

This is an author's draft of an accepted article submitted and published in The Spine Journal DOI:
<http://www.sciencedirect.com/science/article/pii/S1529943016003600>

**Measuring spinal surgical success:
the proportion achieving acceptable symptoms.**

Keep it simple but not simplistic

**Invited Commentary for The Spine Journal on:
What level of pain are patients happy to live with after surgery for lumbar degenerative disorders? Tamás F Fekete, M.D.; Daniel Haschtmann, M.D.; Frank S Kleinstück, M.D.; François Porchet, M.D.; Dezső Jeszenszky, M.D., PhD; Anne F Mannion, PhD**

The field of medical outcomes is often traced to Ernest Codman, a US surgeon interested in “end results” and care quality. Codman promoted transparency and accountability. In 1914 his plan for evaluating surgeon competence was decline by his hospital. So he set up his own hospital.

Subsequently, the scientific field limped along until the 1970's when The RAND Corporation's Health Insurance Experiment and Medical Outcomes Study (from which emerged the SF-36 – MOS Short Form 36-item Health Survey) provided a spring board for developing methods of measuring outcomes using patient-report data. The last decade has seen another surge of interest driven largely by the US Food and Drug Administration's patient-focussed logic that interventions be evaluated in terms of how people “survive, feel and function”. This has underpinned widespread recommendations that clinical outcomes assessments (COAs) are routinely incorporated in studies to measure patient-, clinician-, observer–or performance outcomes. Naturally, people themselves are the most proximal and (sometimes only) valid reporters of how they feel (their symptoms) and function (in daily life). Hence the emphasis on patient-reported outcome measures (PROs -US; PROMS - UK).

Many methods of “measuring” patient reported outcomes have been proposed. One of the key unsolved issues is their interpretation. Specifically, what is a meaningful difference or change in an outcome measure? This takes us beyond statistical significance into the murky waters of clinical significance. Whilst the concept seems simple it is not simplistic and so research into PROs and their interpretation is highly relevant and deeply important for patients, surgeons, health care systems, public expenditure, and society.

This edition of *The Journal* reports a study from Zurich's Schulthess Klinik proposing and investigating a different (albeit borrowed) approach to interpreting surgical success: the proportion of people whose symptoms have reduced to an acceptable level - ie reached a satisfactory state. The study reports some results of three selected questions asked pre and post operatively of over 6,000 people undergoing lumbar spine surgery for a range of degenerative disorders. Specifically reported are: two 0 [none] to 10 [worst imaginable] pain Visual Analogue Scales (leg and back), and the symptom-specific well-being question (SSWB) which asks: “*If you had to spend the rest of your life with the symptoms you have now, how would you feel about it?*” *Very satisfied, somewhat satisfied, neither satisfied nor dissatisfied, somewhat dissatisfied, or very dissatisfied?*

Peoples' responses to the SSWB question administered 12 months post-operatively were used to define the proportions of people whose symptoms were acceptable (very + somewhat satisfied) or not acceptable (*neither satisfied nor dissatisfied + somewhat dissatisfied + very dissatisfied*). Answers were

also used to derive mean pain scores associated with each of the 5 satisfaction levels, and “cut off” pain scores. In brief, the authors found:

- approximately 47% had acceptable symptoms 12 months post-surgery;
- mean pain scores for the worst pain (leg or back) corresponded to the levels of satisfaction;
- the 12 month pain score discriminated acceptable from not acceptable symptom states;
- the cut-off pain score for a satisfactory symptom state was ≤ 3 VAS points.

The authors make two main interpretations. First, spinal surgery improves but does not totally eliminate pain. Second, proportion achieving an acceptable symptom state may represent a more stringent surgical success criterion than standard approaches. Theoretically, and independent of any data, the first statement is logical and the second possible.

The results are indeed striking, if interpreted at face value, and will doubtless stimulate debate. For example, it seems surprising that over 50% of people selected for surgery failed to achieve an acceptable symptom state (PASS) at 12 months. Also, nearly half of those who did achieve PASS rated themselves as only “somewhat satisfied”. Only 1 in 4 people rated themselves “very satisfied”. In simple terms, and at face value, these results could be interpreted as demonstrating – in a very large sample – that spine surgery is a poor pain treatment for the lumbar pathologies studied, and is more often ineffective than effective. The potential implications and ramifications for patients, surgeons, and health care provision could be substantial. If, of course, the results are valid.

Herein lies the crux. In order to interpret these results we must be confident that the measurement processes and study design were fit-for-purpose. There are four related but different requirements for this. The method used to define people's symptoms dichotomously as either acceptable or not acceptable post operatively did so validly. The method used to measure pain quantified accurately the pain related to their lumbar pathology. The study design enabled assessments at 12 month post-operatively to be related to the surgery. The impact (or not) of the surgery was accurately represented.

There is a case for significant reservations. For example:

- The symptom-specific well-being question and its response categories are non-specific, ambiguous, and in part require abstraction (“...for the rest of your life”). Which symptoms does the question refer to? Which symptoms are participants reporting? What constitutes “somewhat” satisfied or dissatisfied? What is the evidence that these judgements relate directly to the symptoms for which the surgery was undertaken 12 months previously? This is

particularly pertinent complex symptoms in a sample of people of whom 35% were aged over 70yrs and in whom comorbidities are common;

- There are long standing concerns about the discriminant ability of patient satisfaction questions. These concerns are supported in the study data - the pain VAS mean difference across 3 of the 5 categories (“somewhat satisfied” – “neither satisfied nor dissatisfied” - “somewhat dissatisfied”) is only 2 VAS pts: 20% of the total scale range. This also highlights that statistical significant findings in large sample should be interpreted cautiously;
- There are concerns about using “middle” response categories labelled “neither satisfied nor dissatisfied”. These concerns are supported by the “U-shaped” bimodal distribution of SSWB responses indicating that the response categories may not be mapping out the intended continuum of more to less;
- The performance of VASs as measurement instrument has been questioned, despite their popularity, because their performance can't be adequately tested;
- The process of re-categorising peoples' SSWB 5-level satisfaction responses as dichotomous PASS/not PASS makes an assumption as to how they would have responded to a different question with a different number of different response options;
- Although cut-off scores are an appealing concept, and seemingly simple solution, there are measurement concerns about using them to divide a continua (here pain) and classify individuals as PASS/ not PASS. This is because single time-point, individual-person VAS measurements are highly likely to be associated with considerable error. [Fig 2 shows unlabelled bars implying notable variance and/or error]
- The cut-off for PASS was ≤ 3 points. The baseline VAS data show that 68% of the sample had a baseline pain score ≥ 7 , and 25% of ≥ 9 . To achieve PASS these people would have to have a minimum pain reduction of between 4 and 7 points (40-70% of the possible scale range). So people could have had a substantial benefit and still be interpreted as surgical “failures”.
- Evaluations of interventions for complex symptoms in complex groups when total symptom resolution is not expected require paired pre and post-operative data analysis to quantify the magnitude of change.

It makes sense, and is important, to evaluate surgical success by incorporating assessments of how people feel and function. Computing the proportions of people achieving acceptable symptom states has intuitive appeal as a clinically meaningful concept. However, it is not simple to articulate this

This is an author's draft of an accepted article submitted and published in The Spine Journal DOI:
<http://www.sciencedirect.com/science/article/pii/S1529943016003600>

concept as a fit-for-purpose measurement process. This study is an exemplar of that fact and how a simple process can be simplistic. It is also a demonstration of how study results and interpretations are hostage to fortune to their measurement instruments, why due diligence is paramount, that we should draw on existing knowledge, embrace and apply the sophisticated developments that have occurred in social measurement in the last 50yrs to achieve these goals. Finally, the study demonstrates how seriously spinal surgeons need to take their COAs, and why they need to invest in measure processes.

Conflicts of interest: none

Funding: none

Jeremy Hobart PhD FRCP
Consultant Neurologist Plymouth Hospitals NHS Trust
Professor of Clinical Neurology and Health Measurement
Plymouth University Peninsula Schools of Medicine and Dentistry
Room N13, ITTC Building,
Plymouth Science Park,
1 Davy Road,
Plymouth PL6 8BX, UK
jeremy.hobart@plymouth.ac.uk

**FIT-4-PURPOSE
MEASUREMENT
WITH
PLYMOUTH
UNIVERSITY**