Guest Editorial

A reconceptualization of the pain numeric rating scale: Anchors and clinically important differences

Introduction

Patient-reported outcomes (PROs) have become a recognized standard in patient-centered medical and rehabilitative care. PROs have been used to replace or compliment hard objective findings in clinical management of many conditions that lack a clear gold standard, including emotional distress and pain.[1] PROs provide insights into a patient's pain experience, or their perception of their health condition, that, to date, has yet to be consistently gleaned from available diagnostics (eg, imaging or blood tests).[2,3]

PROs for pain became especially popular following McCaffery's pronouncement in the 1960s that "*pain is what the patient says it is, and occurs when he or she says it does.*"[4] If this statement is accepted, clinicians must then respect that the patient's ratings of their experience is the closest thing to a gold standard in pain management.

Survey data indicate that pain ratings are particularly popular among clinicians—the 0-10 Numeric Rating Scale (NRS) has been described as the most frequently used PRO in clinical practice globally.[5] It is easy to administer both orally and in writing, has been used across all pain populations, appears to be equally useful across languages and cultures, and when compared to other more complex pain measurement tools has demonstrated similar or better psychometric properties.[6,7] Pain intensity ratings have demonstrated value for predicting recovery from acute trauma,[8,9] identifying subgroups of pain conditions,[10] and evaluating response to intervention.[11] However, although it is popular, simple to administer, and has value for clinical decision-making, we would argue that, despite its ubiquity, there are challenges to interpretation that have yet to be resolved.

The purpose of this short editorial is to describe two currently unresolved issues with interpreting scores on the 0-10 NRS: (1) the upper anchor has yet to be universally standardized and may therefore lead to bias in interpretation and (2) meaningful change is more complex than current guidelines would suggest.

Issue 1: The upper anchor controversy

One issue that likely occupies little mindshare in the clinical and research worlds is consideration of what the description of the extreme high end of the scale should be. In most uses of NRS, only the lower and upper ends receive verbal descriptors, though we are aware of some implementations where the intervening numbers may also be labeled (eg, 5 = moderate pain). However, we will focus on the common use. In these cases, the lower end, "0," seems relatively easy to accept. In rational mathematics, 0 represents the complete absence of a thing: no apples, no speed, no pain. So the lower end being labeled "no pain" seems logical and intuitive. But what should the upper anchor be: "The worst pain ever experienced, imagined, or feared"? In mathematical terms, the opposite of zero is 1/0 or undefinable (infinite). Clearly, the 0-10 scale is not useful with an upper anchor that is irrational or infinite, so for clinical use, some boundary needs to be used. This is no small question as the anchor can have dramatic effect on interpretation of patient ratings. Anecdotally at least, the 2 most common anchors are: pain as bad as can be imagined, or worst pain ever experienced. In the case of the former (worst pain imaginable), interpretation of, for example, a score of 6/10 would only be meaningful for treatment decisions if the clinician also had a sense of what that 10 comparator actually represented—in other words, 6 is only interpretable in terms of the worst pain the patient can possibly imagine. How creative is the patient's imagination? How far into the depths of their darkest imagination are they willing to go? We would argue that many readers of this piece would not willingly explore the depths of their own imaginations in regard to the *worst pain imaginable*, and so asking patients to do so may in fact be tantamount to a cruel and unusual request.

Similarly, asking the patient to rate their current pain in relation to the worst pain they have ever experienced is only interpretable if the clinician has knowledge of the patient's pain history. Perhaps, the patient has led a somewhat blessed life, so this current pain from an ankle sprain may well be the worst they have ever experienced. In that case, it should be 10 based on the scale with which they have been presented, though they may not appear to the outside observer to be writhing in pain. This is a distinct but related bias, that clinicians very likely have their own implicit understandings of what behaviors should be observed at different levels of pain intensity. Our experience in speaking with clinicians would suggest that they are highly skeptical if/when a patient reports having 10/10 pain, especially when other indicators or observations are not in accordance with such a high rating. Similarly, this patient may have already experienced unspeakable atrocities in their lives, so this current pain, while intense and requiring intervention, may in fact pale in comparison to the "worst pain ever" anchor, only receiving a rating of 3/10. In this case, they (unknowingly) risk being undertreated due to low pain intensity, when in fact the low number is a function of the scale, not their actual experience.

Pain in relation to the scope of one's imagination or the length of one's life is two very different things and should be expected to result in different patient ratings. So, perhaps, the most important caution being presented here is the importance of choosing and using a consistent anchor. Switching between anchors will render the validity of the scale almost meaningless. But on a broader scale, perhaps it is time for the pain assessment community to agree on a single universal anchor that has cross-cultural meaning. As an example, we currently like the anchor of "extreme pain" because, not only does it hold at least some degree of meaning, but it is, by definition, the correct label for the extreme end of a scale. Even here of course the interpretation of "extreme" will vary by patient.

To our knowledge, there has yet to be an empirical evaluation of the effect of different "10" anchors on patient pain ratings. In the absence of empirical evidence, we are forced to rely on theory and common sense. We therefore urge clinicians to give due attention to their choice of a "10" anchor and use it consistently within patients and to consider carefully how their choice of 10 can be interpreted differently between patients and also between the patient and clinician.

Issue 2: The meaning of change

Although the NRS can be useful for prognosis, subgrouping, and treatment planning (with the requisite caveats previously mentioned), it is arguably most commonly used as an evaluative tool. That is, the NRS best provides a semi-standardized marker of change in pain between a preintervention and postintervention condition. Traditionally, researchers would have evaluated the significance of that change through inferential statistics, comparing the mean of time 1 or group 1 with the mean of time 2 or group 2. If the 2 means differed by an extent that was deemed to be greater than that would occur by chance, that is, if the probability that the two means were not the same was at least 95% (translated to a $P < .05$), most researchers would have confidently deemed the change to be significant. But in 1989, Jaeschke[12] introduced the concept of clinically important difference (CID) (sometimes termed minimal clinically important difference or MCID). This came out of recognition that with sufficiently large samples even a very small change in pain rating could reach statistical significance by virtue of sample size rather than treatment effectiveness, but the question remained of whether that very small change was actually important to the patients experiencing the pain. Since the introduction of the concept of a CID, academics have engaged in a series of evaluations of several scales to determine the optimal cut score for discriminating between those who have changed an important amount, and those who have not changed an important amount. Interestingly, the determination of important change, the "gold standard" for change, has itself most commonly been a separate PRO, often referred to as a Global Rating of Change (GROC).[13] This begs the question of why the GROC has not simply become the universal PRO for all evaluation of change but that is a discussion for another forum. For now, it is worth noting use of an anchor-based method such as that offered by the GROC, and then using the GROC as a comparator against which to identify the best NRS cut score, itself presents interpretative challenges. The first is the cut score chosen will be the value that best discriminates between improved/not improved groups in a sample, but will not necessarily be a good discriminator. Readers can evaluate this by interpreting either the magnitude of effect size reported in a study or the area under the receiver-operating characteristic curve (abbreviated as "area under the curve" or AUC). When interpreting the area under the curve, a value of 0.50 indicates discriminative ability no greater than chance, whereas a value of 1.00 indicates perfect

discriminative accuracy. So even though a single value may be endorsed as the "best," there is no universal standard for what good discriminative accuracy should be. In the case of the NRS, current practice is to consider 2 points[14] or 30% of baseline as meaningfully important change.[15,16] The field has yet to reach consensus, and regardless of the value chosen, interpretation of change scores where the starting score is 2/10 or less remains unresolved.

But there is a more important and slightly more complex issue to consider when interpreting change scores—that important change is very likely not consistent across the entire breadth of a scale. This can be demonstrated both theoretically and statistically. Theoretically, consider the space between 0 and 1 on the NRS. While not meeting the threshold for a meaningful change as derived in the examples previously mentioned, this change should arguably be considered an important threshold. The experience of the patient has gone from nothing to something (or, in the other direction, from *something* to *nothing*). Similarly, consider a score that starts at the extreme other end of the scale, a change from 10 to 9. Here, the experience of the patient has gone from the absolute worst experience possible (or ever, depending on the anchor) to something less than that. In both cases, it would seem that a 1-point change is meaningful when the experience (pain intensity) starts at one of the extreme ends. In contrast, a 1-point change in the mid-range, say from a 7 to a 6, likely holds less inherent meaning.

This phenomenon of more meaningful change at the poles than in the middle of a scale can be demonstrated through Rasch analysis.[17] In brief, Rasch analysis, as a means of exploring the linearity of a scale, can provide a means to transform an ordinal-level scale score (such as a 0-10 NRS) into a linear score using a logarithm-based transformation procedure. One of many outputs of Rasch analysis, assuming good model fit, is creation of a conversion matrix. Using the matrix, a "raw" ordinal-level score can be plotted on a "transformed" ratio-level scale. In doing so, the distance between adjacent scores becomes constant. While the distance between scale points on an ordinal scale lack any inherent mathematical meaning (the distance between 1 and 2 cannot be assumed to be the same as the distance between 2 and 3), the distance between points on a ratio- or interval-level scale does hold mathematical meaning and is constant (so, the distance between a 1 and 2 *is* the same as the distance between a 2 and 3). The means through which this occurs is a large subject area best described in dedicated manuscripts and texts.[17-19] Further, it should be noted that Rasch (and a conceptually and mathematically similar approach called Item Response Theory) has not found universal support within the measurement science field,[20] but over 2500 published manuscripts returned in a PubMed search of "Rasch analysis" (searched 2017-11-11) suggests it is adequately accepted by the scientific community at large.

Rasch cannot be conducted with a single-item scale; part of the analysis requires use of 1 source of information (at least 1 item) to predict location on another source of information (at least 1 other item). For the purposes of this editorial example, we used Rasch analysis (RUMM 2030; RummLab Inc.) to evaluate the "current pain" 0-10 NRS of the "pain severity" subscale of the Brief Pain Inventory, against the information from the other 3 items on that subscale (pain at its worst over 24 hours, pain at its least over 24 hours, average pain over 24 hours) in a sample of 109 people with acute trauma from one of our existing databases. The "pain at its worst over 24 hours" item did not fit with the remaining 3 items (fit residual = 4.12) and was removed. Fit to the Rasch model for the remaining 3 items was acceptable (mean fit residual = 0.24, standard deviation = 1.02, total item $\chi^2 = 3.53$, $P = .74$) suggesting that the results of the analysis appear to be adequately trustworthy, though the sample skewed slightly toward lower pain ratings.