Success in Spine Care: The Proof Is in the Measurements, Part III

Daniel C. Norvell1  Joseph R. Dettori1  Jens R. Chapman2

1Spectrum Research, Inc., Tacoma, Washington, United States
2Swedish Neuroscience Institute, Swedish Medical Center, Seattle, Washington, United States

Address for correspondence Daniel C. Norvell, PhD, Spectrum Research, Inc., 705 S. 9th Street, #203, Tacoma, WA 98405, United States (e-mail: dan@specri.com).

In the first Science in Spine article for this series, we covered the ABCs of spine measurements, including (A) baseline factors, (B) treatment factors, and (C) perioperative/immediate posttreatment events. Without these considerations, it is difficult to make much sense of the outcomes we measure. In the second article in this series, we discussed the importance of identifying and measuring “clinically important” outcomes. Selecting clinically important outcomes is a challenging task; however, much thought should go into this decision and it should be tied directly to project objectives and desired claims. Given the increasing costs of health care, health care purchasers, payers, and hospital systems are adopting the concept of value-based purchasing, which is having a significant impact on low-quality providers and hospitals. Quality rankings are now being publicly reported. True measures of quality, such as surgical complications and validated patient-reported outcomes (PROs) of effectiveness, may be burdensome and costly to collect. Therefore, the selection of the appropriate measures should be done judiciously, with an understanding of what makes a quality measure while considering the burden and the yield of such selection. In this final article of this series, we will discuss the selection of PROs, the anatomy of a quality outcome measure, and the importance of understanding why you are collecting the measures you are.

Selecting Outcomes

When selecting a PRO for the clinical setting or a research study, consider the anatomy of a quality measure, which includes the following attributes: validity, reliability, responsiveness, and clinical utility (e.g., patient and clinician friendliness).

Validity

Validity is commonly defined as the extent to which an instrument measures what it is intended to measure. Because validity is not a fixed measure, an instrument should be considered valid for use in relation to a specific purpose or set of purposes and in a specific patient population.1 For example, a valid measure of disability for patients with cervical myelopathy following laminoplasty cannot automatically be considered valid for use in patients with cervical odontoid fractures. Furthermore, validity is not summed up by one concept but can be subdivided into the following concepts: content, criterion, and construct validity. All three are discussed in detail in the AOSpine book, Spine Outcomes Measures and Instruments.2 The focus of this article section will be on content validity. If an instrument does not possess the content to support the objectives or claims, its subsequent validity, reliability, and responsiveness are less important. The basis of the content typically starts with an overarching domain. For example, spine outcome instruments can be divided into the following “core” domains:2,3

- Function
- Pain
- Disability (physical)
- Disability (psychosocial)

Content validity is so important that the International Society for Pharmacoeconomics and Outcomes Research Health Science Policy Council assigned a PRO Task Force whose early deliberations centered around the importance of content validity because PROs are used to evaluate the effect of medical products on how patients feel and function.4 They reported four important threats to content validity:

- Unclear conceptual match between the PRO instrument and intended claim (which can also be applied to a clinical study’s objectives)
- Lack of direct patient input into PRO item content for the target population in which the claim is desired
- No evidence that the most relevant and important item content is contained in the instrument
- Lack of documentation to support modifications to the PRO instrument
When selecting the appropriate PRO, one must first ask the following questions: Does the outcome selected match the intent of the study objectives or clinical questions? Are the claims I hope to make supported by the outcome selected? If not, a new outcome should be considered. This consideration strikes at the heart of the study’s intent. What are you trying to ask, evaluate, or claim? Is it pain relief? Is it function or disability? Is it quality of life? Or is it a combination?

Reliability

Reliability is concerned with the consistency of the instrument. In other words, it is the ability to measure something the same way twice. Reliability can be divided into reproducibility and internal consistency. Reproducibility can be further subdivided into interobserver and test–retest reproducibility, both of which are discussed in more detail in the AOSpine book, Spine Outcomes Measures and Instruments. Internal consistency is a measure of how homogenous or consistent the questions in the scale are and to what extent they are measuring the same thing. Most instruments employ several questions or items to assess a single construct or dimension (e.g., pain, disability), because several related observations typically produce a more reliable estimate than a single observation. To accomplish this task, the questions all need to be similar, measuring aspects of a single attribute. The end result is that individual questions should correlate highly with each other and with the total score of items in the same scale.

Responsiveness

Responsiveness, also known as “sensitivity to change,” is a measure of how well an instrument can detect changes as a result of an intervention. It is possible for an instrument to be both valid and reliable but not responsive, which is problematic when applying an instrument to evaluate a patient’s progress or the effects of a particular treatment. A valid and reliable instrument that does not reflect changes as the patient gets worse or improves is of little clinical or research value.

Evidence suggests that a statistically significant score change from a validated outcomes instrument does not necessarily mean that the change is clinically important. Statistical significance is based on many things, not the least of which is the sample size. Therefore, a small change in patients being evaluated may be statistically significant if the sample size in the clinical trial is sufficiently large, yet the change may not represent a large enough clinical benefit to the individual patient. For example, if a study of two antihypertensive drugs finds a statistically significant difference between treatment groups favoring one drug but the difference in blood pressure is 1 mm Hg, this change is not clinically significant. From a clinical perspective, we need to evaluate whether a treatment effect is worthwhile or important when making evidence-based treatment decisions. For research purposes, we need to know the size of the score change that is clinically important to estimate necessary sample sizes for future studies.

As a result, there is increasing awareness that responsiveness should include the ability to “measure a meaningful or important change in a clinical state.” The concept of minimal clinically important difference (MCID) has been pioneered in an effort to define the smallest meaningful score change. Jäschke et al defined the MCID as the smallest change that the patient perceives as beneficial. A description of how the MCID is determined and calculated can be found in the AOSpine Book of Spine Outcomes Measures and Instruments.

Clinical Utility

For clinical utility, when considering patient friendliness, the following questions are worth answering:

- Can the instrument be completed in a relatively short amount of time?
- Are the questions clear, concise, and easy to understand?
- Will patients be uncomfortable answering the questions?

With respect to whether an instrument is deemed clinician friendly, the following questions should be considered:

- Is this instrument completed by the staff or is it self-administered?
- What is the staff effort and cost in administering, recording, and analyzing?
- How much time is required to train the staff in administering the instrument?

Why Are You Collecting the Measurements?

Before finalizing your battery of instruments, it is important to put them into a summary table or matrix that lists the measures and when they will be collected. This broad overview will allow you to consider whether you are missing key measurements, measuring them at inappropriate intervals, or more commonly, collecting too many measurements. Each of these measurements represents an important piece to the puzzle that will need to be considered whether planning a quality improvement program, a retrospective medical record review, or a multisite randomized controlled trial to ensure that the well-intended and well-conceived study questions are answered. These considerations will help you avoid coming up incomplete or empty at the end of the study.

- Does each measurement directly or indirectly map on to the stated objectives?
- Are there measurements that do not map on to the objectives that are extraneous?
- Are there measurements that are actually surrogates for what you really want to measure that are not necessary and may be redundant?
- Can you eliminate these redundancies?

There is a tendency in measure selection to collect extraneous data that will never be used. The old adage of “Let’s collect it now and decide later if we need it” has the tendency to backfire and potentially create data that is incomplete or not valid due to both respondent and research burden. Statistical problems also can arise from data mining and multiple testing. Measurements should be conceived a priori, tied directly to study or program objectives, and targeted with a clear purpose for collecting them. Your final choices and the rationale for them should be...
documented in your study or program protocol, together with specifics of how and when they will be implemented, before you even begin your data collection.

In a recent systematic review evaluating the change in condition-specific pain, function, and general quality of life after spine surgery, several validated outcome instruments that measure a variety of constructs and domains were available to assess the success of treatment for chronic low back pain. Little correlation was found between the change in pain outcomes and the change in health-related quality-of-life (HRQoL) outcomes after spine surgery for low back pain, indicating the authors were measuring different constructs. Pain and functional outcomes instruments were the most responsive to surgery for low back pain and the only outcomes instruments that demonstrated a large effect size. None of the HRQoL tools (including the Short Form-36) were as sensitive to the treatment. The authors recommended administering a visual analog scale for pain and a physical measure such as the Oswestry Disability Index before and after surgical intervention because these outcomes are the most treatment-specific and responsive to change. They recommend against routinely administering an HRQoL measure or selecting a shorter version (e.g., the Short Form-12) in the clinical and research setting to maximize clinical utility, because the measures are the least responsive to spine surgery. As shown in this example, adding measures that are typically thought to be important may not be useful in your clinical setting.

Summary

A well-planned study includes careful consideration and selection of appropriate measurements that are directly linked to study objectives and hypotheses: baseline, treatment factors, perioperative/immediate posttreatment events, and outcomes. Not accounting for these factors may lead to bias for treatment comparisons where an unequal distribution of factors exists. Not selecting appropriate outcomes may lead to results that do not support the intended claims or objectives of the study.

When selecting outcomes, the consideration of the appropriate PROs is paramount as policy makers, regulatory bodies, and patient groups are requiring outcomes to measure the patient’s perspective.

When selecting a PRO for a clinical study, consider the validity, reliability, responsiveness, and clinical utility (e.g., patient and clinician friendliness). The most important consideration is content validity. Does the measure being selected have items that match the intended study objectives? If not, then the other properties are less important. Once measures with appropriate content are selected, then other aspects of validity, reliability, and responsiveness should be considered. The concept of MCID has been pioneered in an effort to define the smallest meaningful score change. When selecting an outcome, ensure that you measure it at baseline and follow-up so that such a change can be calculated. A review of the literature for the MCID with respect to the measure and population of interest should also be performed.

Finally, be careful in the number of measurements you select. Although you want to make sure that important factors are accounted for, you also have to consider patient and clinician burden. Meeting with your study or clinical research team to review a full matrix of measurements over time for your study will go a long way in making sure you have collected the most important measures without overdoing it prior to your data collection procedures.

References

1. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess 1998;2(14):1–iv. 1–74
2. Chapman JR, Hanson RP, Detori JJ, Norvell DC, eds. Spine Outcomes Measures and Instruments. Stuttgart, Germany, New York, NY: Thieme Medical Publishers; 2007
3. Chapman JR, Norvell DC, Herrmsmejer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine (Phila Pa 1976) 2011;36(21, Suppl):S54–568
4. Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report. Value Health 2009;12(8):1075–1083
5. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. 2nd ed. Oxford, UK: Oxford University Press; 1995
6. Wasseërthiel-Smoller S, Smoller J. Mostly about quality of life. In: Biostatistics and Epidemiology. 2nd ed. New York, NY: Springer-Verlag; 1995:163–170
7. Beaton DE. Understanding the relevance of measured change through studies of responsiveness. Spine (Phila Pa 1976) 2000;25(24):3192–3199
8. Bombardier C, Kerr MS, Shannon HS, Frank JW. A guide to interpreting epidemiologic studies on the etiology of back pain. Spine (Phila Pa 1976) 1994;19(18, Suppl):2047S–2056S
9. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12(4, Suppl):1425–1585
10. Deyo RA, Patrick DL. The significance of treatment effects: the clinical perspective. Med Care 1995;33(4, Suppl):AS286–AS291
11. Epstein RS. Responsiveness in quality-of-life assessment: nomenclature, determinants, and clinical applications. Med Care 2000;38(9, Suppl):I91–I94
12. Hägg O, Fritzell P, Nordwall A; Swedish Lumbar Spine Study Group. The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 2003;12(1):12–20
13. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care 2000;38(9, Suppl):I84–I90
14. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality of life research. How meaningful is it? Pharmacoeconomics 2000;18(5):419–423
15. Testa MA. Interpretation of quality-of-life outcomes: issues that affect magnitude and meaning. Med Care 2000;38(9, Suppl):I166–I174
16. Wells G, Beaton D, Shea B, et al. Minimal clinically important differences: review of methods. J Rheumatol 2001;28(2):406–412
17. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertainment of the minimal clinically important difference. Control Clin Trials 1989;10(4):407–415
18. DeVine J, Norvell DC, Ecker E, et al. Evaluating the correlation and responsiveness of patient-reported pain with function and quality-of-life outcomes after spine surgery. Spine (Phila Pa 1976) 2011;36(21, Suppl):569–574