not significant restraints, the SF-36 is preferable to use compared with the SF-12.

Table B: Core Set of Instruments for Multi-Dimensional Measurement of Spinal Surgery Outcomes

SPINAL FUSION (Blount et al., 2002)		**SPINAL DISORDERS** (Bombadier, 2000a)	
Patient-Reported Outcomes	**Recommended Measure**	**Patient-Reported Outcomes**	**Recommended Measure**
General health status	SF-36 or SF-12	**Generic health status**	SF-36 (version 2)
Back specific disability	Oswestry Disability Questionnaire (ODI)	**Back specific function**	Oswestry Disability Questionnaire (ODI) Roland–Morris Disability Questionnaire
Pain level	(Visual) Analog Scale (VAS) (1-10) (for back or lower leg; or neck for cervical fusions)	**Pain**	SF-36 (bodily pain scale)
Return to work	Prolo Economic Scale	**Work disability**	Work Status (10 categories) *e.g. usual job, restricted duties, paid/unpaid sick leave, unemployed due to health/other reasons etc.*
Patient satisfaction	North American Spine Society Patient Satisfaction Index	**Patient satisfaction**	• Patient Satisfaction Scale (*to measure satisfaction with care*) • A global question (*to measure satisfaction with treatment outcome*)
Medication use	1. % of patients using narcotic medication, non-narcotic medication, no medication. 2. % of patients with significant reduction in medication use (>50%) measured post-operatively.	-	-
Surgical Outcomes	**Recommended Measure**	**Surgical Outcomes**	**Recommended Measure**
Fusion status	Radiographic assessment of: 1. Solid fusion rate 2. Nonunion rate 3. Levels fused	-	-
Complications	A generalised complication rate to include: • Percentage of patients with a complication • Breakdown of complications by number	-	-

1. INTRODUCTION

1.1 BACKGROUND

Currently, WorkSafe do not have a robust method for evaluating post-surgery outcomes and are seeking the best available evidence on surgical outcome measurement for spinal surgery and implantable pain therapy/neurostimulation. In particular, WorkSafe is interested in evidence-based best practice criteria for (i) measuring surgical outcomes that are meaningful to the client and (ii) evaluating the effectiveness of elective surgery in terms of client outcomes.

The present snapshot evidence review was commissioned by WorkSafe as an input to assist with the development of policy and guidelines and to support agent decision making and engagement with professional bodies. Results from the review will also be used to support the development of research that examines the health and return-to-work outcomes of surgical interventions.

1.2 REVIEW PURPOSE AND RESEARCH QUESTIONS

The overall aim of the present snapshot review is to identify the best available evidence on the measurement of surgical outcomes. Thus, the review will address the following questions:

Q1: What is the most effective practice to measure surgical outcomes based on best available evidence?

Q2: What is the most effective practice to measure client-focused outcomes resulting from surgery based on the best available evidence?

2. METHODS AND SCOPE

A *'snapshot evidence review'* was conducted to address the research questions. A *snapshot evidence review* provides an overview of the current, publically-available evidence to answer specific research questions. Unlike systematic evidence reviews, the quality of research outputs (e.g., peer-reviewed articles, grey literature reports etc.) is not formally assessed. Findings from a snapshot evidence review may be used to inform knowledge and thinking on a particular topic, or to guide decision-making regarding policy and/or practice, when they are included as one of a number of important inputs rather than as the sole input.

2.1 SEARCH

Initially, a large-scale systematic search of the peer-reviewed academic literature was attempted for the top nine elective surgeries (in terms of total spend) identified by Worksafe and the TAC. However, due to the large number of identified records, the search was restricted to two databases (Ovid Medline and Scopus) and two surgeries of highest priority for WorkSafe (spinal surgery and implantable pain therapy/neurostimulation). In consultation with a surgical advisor, the following combination of search terms was used to interrogate the databases:

Table 1: Search Terms Used for the Systematic Database Search

Key Area	Search Terms
1. Surgical outcome	surgical outcome OR surgical complication OR post-operative outcome OR post-operative complication OR surgery outcome OR clinical outcome OR adverse event OR surgical success OR surgical failure
2. Patient outcome	quality of life OR return to work OR pain OR functional improvement OR patient-reported outcome OR impression of change OR subjective OR objective OR psychological OR mental health
3. Measurement	define OR definition OR measure OR measurement OR grade OR standard OR classification OR classify OR quality assessment
4. Surgery type	spine surgery OR spine operation OR spinal fusion OR spinal disorder surgery OR spinal disorder operation OR implantable pain therapy OR intrathecal infusion OR intrathecal pump for pain OR neurostimulation

Inclusion/exclusion criteria. The following search limits were applied: written in English; published from 2000 until the present (February, 2015); human subjects. To establish a level of scientific rigour to assist with the identification of 'best available evidence', only peer-reviewed articles, systematic reviews and reviews were included; all grey literature was excluded. Please note, the outcomes identified by the WorkSafe as client-focused (e.g., return-to-work, pain reduction, functionality and quality of life) are classified as patient-reported outcomes (PROs) in the literature. Therefore, client-focused outcomes will be referred to as patient-reported outcomes (PROs) in the present snapshot review. To identify best evidence (vs. common practice), articles whose primary purpose was to examine,

compare or review PROs and/or surgical outcomes, and/or the instruments/techniques that measured them, were included. Articles whose primary purpose was to compare different *treatments* were excluded. Furthermore, due to the larger number of instruments developed *ad hoc* for specific studies or medical specialities – a literature review found that over 90 different instruments have been used for low back pain alone (Zanoli et al., 2000) – those instruments that had not been subjected to rigorous psychometric testing and were not widely used were excluded.

Search strategy. To effectively answer the research questions, an iterative approach was adopted to assist with the identification of 'best available evidence'. Specifically, early searches – undertaken to determine the best approach to identify best evidence – revealed that PRO were mostly measured by self-report instruments, which quantify PROs in a standardised manner (Beaton, 2000). Therefore, the psychometric (measurement) properties of tests were used as a proxy criterion for 'most effective practice' and 'best available evidence' (see Section 2.2 for a fuller discussion). Finally, as the HDSG were particularly interested in three specific instruments, a supplementary search was undertaken and results are briefly presented in the addendum to the present review.

2.2 DEFINING EFFECTIVE PRACTICE IN PRO MEASUREMENT

Effectiveness can be measured in multiple ways. For the purposes of this review, 'most effective practice' was determined by the ability of the instruments to effectively measure patient-reported outcomes, that is, by their measurement properties. Three measurement properties identified as critical for assessing the quality of Health-Related PRO (HR-PRO) instruments are reliability, validity and responsiveness (Mokink et al., 2010). These three qualities need to be established prior to selecting an instrument for use in clinical settings (Chapman et al., 2011). [See Table 2 for definitions].

Table 2: Definition of Key Measurement Properties to Assess the Quality of HR-PRO Instruments*

Measurement property	Definition
Reliability*	The degree to which the measurement is free from measurement error
Validity	The degree to which an HR-PRO instrument measures the construct(s) it purports to measure
Responsiveness	The ability of an HR-PRO instrument to detect change over time in the construct to be measured

Note. *Based upon the consensus of a panel of 43 experts (extracted from Mokink et al. (2010)). *Reliability includes the more common notion of 'consistency', that is, an instrument's ability to consistently measure constructs.

Link between Research and Clinical Settings. The responsiveness of an instrument is particularly relevant to the present review as the goal of spinal surgery or treatments such as implantable pain therapy (IPT) is to improve patient outcomes (or at least prevent deterioration). Therefore, the ability of instruments to detect change over time is critical. Responsiveness enables the clinical relevance and importance of change scores (e.g., the difference in scores before and after treatment) to be interpreted (Beaton, 2000). Four key

factors influence the magnitude of change in health outcomes as measured by different instruments (Beaton, 2000):

1. **Target population** (*e.g., patients with acute vs. persistent pain will have different clinical outcomes and, therefore, a different amount of change*).
2. **Treatment type** (*i.e., some treatments will result in bigger changes*).
3. **Timing of data collection** (*i.e., smaller change is likely for shorter intervals between data collection points*).
4. **The construct of change being measured** (e*.g., before-and-after scores for a treatment with known efficacy*).

Therefore, before responsiveness can be applied in clinical settings, the first three points (i.e., target population, treatment type and timing of data collection) must match those used in the research setting. However, the final point – the construct of change – is the 'key to interpreting the meaning of a change score' on HR-PRO instruments (Beaton, 2000, p. 3193). Therefore, results from self-report instruments that measure HR-PRO outcomes "cannot inform policy or treatment decisions until they can be interpreted from the perspective of their clinical relevance and meaning" (Beaton, 2000, p. 3190).

3. RESULTS

3.1 SEARCH RESULTS

A total of 3,790 publications were identified. Of these, 46 records (1.2%) were retained for inclusion in the present review (See Figure 1). None of the implantable pain therapy / neurostimulation articles met inclusion criteria. The majority of included articles were reviews (50%, n = 23) – 32% of which were systematic reviews, with the remainder being literature reviews and evaluations – followed by prospective studies (26%, n = 12) and retrospective studies (11%, n= 5). The remainder of articles (13%, n= 6) were classified as 'other' (e.g., survey, mixed).

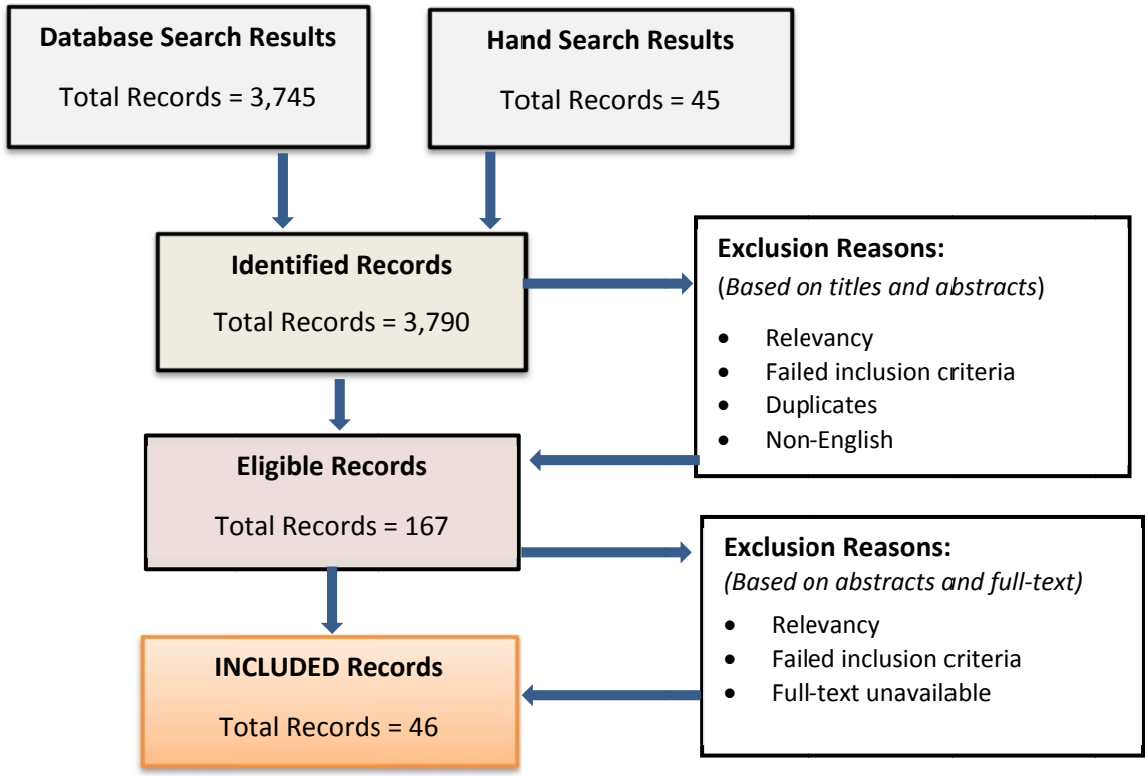

Figure 1: Article Selection Process (PRISMA Diagram)

3.2 SURGICAL OUTCOMES (SPINAL)

3.2.1 Question 1- What is the most effective practice to measure surgical outcomes based on best available evidence?

Summary

The literature presented recommends that spinal fusion be reported as (i) solid fusion rate; (ii) non-union rate; and (iii) levels of fusion. It is recommended that complications be reported as (i) percentage of patients with a complication and (ii) breakdown of

Historically, spinal fusion success has been evaluated in terms of a "solid radiographic fusion" (Glasson et al., 2008). Other traditional indices used for assessing surgical outcomes following spinal surgery are operative measures such as fusion status, reoperation rate, deformity correction or complication rate (Blount et al., 2002; McCormick et al., 2013). The following discussion will focus on fusion status and complications.

a) Fusion Status

As bony union is the major goal of fusion surgery (Tuli et al., 2004), lumbar spinal fusion surgery success has traditionally been determined by the achievement of a solid radiographic fusion (Blount et al., 2002; Glassman et al., 2006). However, plain radiographs have limited ability to effectively determine fusion status (Glassman et al., 2006). For example, plain static radiographs were found to be "generally quite unreliable" for predicting cervical fusion based on the presence or absence of trabecation[1] – one of the most important criteria for predicting fusion (Tuli et al., 2004), which may also be confounded by under- or over- exposure of the radiographs (Sudakawar et al., 2003). Furthermore, the use of spinal implants has made assessment of bone fusion sites more difficult, which can influence evaluation of fusion status (Blount et al., 2002). Unsurprisingly, therefore, only two literature review articles, both of which included a short section on radiographic measurement, met inclusion criteria (Blount et al., 2002; Schoenfeld & Bono, 2011).

The oldest literature review (Blount et al., 2002), which included a short section on the radiographic measurement of fusion status, discussed some specific factors that influence accurate outcome assessment (e.g., spinal implants and imaging techniques). Therefore, Blount et al. (2002, p. 21) recommended that "the criterion and imaging techniques used to establish fusion be clearly delineated and that fusion be evaluated by two independent observers based on the examination of plain radiographs taken 1 year after surgery". Due to a reported correlation between pseudoarthrosis (fusion failure) and functional failure, Blount et al. (2002) recommend that fusion rates are reported in all fusion studies. Furthermore, as the number of operated levels has been shown to significantly influence the fusion rate, this figure should also be reported. Overall, for the sake of consistency and simplicity, it is recommended that fusion be reported only as rates of fusion and rates of non-union (vs.

[1] "On plain radiograph films, bony trabeculae are typically seen traversing the graft and the host bony interface" (Tuli et al., 2004, p. 856). The presence of trabecation was used as the criterion for fusion its absence was used as the criterion for non-union.

fused, partially fused and non-fused). Therefore, when reporting fusion status, the following should be reported:

- Solid fusion rate
- Non-union rate
- Levels of fusion

The most recent literature review (Schoenfeld & Bono, 2011), which briefly discussed radiographic outcome measures for assessing fusion following spinal trauma, focused on the shortcomings of radiographic assessment of treatment. The review concluded that most spine trauma studies still rely on 'soft' (non-validated) measures to assess radiographic results of treatment. Specifically, 10 of the 17 reviewed studies of spine trauma (published from 2001 to 2009) relied upon determinants of focal kyphosis as indicators of fusion failure, despite "significant advances in outcomes research" (Schoenfeld & Bono, 2011, p. 268). Reporting deficiencies were also noted: five of the reviewed studies did not include radiographic parameters when evaluating outcomes and a further two studies did not use validated or accepted radiographic measures for reporting fusion status (relying instead on the authors' own methods).

b) Complications / Adverse Events

Complications following spinal surgery are substantial (Street et al., 2012), affecting an estimated 10-20% of spinal patients (see Nasser et al., 2010). Therefore, the benefit of performing spinal fusion must be greater than the risk of complications (see Blount et al., 2002). However, due to the 'complex nature' of spinal surgery, and the wide speciality-driven variation in definitions, assessment of complications can be difficult. Interpretation of the literature can also be challenging due to variations in the perception, and reporting, of complications (Lebude et al., 2009). Clear definitions of surgical complications, coupled with standardised and reproducible methods for reporting them, are required to reliably measure this type of surgical outcome (Dindo et al., 2004; Lebude et al., 2009; Street et al., 2012). The following section will review the literature on the definition, classification and measurement of spinal surgery complications.

Definition. While the undesirable nature of complications is acknowledged, there is no clear consensus on their definition (Nasser et al., 2010). Some definitions have focused on featural differences of complications such as their unexpected nature (Dindo et al., 2004) or connotations of blame (Mirza et al., 2004) while others have linked them to clinical outcomes such as length of hospital stay (Rampersaud et al., 2006). Furthermore, disagreement as to when a medical event actually constitutes a complication is also apparent. Several factors, such as surgical speciality and experience, may account for this. Specifically, an event was less likely to be deemed a complication by neurosurgeons (vs. orthopaedic surgeons), by surgeons who had practiced more than six years, and who had performed more than 25% of fusions (Lebude et al., 2009). Therefore, a 'working' definition of spinal surgery complications – based upon survey responses of 229 surgeons – expressed complications in terms of their severity:

- "Minor complication: An adverse perioperative event that produces only transient detrimental effect including medical adverse events in the perioperative period".
- "Major complication: An adverse perioperative event that produces permanent detrimental effect or requires reoperation. This entails all medical adverse events occurring in the perioperative period (30 days from time of surgery), regardless of their direct connection to the specific surgical procedure performed" (Lebude et al., 2009, p. 498).

A distinction is also made between complications and adverse events. Although frequently used interchangeably, an adverse event is any unexpected or undesirable event arising from an intervention, which may or may not lead to a complication. Therefore, "all complications are adverse events but not all adverse events are complications" (Ohnmeiss et al., 2010).

Classification. "A major limitation" in reporting complications is the lack of a standardised system for reporting or grading complications (Tevis & Kennedy, 2013). The classification systems reviewed below commonly used some form of severity grading of post-operative complications, although the exact criteria differed among studies. However, one study (Rampersaud et al., 2006) exclusively graded intra-operative complications (i.e., complications that occurred during surgery). Furthermore, classification development differed for general surgery vs. spinal surgery complications. *General surgery systems* grouped complications by broad, clearly defined criteria such as, severity (i.e., level of intervention required to correct complications (Dindo et al., 2004; Mazeh et al., 2014)), etiology (i.e., error diagnosis) (Antonacci et al., 2009; see also Mirza et al., 2006) or major body system (e.g., cardiovascular system, pulmonary system etc.) (Antonacci, 2008). [See Appendix A for a full description.] In contrast, *spinal surgery systems* were less well-developed and defined, possibly due to the complexity of specific spinal pathology (Mirza et al., 2006).

Spinal surgery systems for classifying complications were fragmented and less developed than general surgery systems, indicating that "further work is required in critically assessing spine surgery complications" (Lebude et al., 2009, p. 493). Only one well-organised classification system for spinal surgery complications was identified in the present review. The system graded complications according to severity (Rampersaud et al., 2006) and its applicability was confirmed in a later study.

Severity classification. A five-point complication grading system was used to classify intra-operative adverse events (i.e., those that occurred during surgery) according to their clinical consequences (i.e., length of hospital stay). Length of hospital stay was influenced by whether complications required further treatment.

Grade 0: No complications (no post-operative events)

Grade 1: Minor (none or minimal treatment - minimal effect on length of stay (1 day))

Grade 2: Moderate (treatment required - length of stay increased by 2-7 days)

Grade 3: Major (significant treatment - length of stay increased by more than 7 days)

Grade 4: Death

The above grading system is based on analyses of prospectively collected data from 700 spinal surgeries, 58% of which were fusions. Although most (77%) intra-operative adverse

events do not lead to post-operative complications, some specific adverse events are associated with a high likelihood of their occurrence (e.g., dural tears with persistent cerebrospinal fluid leak, massive blood loss, airway difficulties and esophageal or pharyngeal injuries) (Rampersaud et al., 2006). A later study (Street et al., 2012) used the above system to prospectively estimate the incidence of intra- *and* post-operative complications. Analyses of data from 942 patients (50% of whom underwent elective surgery) confirmed that complications adversely affected length of hospital stay. Furthermore, a comparison of the above system with more traditional data abstraction methods revealed that it was more rigorous and thorough and captured both intra-operative and post-operative events (more traditional methods under-represented post-operative events) (Street et al., 2012).

Despite the 'substantial' complication risk associated with spinal surgery, there is "no clearly defined medical literature on complications" (Proietti et al., 2013, p. 340). For example, adverse events are 'often arbitrarily reported as "device-related," "major," or "preventable"' (Mirza et al., 2006, p. 4). The lack of well-defined and tested complication classification systems for spinal surgery most likely reflects its complexity. Unlike more standardised procedures (e.g., hip or knee replacement), spinal surgery is "individualised for the specific spinal pathology" and needs to consider graft materials and fixation devices and procedural variations (Mirza et al., 2006, p. 2). For example, a systematic review of 105 studies found that the type of complications differed according to the spinal segment reviewed. Specifically, thoracolumbar complications were more often procedure-related (e.g., pseudarthrosis and hardware failures) while cervical complications were more often approach-related (e.g., dysphagia and dysphonia) (Nasser et al., 2010).

Measurement. Some surgeries are riskier than others and are associated with a greater incidence of complications. For example, the invasiveness of spinal surgical procedures (e.g., surgical access route, location of nerve roots decompressed, number of vertebrae fused and instrumented) increases the complication risk (Mirza et al., 2006; Prioetti et al., 2013). Other factors that influence the incidence of complications are study design, anatomical region treated, and follow-up duration. Specifically, significantly higher complication incidence was reported in prospective studies compared to retrospective studies[2], for thoracolumbar spine versus cervical spine, and for studies that used longer follow-up periods (Nasser et al., 2010). For example, although a systematic review of 105 studies estimated an overall incidence of 16.4% for spinal surgery complications, wide variations were reported in individual studies (ranging from <1% to approximately 70%) (Nasser et al., 2010).

Based on a literature review that included 27 studies of spinal fusion, Blount et al. (2002) recommended that complications should be reported as:

- Percentage of patients with a complication
- Breakdown of complications by number

[2] This methodology depends upon the accuracy of record-keeping which has been subjected to few "meaningful quality checks", such that missing data are common. Therefore, retrospective studies likely under-estimate the incidence of complications (Street et al., 2012).

3.2.2 Question 2- What is the most effective practice to measure PRO resulting from surgery based on the best available evidence?

Although assessment of spinal surgery outcomes has traditionally relied upon operative measures such as fusion status, reoperation rate or complication rate, inclusion of patient-based evaluation of outcomes is becoming increasingly common (Blount et al., 2002; Bremerich et al., 2006; DeVine et al., 2011; Glassman et al., 2006). As the patient is the principal source of information (Hagg et al., 2001), patient-reported outcomes (PROs) are typically measured with self-report questionnaires. PRO questionnaires, which can be either generic or disease-specific, provide a quantitative estimate of pain, quality of life and functionality (McCormick et al., 2013). Generic instruments apply across a variety of conditions whereas disease-specific tools attribute symptoms to a specific disease or condition (Kopec et al., 2000). The specific PRO instruments identified in the present snapshot review are summarised in Table 3.

Frie et al. (2012) note that the three most common dimensions of PROs that research has focused on are consistent with the well-established *WHO International Classification of Functioning, Disability & Health*, namely:

(i) *Impairments* - Symptoms (i.e., pain)

(ii) *Activity* - Disability (i.e., functional status)

(iii) *Participation* - Quality of life (i.e., impact on work life)[3]

The following discussion of the best available evidence for the measurement of PROs is organised according to these three dimensions to provide a framework for discussing the literature.

[3] Work indirectly influences quality of life.

Table 3: Summary of the Properties of the PRO Instruments Identified for the Snapshot Review

PRO INSTRUMENT	Type	Psycho-metrics*	DOMAINS COVERED							Reference	Study Design**
			Pain	AoD†	Work	Personal Care	Mobility	Mental health	Other - Multiple		
Oswestry Disability Index (ODI)	Back-specific	R, V, Rs	✓	✓		✓	✓		✓	Chapman et al. 2011 Fairbank & Pysent, 2000 Roland & Fairbank, 2000 Kopec, 2000 Von Korff, 2000 Bombadier, 2000a NIskanen, 2002 Von Korff, 2000 McCormick et al. 2013 Schenfeld & Bono, 2011 Zanoli et al., 2000	S.R. Review Review Review Review Other∞ Other Review Review Review Review
Roland Disability Questionnaire (RDQ)	Back-specific	R, V, Rs	✓	✓		✓	✓		✓	Roland & Fairbank, 2000 Chapman et al. 2011 Von Korff, 2000 Fairbank & Pysent, 2000 Bombadier, 2000a Schoenfeld & Bono, 2011 Kopec, 2000 Zanoli et al., 2000	Review S.R. Review Review Other Review Review Review
SF-36 **(Short-Form-36 Questionnaire)**	Generic	R, V, Rs	✓	✓	✓	✓	✓	✓	✓	Ware, 2000 Garrett et al., 2002 Chapman et al., 2011 Schoenfeld & Bono, 2011 Blount et al. 2002 Bombadier, 2000a Glassman et al. 2006 Von Korff, 2000	Review S.R. S.R. Review Review Other Prospective Review

PRO INSTRUMENT	Type	Psycho-metrics*	DOMAINS COVERED							Reference	Study Design**
			Pain	AoD†	Work	Personal Care	Mobility	Mental health	Other - Multiple		
Sickness Impact Profile (SIP)	Generic	R, V, Rs		✓	✓	✓	✓	✓	✓	Pollard & Johnston, 2001 Nemeth, 2006 Lurie, 2000 Turner et al. 2004 Schoenfeld & Bono, 2011 Blount et al. 2002 Bergner et al. 1981	Retrospective Review Review S.R. Review Review Other
Visual Analogue Scale (VAS)	Generic	R, V, Rs	✓							Von Korff, 2000 DeVine et al., 2000 Schoenfeld & Bono, 2011 Chapman et al. 2011	Review S.R. Review S.R.
Numeric Rating Scale (NRS)	Generic	R, V, Rs	✓							Von Korff, 2000 Chapman et al. 2011	Review S. R.
Verbal Rating Scale (VRS)	Generic	R, V, Rs	✓							Von Korff, 2000	Review
The Prolo Scale (PS)	Back-specific	Rs			✓					Vanti et al. 2013 Blount et al. 2002	Review Review
Occupational Role Questionnaire (ORQ)	Back-specific	R, V			✓					Amick et al. 2000	Review
WL-26	Generic	V			✓					Amick et al. 2000	Review
Neck Disability Index (NDI)	Back-specific	R, V, Rs	✓	✓	✓	✓	✓		✓	Blount et al. 2000 Godil et al. 2013 McCormick et al. 2013	Review Prospective Review
Neck Pain and Disability Scale (NPAD)	Back-specific	R, V	✓	✓	✓	✓	✓	✓	✓	Blount et al. 2002 Bremerich et al. 2008	Review Other
Dallas Pain Questionnaire (DPQ)	Back-specific	R, V	✓	✓	✓			✓		Anderson et al. 2008 Blount et al. 2002 Chapman et al. 2011	Retrospective Review S.R.
Graded Chronic Pain Scale	Generic	R, V	✓	✓						Von Korff, 2000	Review

Notes. *Psychometric Properties: R = reliable; V= valid; Rs = responsive. † AoDL – Activities of Daily Living. **Study design: SR = Systematic Review, 'Review' includes literature reviews and evaluation reviews. ∞ 'Other' category includes survey, mixed methodology. ✓= (aggregate role limitations which icludes work, household, leisure roles etc.).

(i) Impairments / Symptoms (Pain)

Summary

As pain is a multi-dimensional construct that comprises severity, affect and persistence general vs specific pain, the choice of instrument will depend on which aspect of pain is to be measured. ***Pain severity*** – which has been well-researched and is relatively straightforward to measure – comprises two highly-related constructs: pain intensity and pain-related interference with activities. The three most widely-tested and used instruments that measure *pain-related interference with activities* are the ODI, RDQ and the SF-36 (bodily pain scale). All three instruments are reliable, valid and responsive. However, the ODI may be better at detecting change in people with higher disability levels and the RDQ may be more suited to people with minor disability. Being generic, the SF-36 (bodily pain scale) is less responsive than disease-specific instruments and does not link pain to any particular body site. Three psychometrically-sound and widely-used generic instruments that measure ***pain intensity*** (i.e., how much it hurts) are the VAS, NRS and VRS. All three instruments are sensitive to treatments that target pain

Although pain is a leading symptom of spinal disorders (Svensson et al., 2009), there is no gold standard for pain assessment (Hagg et al., 2001). Therefore, "the patient him or herself is the basic reference for outcome, for which the instruments give a more or less exact measurement" (Hagg et al., 2001, p. 1213). Thus, validated self-report instruments are the most common method for measuring pain. Considerable research has focused on pain severity and pain affect, however, much less has been directed at pain persistence (Von Korff, 2000). The ten most common instruments that measure pain – disability-specific pain or generic pain – were extracted from the peer-reviewed articles which focused on pain measurement related to spinal disorders (see Figure 2). Although the psychometric properties of all of the listed instruments have been assessed, some instruments (e.g., the ODI and the RDQ) have been subjected to more rigorous examination. As measurement of ***pain severity*** is 'relatively straightforward' – and many unresolved issues surround measurement of the more complex construct of pain affect – instruments that measure ***pain severity*** are recommended (Bombadier, 2000a).

Pain Severity

Growing evidence suggests that pain severity is a global construct comprising two highly related concepts: pain intensity and pain-related interference with daily activities (Von Korff, 2000). The six most tested and commonly used pain severity instruments will be discussed below: three that measure pain-related interference with daily activities (ODI, RDQ and SF-36[4]) and three that focus exclusively on pain intensity (VAS, NRS and VRS). Where available, results of head-to-head comparisons of different instruments will be presented.

[4] Strictly speaking, the ODI and the bodily pain scale of the SF-36 measure pain severity as they also include a rating scale to assess pain intensity.

Pain-Related Interference with Daily Activities

Two widely used and accepted **disability-specific instruments** that measure the effect of pain on interference with daily activities are the Oswestry Disability Index (ODI) and the Roland-Morris Disability Questionnaire (RDQ) (Chapman et al., 2011; Fairbank & Pysent, 2000). Both instruments assess pain-related limitations in daily activities such as mobility, personal care, sleeping, sitting, lifting and household activities (Roland & Fairbank, 2000). The ODI also includes a pain intensity scale (as did the original form of the RDQ). The RDQ - which was derived from the Sickness Impact Profile (SIP) - comprises 24 'yes/no' questions while the ODI comprises ten items (each with six options) (Kopec, 2000; Roland & Fairbanks, 2000). Both instruments can be completed in five minutes (Von Korff, 2000).

PAIN: 10 Commonly Used Measures

Disability-Specific Measures

Oswestry Disability Index (ODI)

Roland-Morris Disability Questionnaire (RDQ)

The Neck Pain and Disability Scale (NPAD)

Neck Disability (NDI)

Dallas Pain Questionnaire

Generic Measures

Visual Analogue Scale (VAS)

Verbal Rating Scales (VRS)

Numeric rating scale [NRS]

The bodily pain subscale of the SF-36

Graded Chronic Pain Scale

Figure 2: Ten commonly used and tested pain measures (spinal)

- ***Direct Comparison- ODI vs RDQ***

  Responsiveness. The ODI and RDQ are highly correlated with each other (0.77), indicating that they are measuring similar constructs. However, as the ODI tends to score higher than the RDQ, it may be better at detecting change in people with higher disability levels and the RDQ may be more suited to people with minor disability (Bombadier, 2000a; Fairbank & Pysent, 2000). Receiver Operating Characteristic (ROC) analyses – which assess the ability of an instrument to detect change – revealed that the ROC Index for the ODI was 0.76, a figure described as "acceptable" but lower than that found for the RDQ, probably due to sample characteristics (i.e., people with lower disability levels). Further work on the responsiveness of the ODI and the RDQ is recommended at both the group[5] and individual level (Fairbank & Pysent, 2000).

  Reliability[6]. Psychometric testing has demonstrated that both the ODI and RDQ are reliable, with similar reliability co-efficients. Depending on the type of reliability assessed (i.e., test-retest or internal consistency), estimates ranged from 0.71 to 0.91 for the ODI

[5] The magnitude of improvement on the ODI differs according to diagnostic group, therefore, well-defined study populations are critical (Glassman et al., 2009).
[6] Reliability co-efficients, which provide an index of reliability, range from '0' (completely unreliable) to '1' (completely reliable) (Nunnally, 1978). A score within the range of 0.70 to 0.90 is recommended to demonstrate reliability (Roland & Fairbank, 2000).

and from 0.72 to 0.91 for the RDQ (Fairbank & Pynsent, 2000; Niskanen, 2002; Roland & Fairbank, 2000).

Content validity. The RDQ focuses on a limited set of physical activities that are highly likely to be affected by back pain and does not address psychological distress or social problems (although it does include one item on mood) (Roland & Fairbank, 2000). The ODI also (deliberately) does not address the psychological consequences of pain but does have items related to social life (Fairbank & Pynsent, 2000). Therefore, when assessment of the psychological or social consequences of pain is important, it is recommended that the RDQ be used in conjunction with measures that specifically address these areas (Roland & Fairbank, 2000).

Construct validity[7]. Both the ODI and the RDQ possess construct validity. Analyses revealed a moderately strong relationship between the ODI and the VAS (which measures pain intensity) and the SF-36 (Fairbank & Pysent, 2000). Similarly, the RDQ was correlated with the physical subscales of the SF-36 (which include bodily pain) and the SIP (from which it was derived) and pain ratings (Roland & Fairbank, 2000). As the ODI is one of the most widely used disability-specific pain measures (Chapman et al., 2011), it is also used to validate other pain instruments (Fairbank & Pynsent, 2000). However, non-validated versions of the ODI exist (Fairbank, 2007), therefore, use of version 2.0 is recommended (Fairbank & Pysent, 2000) while an expert panel recommended that the original version of the RDQ be used (Roland & Fairbank, 2000).

Timeframe. Both the RDQ and the ODI measure current pain: the RDQ explicitly directs patients to report their pain experience *today* while the ODI assesses pain *at the moment* (the remainder of the ODI questions carry an implicit instruction for assessing pain *now* (Fairbank & Pysent, 2000)). Such an approach is suitable for monitoring short-term changes in back pain in primary care settings or in response to treatment (Roland & Fairbank, 2000). Assessing current pain is also regarded as more robust than reporting average pain over the preceding week as done in some other instruments (Fairbank & Pysent, 2000).

- ***SF-36***

A widely used **generic measure** of pain severity is the bodily pain scale of the SF-36 (Medical Outcomes Study 36-Item Short-Form Health Survey). The bodily pain scale includes two items that measure pain intensity (ranging from 'none' through to 'very severe') and interference with daily activities (ranging from 'not at all' to 'extremely') in the last four weeks. In addition to its brevity – a key factor that influences use in clinical and research settings – the bodily pain subscale has "strong psychometric support and extensive normative data" (Bombadier, 2000a). For example, reliability co-efficients ranging from 0.78 to 0.96 have been reported and validity has been demonstrated by the correlation between the subscale and other measures of pain severity and functional disability (Von Korff, 2000). Furthermore, extensive normative data are available for the bodily pain subscale, which facilitates interpretation of scores and enables monitoring of

[7] Construct validity is established by comparing the scores of an instrument with the scores of other established instruments that measure a similar factor (Roland & Fairbank, 2000).

disease groups and treatments over time (Ware, 2000). However, being generic, the SF-36 does not link pain to any particular body site nor does it measure pain persistence as it is limited to pain occurrence in the last four weeks (Von Korff, 2000).

Pain Intensity

Pain intensity is a "quantitative estimate of the severity or magnitude of perceived pain" and measures 'how much it hurts' (Von Korff, 2000, p. 3142). Three **generic instruments** commonly used to measure pain intensity are: Visual Analog Scales (VAS); Numeric Rating Scales (NRS); and Verbal Rating Scales (VRS)[8].

- ***Direct Comparison: VAS, NRS and VRS***

  VAS depict pain intensity along a line (usually 10cm long), with the end points of the line labelled as either 'no pain' or 'pain as bad as it gets'. Patients identify any point along the line that describes the pain they feel. NRS require patients to rate their pain intensity on a scale from 0 – 10 (or 0-20 or 0-100), where '0' represents 'no pain' and '10' represents 'pain as bad as it gets'. Patients either pick a number verbally or circle a number on a scale. VRSs comprise a list of adjectives that describe different levels of pain intensity, with the first and last adjective describing 'no pain' and 'pain as bad as it gets' Patients read the list and select the adjective that best describes the pain they feel (Von Korff, 2000).

  All three generic instruments (i.e., the VAS, VRS and NRS) are easy to administer, their construct validity (i.e. their association with other self-report measures of pain intensity) is well-documented and they all demonstrate sensitivity to treatments that influence pain intensity. Although studies generally report little difference in sensitivity among the three scales, when differences are found, the VAS is typically more sensitive (Von Korff, 2000) and demonstrated high responsiveness (measured as an effect size) after lumbar spine surgery (DeVine et al., 2011). Unlike the other scales, VAS demonstrate a ratio quality for some patient groups (vs. individuals) indicating that scores represent actual differences in the magnitude of pain intensity. For example, a post-treatment decrease in score from 40 to 20 would also mean that pain had halved (Von Korff, 2000). However, some patients at risk for cognitive difficulties (particularly the elderly and those taking high doses of opioid analgesics) have difficulty understanding VAS. Therefore, it is recommended that VAS be used in conjunction with other instruments (Von Korff, 2000).

  Limitations and Considerations - Due to their simplicity, both VRS and NRS are easily understood by patients and task compliance is high. However, the VRS may be unsuitable for people with a limited vocabulary, particularly if the list is long, and none of the adjectives may adequately describe patients' pain. Furthermore, the ranked scoring method – in which adjectives are ranked by severity – (incorrectly) assumes that the intervals between mild, moderate and severe pain are equivalent, which is problematic when interpreting differences in the magnitude of scores. A unique advantage of the NRS – which requires the patient to 'pick a number from 0 to 10' – is that it can be administered over the phone. However, NRS do not have a ratio quality and a change in

[8] Although these three scales are generic, they have also been adapted to specific pain types, for example, back pain (BP-VAS; BP-NRS) and leg pain (LP-VAS; LP-NRS) (e.g., Carreon et al., 2009; Godil et al., 2013).

pain score from ‘9’ pre-treatment to ‘6’ post-treatment does not represent a 33% decrease in pain. Nevertheless, Von Korff (2000) recommends the use of NRS in clinical and research settings.

(ii) Activity / Disability (Functional Status)[9]

Summary

Although a distinction is made between disability due to pain and disability due to other causes, this distinction is not always apparent in the literature. For example, the ODI – a disease-specific measure of *pain-related* disability – is frequently used to measure disease-specific disability or dysfunction and some other measure (e.g., the VAS) used to measure pain (e.g., Chapman et al., 2011). Furthermore, disease-specific instruments abound, however, few have been widely tested and used (Kopec, 2000). Therefore, the above discussion focused on the two most evaluated and widely-used generic instruments: the SF-36 and the SIP (Garrett et al., 2002). Their reliability, validity and responsiveness have been established. However, the SF-36 is more responsive to back pain and has been more widely used for spine trauma than the SIP. Nevertheless, the SIP may be useful for severely ill patients as it includes items on profound disability. The SF-36 is most usefully used as part of a 'core' set of instruments and, due to its brevity, is more practical and less burdensome to patients and staff than the SIP.

Although a distinction is made between disability due to pain (discussed above) and disability due to other causes (discussed below), this distinction is not always apparent in the literature, particularly for disease-specific measurement of disability. For example, the ODI – a disease-specific measure of *pain-related* disability – is frequently used to measure disability or dysfunction and some other measure (e.g., the VAS) is used to measure pain (e.g., Chapman et al., 2011).

Therefore, the following discussion is limited to widely-used generic measures of disability. Two extensively used generic instruments that assess global functional status are the SF-36 and the Sickness Impact Profile (SIP) (Chapman et al., 2011; Garrett et al., 2002; McCormick et al., 2013). Although generic health status instruments – which broadly assess health and disability and do not measure symptoms specific to a particular condition – exhibit less responsiveness than disease-specific instruments, they enable population-based comparisons of the relative impact of different conditions or treatments (Lurie, 2000).

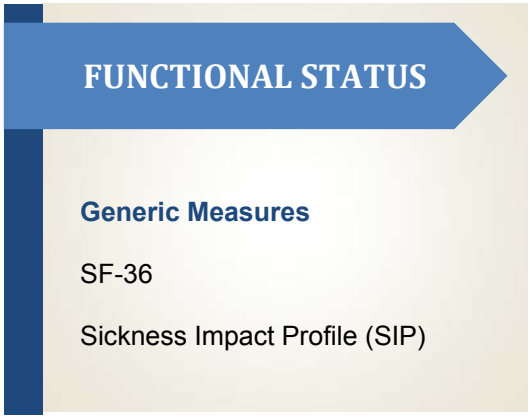

Figure 3: Commonly used and tested functional (general health) status measures (spinal)

[9] The two instruments reviewed in this section - the SIP and the SF-36 - measure health as the absence of dysfunction, although the SF-36 also has an item on vitality (Lurie, 2000).

- ***SF-36***[10]

  In addition to its use as a generic measure of pain severity, the SF-36 is one of the most widely used measures of general health status. The SF-36 has been examined in over 4,000 articles and applied to over 200 diseases and conditions including, spinal disorders, back pain and low back pain (Ware, 2000; n.d.). The SF-36 comprises 36 questions which are classified into two overarching scales: the physical health scale and the mental health scale. The physical health scale comprises four subscales (physical functioning, role-physical, bodily pain, general health) and the mental health scale also comprises four sub-scales (vitality, social functioning, role-emotional, mental health). Although the SF-36 is a generic measure, systematic comparisons have revealed that it includes eight of the most frequently measured health concepts (Ware, 2000).

  Reliability and Validity. Extensive testing has established the psychometric properties of the SF-36. Most studies demonstrate that reliability estimates for the whole test exceed the recommended minimum value of 0.70, with figures typically over 0.80. Reliability estimates for the physical health scale and the mental health scale usually exceed 0.90. The validity of the SF-36 has also been demonstrated, with multiple studies providing evidence of its content, construct, concurrent, criterion-related and predictive validity (the SF-36 has been compared to 225 other health measures).

  Responsiveness. When comparing before-and-after treatment effects, three of the subscales that comprise the physical health scale (i.e., bodily pain, physical functioning and role-physical) are the most responsive to treatments that change physical morbidity (e.g., knee replacement, hip replacement). In contrast, three of the subscales that comprise the mental health scale (i.e., mental health, role-emotional and social functioning) are the most responsive to drugs and therapies that target mental health (e.g., depression) (Ware, 2000). Evidence from "more than 250 longitudinal studies suggests that the SF-36 is also a useful tool for evaluating the benefits of alternative treatments" (Ware, 2000, p. 3137).

  Although frequently used as the sole measure of health outcomes, the 'most useful' studies include the SF-36 as part of a 'generic core' (Ware, 2000, p. 3137). Due to its brevity, it can be used in conjunction with other general measures or more precise specific measures, and is a practical alternative to the much longer Medical Outcome Study (MOS) measure from which it was derived. Although some studies have reported that the SF-36 has 10-20% less precision than the MOS measure, it "rarely miss(es) a noteworthy difference in physical or mental health status in group comparisons" (Ware, 2000, p. 3137). In contrast, longer measures of general health status can place a 5-10 fold greater burden on the patient. Furthermore, compared to the longer Sickness Impact Profile (SIP) – discussed below – the SF-36 is at least as effective in detecting health differences (Ware, 2000).

[10] The SF-36 (Short-Form-36) was derived from the Medical Outcomes Study (MOS). The eight health concepts covered in the SF-36 were selected from the 40 concepts included in the MOS (Ware, 2000).

- ***Sickness Impact Profile (SIP)***

  The Sickness Impact Profile (SIP) was designed to measure outcomes of health care for use in evaluation, program planning and policy formulation (Bergner et al., 1981)[11]. The SIP is a comprehensive, behaviourally-based measure of general health status that ranges from minimum to maximum dysfunction (see Pollard & Johnston, 2001). Minimum levels of dysfunction were included as sensitivity to low levels of dysfunction (or the ability to detect low-level sickness impacts) is critical for comparing differences in diagnostic and treatment groups as well as changes over time (Bergner et al., 1981). The SIP comprises 136 items that assess the impact of illness or disability on 12 areas of activity: sleep and rest, eating, work, home management, recreation and pastimes, ambulation, mobility, body care and movement, social interaction, alertness behaviour, emotional behaviour, and communication. These items are organised into three categories: physical, psychosocial and independent. The SIP can be self-administered or interviewer-administered and only those statements that describe patients on that particular day and are related to their health are checked (Bergner et al., 1981). Thus, the SIP measures performance of specific behaviours versus judgements of capacity (Lurie, 2000). Multiple studies have confirmed the reliability and validity of the SIP, and it has been used in a range of spinal studies (Nemeth, 2006) including assessment of spinal cord stimulation (Turner et al., 2004).

  <u>*Limitations of SIP*</u> - Some problems with the SIP have been identified. Specifically, illogical scoring, the nature and meaning of overall scores, item ambiguity, item order, questionnaire length and the work category (Pollard & Johnston, 2001). For example, due to item weighting, it is possible for a person with paraplegia (unable to walk) to obtain a similar score to a person with arthritis (walk with difficulty) for the ambulation category. Similarly, as some items are mutually exclusive (e.g., *'I do not use the stairs at all'* precludes checking another item, '*I go up and down stairs more slowly'*), the score of a person with the most severe limitations would not reflect their actual level of dysfunction because the maximum score could not be achieved. Finally, unemployed or retired individuals could not check any of the items in the 'work' category and, therefore, would receive a score of '0' indicating no functional limitation. In addition to inaccurately reflecting individual dysfunction, group comparisons would be compromised (Pollard & Johnston, 2001).

- ***Direct Comparison of SIP vs SF-36***

  A systematic search of the instruments used to assess PROs from 1990 to 1999, ranked the SIP as the second most widely evaluated (the SF-36 was the most widely evaluated) (Garrett et al., 2002), however, it is not as widely used as the SF-36 for spine trauma research (Schoenfeld & Bono, 2011). Furthermore, although the SIP is responsive to back pain (see Blount et al., 2002), it is less responsive than the SF-36 (Lurie, 2000). The SIP measures more dimensions than the SF-36, however, its length is a major disadvantage as it limits its practicality and contributes to patient burden. Specifically, the average completion time for the 136-item SIP is 20 to 30 minutes – probably longer for

[11] Bergner et al. (1981) were the creators of the SIP. Therefore, this study (which is earlier than 2000) was included.

chronically ill people – compared to 10 minutes for the 36-item SF-36 (Blount et al., 2002; Pollard & Johnston, 2001). Nevertheless, the SIP was evaluated as easier to understand than the SF-36 (Lurie, 2000). A review of PRO instruments identified the SF-36 and the SIP as potential measures of general health status, however, the SF-36 was recommended (Blount et al., 2002). Nevertheless, as the SIP contains items on profound disability, it would be "useful in severely ill populations in which other measures may display substantial floor effects" [12] (Lurie, 2000, p. 3128).

(iii) Participation / Quality of Life (Return-to-Work)

Summary

The (few) dedicated instruments that measure return-to-work require further testing to demonstrate their effectiveness as only limited evidence is available on the psychometric properties of the Prolo scale, the WL-26 and the ORQ. Nevertheless, the Prolo scale has been widely used in clinical settings over the last 20 years. Two generic (and psychometrically-sound) instruments that include work-related questions and are frequently used to measure return-to-work are the SF-36 and the SIP. However, neither instrument is satisfactory for this purpose. Specifically, the SF-36 assesses aggregate role limitations (e.g., work, leisure and household roles) without specifying the work role while the SIP work-role questions do not accurately reflect individual dysfunction. Overall, further validation of work outcome measures is required to increase their interpretability (i.e., meaningfulness) to consumers (Amick et al., 2000).

Although job loss is the "most economically important consequence of spinal surgery" (Blount et al., 2002, p. 20), there are few validated instruments that measure return-to-work following spinal surgery (Amick, et al., 2000). Three reviews that assessed return-to-work instruments, and met inclusion criteria, were identified for the present report. Two reviews assessed the Prolo Scale - a widely used back-specific tool that assesses return-to-work following spinal surgery – while the third review discussed evidence for two instruments that measure work capacity for musculo-skeletal disorders: the WL-26 and the Occupational Role Questionnaire (ORQ).

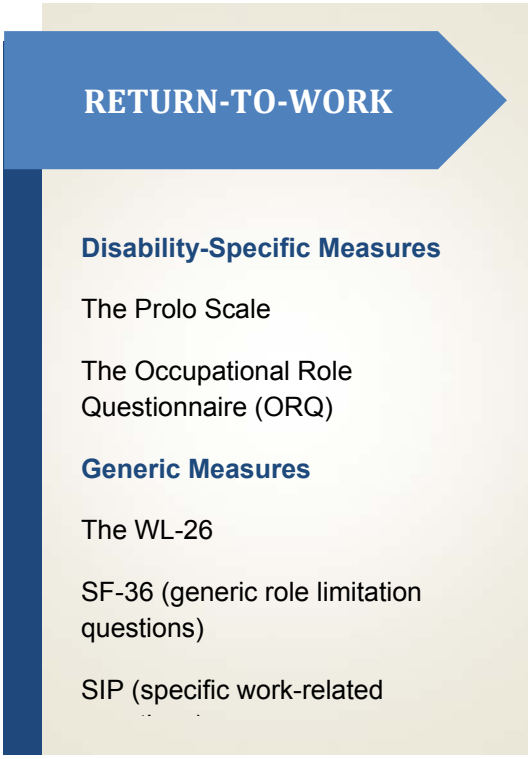

Figure 4: Return-to-Work measures (spinal)

[12] Floor (and ceiling) effects refer to the inability of an instrument to detect changes that occur at the low (or high) end of the construct (e.g., pain or activity level) being measured (Skolasky et al., 2007).

- **The Prolo Scale**[13]

There are four versions of the Prolo Scale (PS). The original PS, which was designed to assess lumbar spine surgery outcomes, comprised two self-report scales: an economic status scale and a functional status scale. Each scale contained five possible Likert responses (see Table 4). The functional status scale ranged from F1 (*total incapacity or worse than before operation*) to F5 (*complete recovery, no recurrent episodes of low back pain, able to perform all previous sport*) while the economic scale ranged from E1 (*completely invalid*) to E5 (*able to work at previous occupation with no restrictions of any kind*). The total score is obtained by adding the scores of each scale (range =1 to 5), resulting in a minimum score of '2' and a maximum score of '10'. Total scores are categorised as excellent (9-10), good (8-7), fair (6-5) and poor (4-2) (Vanti et al., 2013).

Table 4: The Prolo Scale **(Prolo et al., 1986)**

Economic status scale	
E_1	Completely invalid*
E_2	No gainful occupation including ability to do housework or continue retirement activities
E_3	Able to work but not at previous occupation
E_4	Working at previous occupation part time or limited status
E_5	Able to work at previous occupation with no restrictions of any kind
Functional status scale	
F_1	Total incapacity (or worse than before operation)
F_2	Mild-to-moderate level of low back pain and/or sciatica (or pain same as before operation but able to perform all daily tasks of living)
F_3	Low level of pain and able to perform all activities except sports
F_4	No pain but patient has had one or more recurrences of low back pain or sciatica
F_5	Complete recovery, no recurrent episodes of low back pain, able to perform all previous sport activities

Note. * Completely housebound

The PS has been frequently used to assess degenerative pathologies of the spine, and has sometimes been adapted for use in other spinal districts such as the thoracic and cervical spine. Although the PS has sometimes been used as the main (primary) outcome when assessing surgery outcomes, it is more commonly used in association with other measures (Vanti et al., 2013).

An earlier literature review – which included 27 studies – focused on identifying the most validated and frequently used tools to measure spinal fusion outcomes across a range of dimensions (e.g., general health, spine-specific disability, return to work) with a view to recommending a core set of the best available instruments (Blount et al., 2002). The

[13] The Prolo Scale is also known as the Prolo score, Prolo Economic Scale, Prolo Economic Functional Rating Scale, anatomic economic functional grading system or ''modified'' Prolo Scale (Vanti et al., 2013).

authors recommended the PS to measure the return-to-work dimension, as it was the "only available tool to quantify the area" (Vanti et al., 2013, p. 239). However, only the economic status scale of the PS was recommended, not the functional status scale. Furthermore, Schnee's 1997 modified version of the PS was recommended as it replaced the phrase 'at previous occupation' found in the original PS with the phrase 'working/active at previous level'. This adjustment ensured that it included all spinal patients irrespective of previous occupational status (Blount et al., 2002). Finally, Schnee's interpretation of scores - in which a score of E4 (*working/active at previous level with limitation*) or E5 (*working/active at previous level without restriction*) represented a successful return-to-work - was described as a 'reasonable interpretation' (Blount et al., 2002). In contrast, Blount et al. (2012) recommended other instruments to measure functional status, such as the Oswestry Disability Scale, rather than the functional status scale included in the original PS (Vanti et al., 2013).

A more recent systematic review – which included 64 studies – focused exclusively on the "application, interpretation and accuracy" of the PS (Vanti et al., 2013). In addition to noting the difficulty of comparing studies that use different versions of the PS to assess spinal surgery outcomes, the systematic review located only one study that examined the psychometric properties of the PS. However, this study only demonstrated that the PS is sensitive to changes in pre- and post-operative functioning – it did not provide any evidence for the reliability or validity of the PS. Despite the dearth of validation studies, the PS has been adapted for clinical examination for 20 years because the scale is a "simple and useful tool for standard evaluation of the efficacy of different surgical techniques in opposition to self-report measures" (Vanti et al., 2013, p. 237).

- **Lower Back Pain Measures: The ORQ and the WL26**

The third review identified for this report discussed two instruments that measure work capacity for musculo-skeletal disorders: the ORQ and the WL-26 (Amick et al., 2000). The authors state that, although these instruments were designed for upper extremity musculo-skeletal disorders, they can be applied to lower back pain.

The Occupational Role Questionnaire (ORQ) measures two aspects of working life: job performance and job satisfaction. Job performance includes work speed, work load and work breaks. Job satisfaction includes opportunities to upskill, job satisfaction and job retention (Amick et al., 2000). In addition to face validity (which increases user acceptance), the scales have been assessed as psychometrically sound (albeit by one study only) and they correlate moderately with pain (r = .037) and the Roland-Morris Disability Scale (r = .057). These two correlations indicate that pain and disability influence job performance and job satisfaction. Nevertheless, the ORQ has drawbacks – it is cognitively demanding as it requires "extensive recall and cognitive comparisons" and it also requires further testing to assess its generalisability to other injury groups (Amick et al., 2000).

The WL-26 conceptualises the work role as having five work demand categories, which are operationalised in the following scales: work scheduling, physical demands, mental demands, social demands and output demands (Amick et al., 2000). The scales have high reliability (ranging from .88 to .92) and high scaling success (i.e., they correlate

more highly with their hypothesised scale than other scales). The review authors note that evidence is emerging on the construct validity of the instrument (Amick et al., 2000). However, at the time of the review, there was no evidence on the clinical responsiveness of the WL-26, therefore, the clinical utility of the instrument is unknown.

SF-36 and SIP (Work-Related Questions). The SF-36 contains questions on role limitations, without specifying specific roles. Therefore, the questions reflect aggregate limitation across various roles including, but not limited to, work roles (Amick et al., 2000). The limitations of the SIP work questions have been briefly discussed elsewhere in this document (see p. 24).

3.2.3 *Multi-Dimensional Outcome Measurement (Spinal)*

Due to the complexity of spinal surgery (Street et al., 2012) and the need for comprehensive evaluation (Garrett et al., 2002), calls have been made for the multi-dimensional assessment of spinal surgery outcomes (Blount et al., 2002; Bombadier, 2000a; McCormick et al., 2013). The following section is divided into two parts. Part A discusses a set of core domains for spinal surgery outcomes and instruments to measure those core domains while Part B discusses the 'minimal acceptable outcomes' approach to assess spinal fusion success using multiple outcomes.

PART A: Core Domains and Instruments

Despite the lack of consensus on a standard set of PROs (Tevis & Kennedy, 2013) or the most reliable measurement tool (Vanti et al., 2013) - making it difficult to draw definite conclusions about the efficacy of different treatment strategies (Carreon et al., 2008) – two reviews collated 'best evidence' on spinal surgery outcome measurement to identify a 'core set' of domains and instruments. Although Bombadier (2000a) focused solely on patient-reported outcomes and Blount et al. (2002) recommended inclusion of both surgical and patient-reported outcomes, there was considerable agreement between both literature reviews on (a) the dimensions to be measured and (ii) the instruments to be used (Table 5).

Core Domains. Based on extensive literature reviews, an expert panel – with practical and methodologic expertise of PRO measures – recommended that PROs for spinal surgery should cover a core set of five domains:

(i) generic health status,
(ii) back specific function / disability,
(iii) pain,
(iv) work disability / return-to-work and
(v) patient satisfaction (Bombadier, 2000a, 2000b).

The second literature review of 27 articles – which focused specifically on spinal fusion – recommended inclusion of both subjective PRO measures and objective surgical measures. This recommendation was made because spinal fusion success rates that are based solely on subjective measures (e.g., pain levels and/or patient satisfaction) are substantially higher than success rates based on objective measures (e.g., return to work, mobility, and/or use of

analgesics) (Blount et al., 2002). Therefore, the following additional domains were recommended to be included to assess spinal fusion outcomes:

(vi) medication use,
(vii) complications and
(viii) fusion status (Blount et al., 2002).

Instruments to Measure Core Domains. Based upon analyses of the properties of existing instruments, both reviews recommended a 'toolkit' of measures to assess the above core dimensions (see Table 4). The use of multiple measures to assess spinal surgery outcomes is also congruent with the acknowledgement that some PRO measures are best used in conjunction with others (e.g., Von Korff, 2000; Ware, 2000). Inclusion of a generic health status instrument (i.e., the SF-36) in the 'toolkit' is recommended to address broader issues (e.g., co-morbidities) not considered by disability-specific instruments and to provide a more comprehensive picture of patients' health (Blount et al., 2002; Bombadier, 2000a). Inclusion of disease-specific instruments is recommended as these are typically more responsive than generic instruments (Lurie, 2000). Possibly due to the difficulty in defining patient satisfaction – and which dimensions to measure (e.g., expectation/satisfaction) – recommendations differed. Blount et al. (2002) recommended a subsection of the North American Spine Society (NASS) low back pain questionnaire (the Patient Satisfaction Index) while Bombadier (2000a) recommended assessment of satisfaction with the care received and treatment outcome.

A more recent systematic review of RCTs, although restricted to persistent low back pain, made similar recommendations (Chapman et al., 2011). Specifically, recommended domains were pain, function and quality of life. The recommended instruments for pain were the VAS and the NRS, for function the ODI and RDQ, and for quality of life the SF-36. Measurement of return-to-work and medication was *not* recommended as they are "complicated outcome measures", however, assessment of complications was recommended as a "standard of clinical practice". Practical issues, such as the burden on staff and patients, need to be considered when selecting multiple outcome measures (Chapman et al., 2011).

Part B: Minimum Acceptable Outcomes

Improvements in PROs (e.g., pain, function and quality of life) are typically expressed as statistically significant changes in scores on PRO measures (e.g., a 3-point decrease on a 10-point VAS pain scale). However, these results are usually reported at the group level and may not translate to significant changes at the individual level. Recently, composite scores derived from the minimal threshold for success across multiple domains – which commonly include pain, function, medication use and work status – have been suggested. However, the magnitude of change that is acceptable to individuals may vary (Carragee & Cheng, 2010).

Therefore, an alternative method of measuring surgical success in terms of '**minimum acceptable outcomes**' has been proposed. Patients individually identify the minimum outcome that they are willing to accept before they would consider undergoing a surgical procedure using standard metrics. Patient-identified minimum acceptable outcomes can then be compared at follow-up with those achieved from surgery to determine how frequently patients' goals were met. Patient-determined minimum acceptable outcomes can be used as

a “validity check against post hoc patient satisfaction or surgeon assessment scores” (Carragee & Cheng, 2010, p. 314).

A study of 165 consecutive patients who underwent lumbar fusion for either isthmic spondylolisthesis or disc degeneration reported a high level of improvement as the minimum acceptable outcome for spinal fusion (i.e., a decrease in pain intensity to 3/10 on the VAS, an improvement of 20 or more points on the ODI, discontinuing opioid medication and a partial return-to-work). Achievement of minimum acceptable outcomes was strongly associated with patient satisfaction at the 2-year follow-up. However, patients with abnormal psychometrics (i.e., presence of psychological distress and other psychosocial stressors) or with compensation claims reported satisfaction in the absence of minimum acceptable outcome achievement. As an aside, few patients considered the more commonly used metric of ‘minimal clinically important difference’ (MCID)[14] to be an acceptable spinal fusion outcome (Carragee & Cheng, 2010).

[14] The minimal clinically important difference (MCID) is ‘‘the smallest change that is important to patients’’ (see Copay et al., 2008, p. 969).

Table 5: Core Set of Instruments for Multi-Dimensional Measurement of Spinal Surgery Outcomes

SPINAL FUSION (Blount et al., 2002)		**SPINAL DISORDERS** (Bombadier, 2000a)	
Patient-Reported Outcomes	**Recommended Measure**	**Patient-Reported Outcomes**	**Recommended Measure**
General health status	SF-36 or SF-12	**Generic health status**	SF-36 (version 2)
Back specific disability	Oswestry Disability Questionnaire (ODI)	**Back specific function**	Oswestry Disability Questionnaire (ODI) Roland–Morris Disability Questionnaire
Pain level	(Visual) Analog Scale (VAS) (1-10) (for back or lower leg; or neck for cervical fusions)	**Pain**	SF-36 (bodily pain scale)
Return to work	Prolo Economic Scale	**Work disability**	Work Status (10 categories) *e.g., usual job, restricted duties, paid/unpaid sick leave, unemployed due to health / other reasons etc.*
Patient satisfaction	North American Spine Society Patient Satisfaction Index	**Patient satisfaction**	• Patient Satisfaction Scale (*to measure satisfaction with care*) • A global question (*to measure satisfaction with treatment outcome*)
Medication use	1. % of patients using narcotic medication, non-narcotic medication, no medication. 2. % of patients with significant reduction in medication use (>50%) measured post-operatively.	-	-
Surgical Outcomes	**Recommended Measure**	**Surgical Outcomes**	**Recommended Measure**
Fusion status	Radiographic assessment of: 1. Solid fusion rate 2. Nonunion rate 3. Levels fused	-	-
Complications	A generalised complication rate to include: • Percentage of patients with a complication • Breakdown of complications by number	-	-

3.3 IMPLANTABLE PAIN THERAPY / NEUROSTIMULATION

Summary

No articles met inclusion criteria for best available evidence for measuring PROs or surgical outcomes for implantable pain therapy (IPT) or neurostimulation. It was recommended that future studies use valid and reliable instruments to measure PROs, namely, pain, physical and psychosocial functioning and work status (Turner et al., 2004).

None of the articles identified in the systematic search met inclusion criteria for best available evidence for measuring surgical or patient-reported outcomes of implantable pain therapy (IPT) or neurostimulation. Therefore, a scan of eligible (but rejected) articles was undertaken to identify *common practice* for measuring outcomes. Results reported below represent common practice for measuring surgical and patient-reported outcomes for IPT and neurostimulation and **cannot be regarded as 'best evidence'** without further assessment.

3.3.1 Question 1- What is the most effective practice to measure surgical outcomes based on best available evidence?

Surgical Complications / Adverse Events

For the sake of clarity and to aid comparison, the following section will follow the same structure as the spinal fusion complications' section namely definition, classification and measurement.

Definition: No clear definition of complications or adverse events related to IPT or neurostimulation was evident. However, the systematic reviews, narrative reviews and studies that examined complications usually did so within the context of assessing the safety of IPT or neurostimulation and they were variously labelled as complications, adverse events, or adverse occurrences (Deer et al., 2012; Hayek et al., 2011; Smith et al., 2008; Turner et al., 2004).

Classification: No clear, well-ordered classification system was found for IPT or neurostimulation; most studies reported a catalogue of adverse events and complications. Articles commonly graded complications by severity, which was generally a binary system: minor *versus* severe/serious/life-threatening (Falco et al., 2009; Liem et al., 2013) or morbidity *versus* mortality (Deer et al., 2012). One systematic review of intrathecal infusion systems classified complications by their source. Thus, complications were divided into three broad groups: technical, biological or medication-related (Falco et al., 2009). A more detailed division of complications by source attributed complications to (i) implantation or management procedures, (ii) drug reactions or side effects, (iii) device malfunction and, (iv) human error in programming or refilling the device (Deer et al., 2012). More commonly, however, articles made the simple distinction between device-related complications and other complications for both IPT and neurostimulation, which were frequently grouped by severity (major vs. minor) (Buchser et al., 2007; Liem et al., 2013).

Measurement: A range of factors influence reporting of adverse events and complications. For example, the reported incidence of device-related complications was higher in earlier publications. The lower incidence in more recent publications is likely due to "improved education, techniques and experience in the implanting community" (Hayek et al., 2011, p. 238). Other factors include the heterogeneity of studies, differences in equipment, length of follow-up (Turner et al., 2004), national differences in reporting and, for mortalities, difficulty in determining cause of death (Deer et al., 2012). Consistent with the recommendations of Blount et al. (2002), the incidence of adverse event/complications were often reported as the percentage of patients with a complication and a breakdown of complications by number (Liem et al., 2012; Hayek et al., 2011; Turner et al., 2004).

3.3.2 Question 2- What is the most effective practice to measure PRO resulting from surgery based on the best available evidence?

As IPT and neurostimulation are an 'advanced stage therapy' to manage persistent, intractable pain (Patel, 2009), pain was the most frequently measured PRO. Reduction in opioid intake – although an objective pain measure (Blount et al., 2002) – was a secondary outcome. Other secondary outcomes included functional status, psychological status and return to work, however, few were discussed. It was recommended that future studies measure PROSs (e.g., pain, physical and psychosocial functioning and work stat us)with valid and reliable instruments (Turner et al., 2004). The following discussion relates to non-cancer pain.

(i) Impairments / Symptoms – Pain

The most frequently reported pain instrument was the generic Visual Analog Scale (VAS), which measures pain intensity (i.e., the severity or magnitude of pain (Von Korff, 2000)). Other (less) frequently recurring instruments mentioned in systematic reviews and one review were the ODI, NRS, the McGill Pain Questionnaire and the Brief Pain Inventory (BPI)[15] (Falco et al., 2013; Hayek et al., 2001; Patel et al., 2009; Smith et al., 2008; Turner et al., 2004). Four of the systematic reviews reported pain as 'pain relief' which was divided into short-term pain relief (less than one year) and long-term pain relief (more than one year). A fifth systematic review compared acute pain (≤30 days) with persistent pain (≤1 year) (Abou-Setta et al., 2011).

According to Von Korff (2000), the experience of pain is multi-dimensional. As some of the instruments mentioned above measure different facets of pain, direct comparisons among studies would be difficult. For example, use of the VAS to measure pain at different intervals suggests that a reduction in pain intensity is the primary aim of therapy, use of the ODI suggests that a reduction in pain severity (i.e., pain intensity *and* interference with daily activities) is the main goal of therapy while use of the McGill Pain Questionnaire suggests

[15] The psychometric properties of a number of these instruments are discussed in the section entitled, Patient-Reported Outcomes (Spinal) (see pp. 9-18).

that a reduction in pain affect (i.e., emotional arousal and disruption caused by the pain experience (Von Korff, 2000)) is the primary goal. Few instruments specifically measure pain persistence although evidence suggests that pain intensity (how much it hurts) and pain persistence (how long it hurts) are more usefully viewed as "independent facets of pain status" (Von Korff, 2000, p. 3140). Nevertheless, a systematic review of pain intensity following traumatic musculo-skeletal injury reported that the VAS (which measures pain intensity) was the most frequently used (Rosenbloom, 2013).

(ii) Activity / Disability – Functional Status

Articles rarely discussed disability or functional status. Nevertheless, some of the systematic reviews included a table that listed the instruments used in studies that assessed PROs of IPT and neurostimulation, specifically, the ODI and the SF-36 (Falco et al., 2009; Patel et al., 2009). The Sickness Impact Profile was mentioned in a very few studies.

(iii) Participation / Quality of Life – Return-to-Work

Although return to work was cited as a secondary outcome of interest in multiple systematic reviews of IPT and neurostimulation (Hayek et al., 2011; Patel et al., 2009; Falco et al., 2009), only one explicitly discussed this topic (Turner et al., 2004). However, no instruments for measuring return to work were mentioned. Turner et al. (2004) noted that no articles systematically reported pre- and post-treatment work status or post-treatment return-to-work rates.

4. CONSIDERATIONS AND RECOMMENDATIONS

Three critical intervention goals for spinal pathology are to "improve the patient's quality of life, restore function, and relieve pain" (McCormick, Werner & Shimer, 2013, p. 99). Spinal researchers have been leaders in the field of PRO measurement and have produced a prolific number of instruments. Unfortunately, the psychometric (measurement) properties of many of these instruments have yet to be rigorously assessed and gold standards have yet to be established (Carreon et al., 2008; Hagg et al., 2001; Niskanen, 2002; Zanoli et al., 2000). Consequently, a great deal of literature was reviewed to assess 'best evidence' for measuring surgical and patient-reported outcomes following spinal surgery. However, 'best evidence' for measuring outcomes of IPT and neurostimulation therapies has not yet been established.

Overall, the present snapshot review provides an expedited view of the best available evidence in the literature and is a first step to identifying the full spectrum of evidence. Useful guidance is available in the literature to commence a discussion with clinicians/surgeons and patients about measuring outcomes for spinal fusion. However, guidance on IPT and neurostimulation is limited to a brief summary of common practice (vs. best evidence). Use of standardised approaches and validated instruments to measure the outcomes of spinal surgery will enable comparison among groups and treatment methods (Glassman et al., 2006; Schoenfeld & Bono, 2011).

A summary of findings and recommendations across surgical and PROs is provided below.

4.1 Measurement of surgical outcomes and patient-reported outcomes

4.1.1 As the patient is the principal source of information for subjectively-measured client outcomes (Hagg et al., 2001), self-report questionnaires were overwhelmingly used to assess patient-reported outcomes (PROs). In contrast, surgical outcome measurement typically relied on objective criteria such as imaging (for fusion status) and surgical complications.

4.1.2 The quality of instruments that assess specific PROs is determined by three key measurement properties: *reliability* (i.e., freedom from measurement error); *validity* (the instrument measures what it is supposed to measure) and *responsiveness* (ability to detect change over time). These three qualities need to be established prior to selecting an instrument for use in clinical settings (Chapman et al., 2011).

4.1.3 Selection of a specific instrument to measure pain will be influenced by which facet of pain is to be assessed. Pain severity (i.e., pain intensity and pain-related interference with daily activities) is most commonly measured. Pain intensity instruments are also used to measure persistent pain.

4.1.4 Although many disease-specific measures of disability (functionality) exist, few have been subjected to widespread, rigorous testing. Furthermore, the two most widely tested and used disease-specific instruments (the ODI and the RDQ) measure pain-related disability versus disability due to other causes. Use of generic instruments – although less responsive than disease-specific instruments – enables comparison of the impact of treatment and different injury groups.

4.1.5 Despite its economic importance, few validated instruments specifically measure return-to-work following spinal surgery (or IPT / neurostimulation).

4.1.6 Reporting complications is considered essential for assessing spinal fusion outcomes. Some (limited) guidance is available in the literature: a generalised complication rate that includes the percentage of people with complications and a breakdown of complications by type is recommended (Blount et al., 2002).

4.2 Multi-dimensional measurement of surgical outcomes and PROs

4.2.1 Due to the complexity of spinal surgery and the need for comprehensive assessment, multi-dimensional measurement of (i) PROs or (ii) PROs *and* surgical outcomes has been suggested.

4.2.2 Two reviews (one of which included an expert panel) collated best evidence on spinal surgery outcomes and identified a parsimonious core set of domains to be measured, as well as recommending a 'toolkit' of instruments/techniques to measure them (**see Table 4 for details**).

4.2.3 For PRO measurement, five core domains identified as the most useful and practical in a clinical setting were: generic health status; back-specific disability; pain; patient satisfaction; and return-to-work.

4.2.4 For the combined assessment of PROs *and* surgical outcomes, an extended list of domains included the five mentioned above as well as: medication use; fusion status and complications.

4.3 Review mode

A full systematic review is strongly indicated for the following reasons:

- ✓ To access and thoroughly assess all bodies of relevant evidence (due to the longer timeframe associated with a systematic (vs. snapshot) literature review mode).
- ✓ To enable a careful analysis of the most appropriate PRO instruments for specific spine pathologies (e.g., the Neck Disability Index (NDI) for cervical spine).
- ✓ To allow a thorough analysis of the very large literature on pain, as this particular PRO is critical for spinal surgery and is the focus of IPT / neurostimulation.

5. ADDENDUM

5.1 RATIONALE & PURPOSE

The HDSG indicated that they were interested in promoting the use of ePPOC (the electronic Persistent Pain Outcome Collaboration) to surgeons to allow national benchmarking for the measurement of outcomes of spinal surgery and IPT. Therefore, HDSG expressed an interest in information about the BPI (used by ePPOC to measure pain) as well as a patient Global Impression of Change Instrument (PGIC) and the SF-12.

The purpose of the Addendum is to provide the HDSG with a brief overview of information on the BPI, SF-12 and the PGIC. Three brief searches were conducted to identify information on the validation of these three instruments and their use for spinal disorders. The searches were not intended to be systematic or exhaustive. Results are reported as an Addendum to the present Snapshot Evidence Review.

5.2 METHOD

Google Scholar was searched using the following terms:

1. Brief Pain Inventory, spine surgery, spinal surgery, validation [16]
2. SF-12, spine surgery, spinal surgery, validation
3. Global Impression of Change, spine surgery, spinal surgery, validation

The first ten pages of results for each of the three searches were scanned for relevancy (i.e., a total of 30 pages). As few relevant records were obtained for the Brief Pain Inventory and the Global Impression of Change instruments, another search was performed using the name of each instrument only. The first five pages of results for each instrument were scanned for relevancy (i.e., a total of ten pages). Overall, a total of 21 full-text articles and reports were retrieved for the three instruments. Of these, 16 (15 articles and 1 report) were retained as relevant and are discussed below (see Table 6 for a breakdown).

Table 6: Records Included in the Addendum

Instrument	Results
Brief Pain Inventory (BPI)	Three peer-reviewed articles and one report [a]
12-Item Short Form Health Survey (SF-12)	Five peer-reviewed articles
Global Impression of Change (GIC)	Two peer-reviewed articles [b]
Multiple (*i.e., articles reviewed multiple instruments including the three of interest*)	Five peer-reviewed articles

Note. [a] The report was the BPI Users' Guide (Cleeland, 2009). [b] Although these two articles did not address the psychometric properties of the GIC, they were included as they provided information using a different methodological framework (i.e., the individual patient's perspective).

[16] The term, 'validation' was used in preference to 'validity' and 'reliability' as validation refers to the process of assessing the psychometric properties of instruments.

5.3 RESULTS

5.3.1 The Brief Pain Inventory (BPI)

Description. The BPI was originally developed to measure cancer-related pain, although it has since been used for other conditions including, low back pain, chronic back pain and spinal disorders (Atkinson et al., 2010; Keller et al., 2004; see also Cleeland, 2009). Consistent with the multi-dimensional nature of pain, the BPI focuses on two components of pain: sensory pain and reactive pain (Atkinson et al., 2010; Cleeland, 2009). The sensory pain dimension is measured as 'pain intensity' (current, worst, least and average pain) using an 11-point NRS which ranges from '0' *no pain* to '10' *pain as bad as you can imagine*. The reactive pain dimension is measured as 'interference with everyday function' (i.e., general activity, mood, walking ability, normal work, relations with other people, sleep and enjoyment of life) and is also assessed with an 11-point NRS which ranges from '0' *does not interfere* to '10' *completely interferes* (Atkinson et al., 2010).

Psychometric properties of the BPI. Factor analysis has confirmed the two-factor structure of the BPI, that is, 'pain intensity' and 'interference with daily function' (Cleeland, 2009; Jensen, 2003). The two-factor structure was replicated in a sample that included patients with low back pain (Keller et al., 2004). Factor analysis also revealed that the 'interference with daily function' factor comprises two sub-components: *activity* and *affect* (emotions). *Activity* items include 'walking', 'general activity', 'working', and 'sleep' 'while *affect* items include 'enjoyment of life', 'relations with others' and 'mood' (Atkinson et al., 2010). Although some clinical trials use only one of the items from the pain intensity factor (e.g., 'worst' or 'average' pain), the test developers recommend use of all four pain intensity items (Cleeland, 2009). Use of all four items (current, worst, least and average pain) more effectively captures the different components of pain intensity, thus increasing the content validity of this factor of the BPI (Jensen, 2003). Multiple studies have demonstrated the reliability (internal consistency) [17] of the BPI (Cleeland, 2009), with the 'pain at its worst in the last 24-hours' item being the most stable (reliable) (Atkinson et al., 2010). The reliability of the BPI has also been demonstrated for a sample that included patients with low back pain (Keller et al., 2004).

Use of the BPI for spinal disorders. Two studies (Keller et al., 2004; Tan et al., 2004) were located that examined the BPI for samples that included patients with low back pain. Keller et al. (2004) formally assessed the reliability and validity of the BPI for non-cancer patients (i.e., those with arthritis or low back pain with and without workers' compensation) "under clinical conditions that were as close as possible to normal practice" (p. 311). For lower back pain patients, the BPI was found to be reliable (i.e., possess high internal consistency) and the two-factor structure of the BPI was confirmed. The construct validity of the BPI was demonstrated by the high correlations with the Roland Morris Disability Questionnaire (a disease-specific measure of back pain-related disability) and the bodily pain scale of the SF-36 (this scale provides a generic measure of pain). Finally, the BPI also demonstrated sensitivity to changes in disability. Overall, the results support the use of the BPI for patients

[17] As an aside, the test-retest form of reliability may be inappropriate for instruments that measure pain as this form of reliability assumes stability over time in the factor being measured and pain levels may fluctuate over time (Jensen, 2003).

with low back pain. However, the authors note that their sample was not representative as it was necessary to select patients whose back pain was likely to change to enable an assessment of the sensitivity of the BPI (Keller et al., 2004). Similarly, the second study (Tan et al., 2004) confirmed the two-factor structure of the BPI for patients with a primary diagnosis of back pain or pain at multiple sites (including back pain). The BPI was found to be reliable (high internal consistency), valid (highly correlated with the Roland Morris Disability Questionnaire) and responsive (the two BPI scales showed statistically significant improvement following treatment).

5.3.2 The 12-Item Short-Form Health Survey (SF-12)

Description. Due to its length, the 36-Item Medical Outcomes Study Short-Form Health Survey (SF-36) may be too burdensome and costly for large-scale health measurement and monitoring studies, particularly those that include multiple PROs (Luo et al., 2003; Ware et al., 1996). To address this issue, the 12-Item Short-Form Health Survey (SF-12) was derived from the longer SF-36.

Item selection. The challenge in constructing a short-form measure is striking a balance between brevity (i.e., the number of items) and other critical factors, such as comprehensiveness and the statistical precision of scores (Ware et al., 1996). Based on regression analyses, the best subset of 12 items that reproduced the two summary scales of the SF-36 were included in the SF-12, namely the Physical Component Scale (PCS) and the Mental Component Scale (MCS). At least one item was selected from each of the eight sub-scales of the SF-36 (four of these sub-scales are subsumed under the PCS and four are subsumed under the MCS). However, two items were selected for the two sub-scales that best predict physical health (Role Physical, Physical Functioning) and the two subscales that best predict mental health (Role Emotional, Mental Health). Thus, the 12 items included in the SF-12 provide a representative sampling of the content of the eight sub-scales of the SF-36 (Ware et al., 1996).

Psychometric properties of the SF-12. The SF-12 is entirely a subset of the SF-36 health survey (Ware et al., 1996). Therefore, in the following studies, the psychometric properties of the SF-12 were examined by extracting the relevant items from data already obtained for the SF-36 and re-analysing it. The test developers used SF-36 data from two sources: (i) the U.S. National Survey of Functional Health Status (NSFHS), a cross-sectional survey used to gather norms for the SF-36 Health Survey, and (ii) the Medical Outcomes Study (MOS), an observational study of adult patients with chronic conditions (Ware et al., 1996). Overall, the SF-12 was found to be reliable and valid, although less so than the SF-36. An Australian study cross-validated the items included in the standard form of the SF-12 using data obtained for the SF-36 by the 1995 Australian National Health Survey and concluded that the SF-12 was suitable for use in an Australian context.

Use of the SF-12 for spinal disorders. Three studies were located that examined the psychometric properties of the SF-12 for use with spinal disorders (Lee et al., 2008; Luo et al., 2003; Singh et al., 2006). Among patients with back pain, the SF-12 was found to be reliable, valid and responsive (Luo et al., 2003). However, the authors state that a limitation of this study was that it did not also examine the psychometric properties of the SF-36.

Therefore, no recommendations as to whether the SF-12 or the SF-36 is more appropriate for use in back pain patients could be made (Luo et al., 2003).

A later study (Singh et al., 2006) partially addressed this limitation by comparing the psychometric properties of the SF-12 with the SF-36 in patients with cervical spondylotic myelopathy who underwent decompression surgery. Analyses focused on a comparison of the PCS and MCS scales for each instrument. Reliability was considerably lower for the PCS and MCS scales of the SF-12 compared to the SF-36, however, reliability coefficients were still well above the recommended threshold of .70. The drop in reliability was most likely to the inclusion of fewer items in the SF-12, however, this was described as an 'acceptable trade-off' in terms of the practicality of the SF-12 which can be completed in 2 minutes. The validity of the SF-12 was demonstrated by the extremely high correlations between the PCS component of the SF-12 and the SF-36 and the MCS component of the SF-12 and the SF-36. Both instruments were responsive (measured as the change in pre- and post-surgery scores), although the SF-12 was less responsive to change than the SF-36. A major caveat associated with the SF-12 is that, due to the inclusion of fewer items, the SF-12 cannot be "reliably reported in terms of the eight domains, but only as PCS and MCS summaries" (Singh et al., 2006, p. 642).

The third study (Lee et al., 2008) partially addressed the above caveat by comparing the validity of version 2 of the SF-12 and the SF-36 in patients undergoing elective spinal surgery (cervical and lumbosacral). The SF-12 v2 incorporated changes to the wording of items and an increased range of responses, which "minimizes the ceiling and flooring effects, thus allowing for the scoring of the 8 scales in addition to the 2 summary scores" (Lee et al., 2008, p. 829). For both spinal groups, strong correlations were found between the SF-36 v2 and SF-12 v2 for the eight sub-scales and the two summary scores (PCS and MCS), indicating that they are valid alternatives in patients with spinal disorders. Availability of the more detailed information in the eight sub-scales in the SF-12v2 enables clinicians make a "detailed yet efficient and valid assessment of individual health profiles" (see Lee et al., 2008, p. 832). However, a limitation of this study was that the SF-12v2 and the SF-36v2 were not administered separately, thus, the influence of survey length on the outcome could not be determined (Lee et al., 2008).

Use of the SF-12 vs the SF-36

Although the SF-12 captures less health-related information than the SF-36, it is comparable to the SF-36 in terms of reliability, validity and responsiveness for spinal disorders (Singh et al., 2006). Therefore, the choice of which form is most suitable will depend on its intended use. Due to its brevity, the SF-12 is 'most justified' for large-scale studies, particularly when multiple measures are used, or in busy clinical practices, as it takes two minutes to complete (Luo et al., 2003; Lee et al., 2008; Ware et al., 1996). However, due to the inclusion of fewer items, the health information obtained from the SF-12 can only be reliably reported as summaries of the PCS and the MCS (Singh et al., 2006). In contrast, due to the inclusion of more items, the SF-36 'defines more levels of health' and the more detailed information captured by the PCS and MCS summary scales, and particularly from the eight sub-scales, can provide 'more reliable estimates of individual levels of health' (Ware et al., 1996, p. 231). Therefore, the SF-36 may be more advantageous than the SF-12 in smaller studies (Ware et al., 1996).

5.3.3 The Patient Global Impression of Change (PGIC)

Description of the GIC. The PGIC instrument is a global self-report scale that assesses patients' own impressions of change. Scale descriptors commonly range from *much better* through to *no change* to *much worse* (see Hurst & Bolton, 2004). Problematically, the PGIC used in the two companion studies included for review differed: one study (Hurst & Bolton, 2004) described the PGIC as a self-report 7-point NRS whereas the other companion study (Bolton, 2004) described the GIC as an 11-point NRS.

Rationale for including the two companion studies. Neither study examined the psychometric properties of the PGIC for assessing spinal surgery outcomes. Instead, the PGIC (which assesses the individual's perception of change) was used to examine the sensitivity of common statistical methods for evaluating clinically important changes in the treatment of back and neck pain. These two studies were included as they evaluate change from the individual patient's perspective (versus group averages and the statistical significance of their differences). For conditions with no directly measurable end point (e.g., pain conditions), assessment of patient experiences and whatever represents a meaningful improvement for the individual patient is 'pivotal' (Hurst & Bolton, 2004, p. 27).

Psychometric properties of the GIC. Hurst and Bolton (2004) note that the reliability and validity of the modified PGIC has not been assessed, and although they used it as a valid external criterion for measuring clinical significant change, this aspect also has not been assessed. Nevertheless, in the absence of a 'true gold standard', use of the GIC as an external criterion was deemed 'conceptually reasonable' and 'clinically relevant' (p. 34).

Use of the GIC for spinal disorders. Hurst and Bolton (2004) assessed the clinical significance of change scores on the Bournemouth Questionnaire (a multi-dimensional questionnaire validated for use in neck and back pain) following treatment for neck or back pain. Using a cut-off score that included '6' *better* or '7' a *great deal better* on the PCIG to denote clinical significant improvement, corresponding cut-off scores for three common statistical methods (effect size, raw change and percentage change scores, and reliable change index scores) were identified to distinguish between patients who did and did not improve following treatment for neck and back pain. This methodology provides an alternative framework for identifying clinically significant change in patients which, in the past, has "relied almost exclusively on the statistical significance of changes in scores" for PRO to interpret treatment effectiveness (Hurst & Bolton, 2004, p. 26). Nevertheless, the corresponding cut-off scores for the common statistical methods analysed above were for the Bournemouth Questionnaire. Therefore, further research is required to identify (statistical) cut-off scores on other PRO measures (Hurst & Bolton, 2004).

5.3.4 Reviews of Multiple Instruments

Five reviews of multiple instruments that mentioned any of the three measures of interest to this addendum were identified (Alexander et al., 2009; Carey & Mielenz, 2007; Chapman et al., 2011; Jensen, 2003; Maughan & Lewis, 2010). To avoid repetition, two reviews are briefly mentioned below: the first recommended specific instruments while the second catalogued commonly used instruments. The spinal cord outcomes partnership endeavor

(SCOPE) reviewed a range of tools to assess spinal cord injury (SCI) outcomes for clinical research studies. Recommended tools were assessed for their metric properties (reliability, validity, sensitivity and accuracy). A number of tools were recommended to measure different aspects of pain. Among others, the 7 Point Guy/Farrar Patient Global Impression of Change scale to measure global changes in pain and the Brief Pain Inventory to measure pain interference were recommended. Similarly, a range of measures were recommended to measure quality of life including the SF-36 and the SF-12 (Alexander et al., 2009). A narrative review (Carey & Mielenz, 2007) noted that, although there is no single or core set of instruments to measure back pain, pain is commonly measured using a numerical, verbal or visual scale (i.e., NRS, VRS or VAS). However, the BPI was also mentioned as a related measure and was described as a "hybrid between a pure pain score and a functional assessment" (Carey & Mielenz, 2007, p. S11). Although the SF-36 and the SF-12 were commonly used to measure quality of life, it was noted that the physical functioning scale of the SF-12 "does not sufficiently predict the SF-36 scores for individual patients with low back pain" (see Carey & Mielenz, 2007, p. S11).

5.4 SUMMARY AND RECOMMENDATIONS

The information presented in the Addendum is based on an abbreviated search of the literature and is intended to provide a brief overview of the BPI, the SF-12 and the PCIG. The BPI is a generic instrument that measures two components of pain: 'intensity' and 'interference with daily function'. Although originally developed for use in cancer patients, the BPI has been found to be reliable, valid and responsive for low back pain patients. The SF-12 was derived from the longer SF-36 and was developed for use in large scale studies in which time and money constraints may preclude the use of the SF-36. The SF-12 has been found to be reliable, valid and responsive for spinal disorders, although less so than the SF-36 due to the inclusion of fewer items. The PGIC has not been validated. However, this scale measures clinically meaningful improvement from the patient's perspective, which represents an alternative framework to the more common method of interpreting the statistical significance of changes in scores on PRO to assess treatment effectiveness.

[See Appendix B for the reference list for the Addendum.]

6. REFERENCES

Abou-Setta, A. M., Beaupre, L. A., Rashig, S., Dryden, D. M., Hamm, M. P., Sadowski, C. A., Menon, M. R. G…Jones, A. (2011). Comparative Effectiveness of Pain Management Interventions for Hip Fracture: A Systematic Review. *Annals of Internal Medicine, 155*, 234-245.

Amick, B. C., Lerner, D., Rogers, W, H., Rooney, T., & Katz, J. N. (2000). A review of health-related work outcome measures and their uses, and recommended measures *Spine, 25*, 3152–3160.

† Anderson, C., Christensen, F, B., & Bunger, C. (2006). Evaluation of a Dallas Pain Questionnaire classification in relation to outcome in lumbar spinal fusion. *European Spine Journal, 15,* 1671–1685. DOI 10.1007/s00586-005-0046-z

Antonacci, A. C., Lam, S., Lavarias, V., Homel, P., & Eavey, R. D. (2008). A morbidity and mortality conference-based classification system for adverse events: Surgical outcome analysis: Part I. *Journal of Surgical Research, 147*, 172–177.

Antonacci, A. C., Lam, S., Lavarias, V., Homel, P., & Eavey, R. D. (2009). A report card system using error profile analysis and concurrent morbidity and mortality review: surgical outcome analysis, Part II. *Journal of Surgical Research, 153,* 95–104.

Beaton, D. E. (2000). Understanding the relevance of measured change through studies of responsiveness. *Spine, 25*, 3192–3199.

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: Development and final revision of a health status measure. *Medical Care, 19*, 787-805.

Blount, K. J., Krompinger, W. J., Maljanian, R., & Browner, B. D. (2002). Moving toward a standard for spinal fusion outcomes assessment. *Journal of Spinal Disorders & Techniques, 15*, 16–23.

Bombadier, C. (2000a). Outcome assessments in the evaluation of treatment of spinal disorders: Summary and general recommendations. *Spine, 25*, 3100–3103.

Bombadier, C. (2000b). *Spine* Focus Issue Introduction: Outcome assessments in the evaluation of treatment of spinal disorders: *Spine, 25*, 3097–3099.

† Bremerich, F. H., Grob, D., Dvorak, J., & Mannion, A. F. (2008). The Neck Pain and Disability Scale: Cross-cultural adaptation into German and evaluation of its psychometric properties in chronic neck pain and C1–2 fusion patients. *Spine, 33*, 1018–1027.

Buchser, E., Durrer, A., & Foletti, A. (2007). Neurostimulation technology for the treatment of chronic pain: a focus on spinal cord stimulation. *Expert Review of Medical Devices, 4.2*, 201.

Carragee, E. J., & Cheng, I. (2010). Minimum acceptable outcomes after lumbar spinal fusion. *The Spine Journal, 10,* 313-320.

Carreon, L. Y., Glassman, S. D., & Howard, J. (2008). Fusion and nonsurgical treatment for symptomatic lumbar degenerative disease: a systematic review of Oswestry Disability Index and MOS Short Form-36 outcomes. *The Spine Journal, 8,* 747–755.

Carreon, L. Y., Glassman, S. D., McDonough, C. M., Ranpersaud, R., Berven, S., & Shainline, M. (2009). Predicting SF-6D utility scores from the Oswestry Disability Index and Numeric Rating Scales for back and leg pain. *Spine, 34,* 2085–2089.

Chapman, J, R., Norvell, D. C., Hermsmeyer, J. T., Bransford, R. J., DeVine, J., McGirt, M. J., & Lee, M. J. (2011). Evaluating common outcomes for measuring treatment success for chronic low back pain*. Spine, 36*, S54–S68.

Copay, A. G., Glassman, S. D., Suach, B. R., Berven, S., Schuler,T. C., & Carreon, L. Y. (2008). Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and Pain Scales. *The Spine Journal, 8,* 968–974.

Deer, T. R., Levy, R., Prager, J., Buchser, E., Burton, A., Caraway, D., Cousins, M…Mekhail, N. (2012). Polyanalgesic Consensus Conference—2012: Recommendations to reduce morbidity and mortality in intrathecal drug delivery in the treatment of chronic pain. *Neuromodulation: Technology at the Neural Interface, 15,* 467–482.

DeVine, J., Norvell, D. C., Ecker, E., Fourney, D. R., Vaccaro, A., Wang, J., & Andersson, G. (2011). Evaluating the correlation and responsiveness of patient-reported pain with function and quality-of-life outcomes after spine surgery. *Spine, 36*, S69–S74.

Dindo, D., Demartines, N., & Clavien, P-A. (2004). Classification of surgical complications: A new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Annals of Surgery, 240*, 205–213.

Fairbank, J. C. T. (2007). Use and abuse of Oswestry Disability Index. *Spine, 32,* 2787–2789.

Fairbank, J. C. T., & Pynsent, P. B. (2000). The Oswestry Disability Index. *Spine, 25,* 2940–2953.

Falco, F. J., Patel, V. B., Hayek, S. M., Deer, T. R., Geffert, S., Zhu, J., Onyewu, O…& Manchikanti, L. (2013). Intrathecal infusion systems for long-term management of chronic non-cancer pain: an update of assessment of evidence. *Pain Physician, 16*, SE185-216.

Frie, K. G., van der Meulen, J., & Black, N. (2012). Single item on patients' satisfaction with condition provided additional insight into impact of surgery. *Journal of Clinical Epidemiology, 65,* 619-626.

Garrett, A., Schmidt, L., Mackintosh, A., & Fitzpatrick, R. (2002). Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ, 324*, 1417-1422.

Glassman, S. D., Carreon, L. Y., Djurasovic, M., Dimar, J. R., Johnson, J. R., Puno, R. M. & Campbell, M. J. (2009). Lumbar fusion outcomes stratified by specific diagnostic indication. *The Spine Journal, 9*, 13–21.

Glassman, S., Gornet, M. F., Branch, C., Polly, Jnr, D., Peloza, J., Schwender, J. D., & Carreron, L. (2006). MOS Short Form 36 and Oswestry Disability Index outcomes in lumbar fusion: A multicenter experience. *The Spine Journal, 6*, 21–26.

† Godil, S. S., Parker, S. L., Zuckerman, S. L., Mendenhall, S. K., & McGirt, M. J. (2013). Accurately measuring the quality and effectiveness of cervical spine surgery in registry efforts: determining the most valid and responsive instruments. *The Spine Journal, 14,* 2885-2891.

Hagg, O., Fritzell, P., Oden, A., & Nordwall, A. (2001). Simplifying outcome measurement. *Spine, 27,* 1213–1222.

Hayek, S. M., Deer, T. R., Pope, J. E., Panchal, S. J., & Patel, V. B. (2011). Intrathecal therapy for cancer and non-cancer pain. *Pain Physician. 14*, 219-248.

Kopec, J. A. (2000). Measuring functional outcomes in persons with back pain: A review of back-specific questionnaires. *Spine, 25,* 3110–3114.

Lebude, B., Yadla, S., Albert, T., Anderson, D. G., Harrop, J. S., Hilibrand, A., Maltenfort, M....Ratliff, J. K. (2010). Defining "complications" in spine surgery: neurosurgery and orthopedic spine surgeons' survey. *Journal of Spinal Disorders & Techniques, 23,* 493-500.

Li, H. L., & Dai, L-Y. (2011). A systematic review of complications in cervical spine surgery for ossification of the posterior longitudinal ligament. *The Spine Journal, 11,* 1049-1057.

Liem, L., Russo, M., Huygen, F. J. M., Van Buyten, J-P., Smet, I., Verrills, P., Cousins, M…Levy, R. (2013). A multicenter, prospective trial to assess the safety and performance of the spinal modulation dorsal root ganglion neurostimulator system in the treatment of chronic pain. *Neuromodulation: Technology at the Neural Interface, 16,* 471–482.

Lurie, J. (2000). A review of generic health status measures in patients with low back pain. *Spine, 25*, 3125–3129.

Mazeh, H., Cohen, O., Mizrahi, I., Hamburger, T., Stojadinovic, A., Abu-Wasel, B., Alaivan, B…Nissan, A. (2014). Prospective validation of a surgical complications grading system in a cohort of 2114 patients. *Journal of Surgical Research, 188,* 30–36.

McCormick, J. D., Werner, B. C., & Shimer, A. L. (2013). Patient-reported outcome measures in spine surgery. *Journal of the American Academy of Orthopaedic Surgeons, 21,* 99–107.

Mirza,S. K., Deyoi,R. A., Heagerty, P. J.,Turner, J. A., Lee, L. A., & Goodkin, R. (2006). Towards standardized measurement of adverse events in spine surgery: Conceptual model and pilot evaluation. *BMC Musculoskeletal Disorders, 7,* 53. doi:10.1186/1471-2474-7-53

Mokink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*, 737-745.

Nasser, R., Yadla, S., Maltenfort, M. G., Harr, J. S., Anderson, D. G., Vaccaro, A. R., Sharan, A. D., & Ratliff, J. K. (2010). Complications in spine surgery. *Journal of Neurosurgery: Spine, 13*, 144–157.

Nemeth, G. (2006). Health related quality of life outcome instruments. *European Spine Journal, 15*, S44-S51.

Niskanen, R. O. (2002). The Oswestry low back pain disability questionnaire a two-year follow-up of spine surgery patients. *Scandinavian Journal of Surgery, 91*, 208 – 211.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Ohnmeiss, D. D., Bodemer, W., & Zigler, J. E. (2010). Effect of adverse events on low back surgery outcome: Twenty-four-month follow-up results from a Food and Drug Administration investigational device exemption trial. *Spine, 35,* 835–838.

Patel, V. B., Manchikanti, L., Singh, V., Schultz, D. M., Hayek, S. M., & Smith, H. S. (2009). Systematic review of intrathecal infusion systems for long-term management of chronic non-cancer pain. *Pain Physician, 12*, 345-360.

Pollard, B., & Johnston, M. (2001). Problems with the Sickness Impact Profile: A theoretically based analysis and a proposal for a new method of implementation and scoring. *Social Science and Medicine, 52*, 921-934.

Proietti, L., Scaramuzzo, L., Schiro, G. R., Sessa, S., & Logroscino, C. A. (2013). Complications in lumbar spine surgery: A retrospective analysis. *Indian Journal of Orthopaedics, 47*, 340-345.

Rampersaud, Y. R., Moro, E. R., Neary, M. A., White, K., Lewis, S. J., Massicotte, E. M., & Fehlings, M. G. (2006). Intraoperative adverse events and related postoperative complications in spine surgery: Implications for enhancing patient safety founded on evidence-based protocols. *Spine, 31,* 1503–1510,

Roland, M., & Fairbank, J. (2000). The Roland–Morris Disability Questionnaire and the Oswestry Disability Questionnaire. *Spine, 25,* 3115–3124.

Rosenbloom, B.N., Khan, S., McCartney, C., & Katz, J. (2013). Systematic review of persistent pain and psychological outcomes following traumatic musculoskeletal injury. *Journal of Pain Research, 6,* 39–51.

Schoenfeld, A. J., & Bono, C. M. (2011). Measuring spine fracture outcomes: Common scales and checklists. *Injury, Int. J. Care Injured, 42*, 265–270.

Skolasky, R. L., Riley, L. H., Albert, T. J. (2007). Psychometric properties of the Cervical Spine Outcomes Questionnaire and its relationship to standard assessment tools used in spine research. *The Spine Journal, 7*, 174–179.

Smith, H. S., Deer, T. R., Staats, P. S., Singh, V., Sehgal, N., & Cordner, H. (2008). Intrathecal drug delivery. *Pain Physician (Opioid Special Issue), 11,* S89-S104.

Street, J. T., Lenehan, B. J., DiPaola, C. P., Boyd, M. D., Kwon, B. K., Paquette, S. J., Boyd, M. D….Fisher, C. G. (2012), Morbidity and mortality of major adult spinal surgery. A prospective cohort analysis of 942 consecutive patients. *The Spine Journal, 12,* 23-34.

Sudhakar, N., Laing, R. J. C., & Redfern, R. M. (2003). Assessment of fusion after anterior cervical discectomy. *British Journal of Neurosurgery, 17,* 54-59.

Svensson, E., Schillberg, B., Ling, A-M., & Nystro, B. (2009).The Balanced Inventory for Spinal Disorders: The validity of a disease specific questionnaire for evaluation of outcomes in patients with various spinal disorders. *Spine, 34*, 1976–1983.

Tevis, S., & Kennedy, G. D. (2013). Postoperative complications and implications on patient-centered outcomes. *Journal of Surgical Research, 181*, 106-113.

Tuli, S. G., Chen, P., Eichler, M. E., & Woodard, E. J. (2004). Reliability of radiologic assessment of fusion: Cervical fibular allograft model. *Spine, 29,* 856-860.

Turner, J. A., Loeser, J. D., Deyo, R. A., & Sanders, S. B. (2004). Spinal cord stimulation with failed back surgery syndrome or complex regional pain syndrome: A systematic review of effectiveness and complications. *Pain, 108,* 137-147.

Von Korff, M., Jensen, M. P., & Karoly, P. (2000). Assessing global pain severity by self-report in clinical and health services research. *Spine, 25*, 3140–3151.

Vanti, C., Prosperi, D., & Boschi, M. (2013). The Prolo Scale: History, evolution and psychometric properties. *Journal of Orthopaedic Traumatology, 14,* 235–245.

Ware, J. E. (2000). SF-36 health survey update. *Spine, 25*, 3130–3139.

Ware, J. E. (n.d). *SF-36® Health Survey Update*. www.**sf-36**.org/tools/**sf36**.

Zanoli, G., Stromovist, B., Padua, R., & Romanini, E. (2000). Lessons learned searching for a HRQoL instrument to assess the results of treatment in persons with lumbar disorders. *Spine, 25*, 3178–3185.

† Articles met inclusion criteria and were used to identify pain instruments (see Figure 2) but were not directly discussed in the evidence review.

7. APPENDIX A: General Surgery Classification Systems for Surgical Complications

General surgery systems grouped complications by broad, clearly defined criteria such as, severity (i.e., level of intervention required), etiology (i.e., error diagnosis) or major body system (e.g., pulmonary, cardiac).

- Severity classification. Two articles reported systems that used a five-point severity grading system based upon the treatment required to correct complications (Dindo et al., 2004; Mazeh et al., 2014). Both systems are variants of an original system proposed by Clavien and Dindo (2004), which 'revolutionised' complication reporting by shifting the emphasis from the presence of a complication to its grading in terms of outcome (Mazeh et al., 2014). The system proposed by Dindo et al. (2004) – which is more specific and detailed than the original Clavien-Dindo system – categorised complications as:

  Grade 1: required no or limited pharmacological intervention.
  Grade 2: required pharmacological treatment, blood transfusion or physiotherapy.
  Grade 3: required surgical, endoscopic or radiological intervention.
  Grade 4: life-threatening and required ICU management.
  Grade 5: resulted in death.

- Etiology classification. Two separate systems classified complications according to errors. The Harvard Medical Practice Study classified the etiology of adverse events and medical errors into five broad categories (see Mirza et al., 2006):

  1. **Diagnostic errors** (e.g., error / delay in diagnosis).
  2. **Treatment errors** (e.g., technical error in operation).
  3. **Preventive errors** (e.g., inadequate monitoring or follow-up).
  4. **System errors** (e.g., equipment failure).
  5. **Other errors** (unclassified).

Antonacci et al., (2009) also classified complications according to five main error types, specifically, errors of: (i) Diagnosis, (ii) Judgment, (iii) Technique, (iv) Communication / supervision and (v) Miscellaneous (equipment & supply). Errors in diagnosis, judgment, communication and miscellaneous were significantly higher in lethal *versus* non-lethal cases.

- Major body system classification. A comprehensive system of classifying complications according to major body system class was based upon data from 29,237 operative procedures. A total of 245 adverse events were classified according to15 body system categories (Antonacci, 2008):

  1. Cardiovascular system
  2. Endocrine system
  3. Gastrointestinal system
  4. Genitourinary system
  5. Hematologic or vascular system
  6. Musculoskeletal system
  7. Nervous system
  9. Wound or skin
  10. Death
  11. Device malfunction
  12. Infection
  13. Injury
  14. Metabolic
  15. Other

8. Pulmonary system

Further analyses revealed that a subgroup of seven of the above categories accounted for 82% of all adverse events (Infection, Hematologic, Pulmonary, Cardiac, Wound, Gastrointestinal, and (Iatrogenic) Injury).

8. APPENDIX B: Reference List For Addendum

Alexander, M. S., Anderson, K. D., Biering-Sorensen, F., Blight, A. R., Brannon, R., Bryce, T. N., Creasey, G…Whiteneck, G. (2009). Outcome measures in spinal cord injury: Recent assessments and recommendations for future directions. *Spinal Cord, 47*, 582–591.

Atkinson, T. M., Mendoza, T. R., Sit, L., Passik, D., Scher, H. I., Cleeland, C., & Basch, E. (2010). The Brief Pain Inventory and its "pain at its worst in the last 24 hours" item: Clinical trial endpoint considerations. *Pain Medicine, 11,* 337–346.

Bolton, J. E. (2004). Sensitivity and specificity of outcome measures in patients with neck pain: Detecting clinically significant improvement. *Spine, 29,* 2410–2417.

Carey, T. S., & Mielenz, T. J. (2007). Measuring outcomes in back care. *Spine, 32,* S9–S14.

Chapman, J, R., Norvell, D. C., Hermsmeyer, J. T., Bransford, R. J., DeVine, J., McGirt, M. J., & Lee, M. J. (2011). Evaluating common outcomes for measuring treatment success for chronic low back pain*. Spine, 36*, S54–S68.

Cleeland, C. S. (2009). *The Brief Pain Inventory: User guide.* www.mdanderson.org > Education and Research > Departments, Programs and Labs > Departments and Divisions > Symptom Research > Symptom Assessment Tools). DOI 10.1007/s00586-005-1044-x.

Hurst, H., & Bolton, J. (2004). Assessing the clinical significance of change scores recorded on subjective outcome measures. *Journal of Manipulative and Physiological Therapeutics, 27*, 26-35.

Jensen, M. P. (2003). Questionnaire validation: A brief guide for readers of the research literature. *The Clinical Journal of Pain, 19,* 345–352.

Keller, S., Bann, C. M., Dodd, S. L., Schein, J., Mendoza, T. R., & Cleeland, C. S. (2004). Validity of the Brief Pain Inventory for use in documenting the outcomes of patients with noncancer pain. *Clinical Journal of Pain, 20*, 309-318.

Lee, C. E., Browell, L. M., & Jones, D. L. (2008). Measuring health in patients with cervical and lumbosacral spinal disorders: Is the 12-Item Short-Form Health Survey a valid alternative for the 36-Item Short-Form Health Survey? *Archives of Physical Medicine Rehabilitation, 89*, 829-833.

Luo, X., George, M. L., Kakouras, I., Richardson, W., & Hey, L. (2003). Reliability, validity, and responsiveness of the Short Form 12-Item Survey (SF-12) in patients with back pain. *Spine, 28*, 1739–1745.

Maughan, E. F., & Lewis, J. S. (2010). Outcome measures in chronic low back pain. *European Spine Journal, 19,* 1484–1494. DOI 10.1007/s00586-010-1353-6.

Sanderson, K., & Andrews, G. (2002). The SF-12 in the Australian population: Cross-validation of item selection. *Australian and New Zealand Journal of Public Health, 26,* 343-345.

Singh, A., Gnanalingham, K., Casey, A., & Crockard, A. (2006). Quality of life assessment using the Short Form-12 (SF-12) Questionnaire in patients with cervical spondylotic myelopathy: Comparison with SF-36. *Spine, 31,* 639–643.

Tan, G., Jensen, M. P., Thornby, J. I., & Shanti, B. F. (2004). Validation of the Brief Pain Inventory for chronic nonmalignant pain. *The Journal of Pain, 5,* 133-137.

Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. Med Care, 34, 220–233.