Reviewer's report

Title: Measuring patient reported outcomes: Moving beyond misplaced common sense to hard science

Version: 2 Date: 3 January 2011

Reviewer: Stephen Joel Coons

Reviewer's report:

I very much agree with the authors’ basic premise that, all too often, the development of patient-reported outcomes measures is undertaken too cavalierly. Many people who should know better, do not consider the scientific rigor that is necessary to produce a PRO instrument that can provide valid and reliable measurement of the target constructs. However, I have some very serious concerns about the path the authors take in this manuscript.

Major Compulsory Revisions

1. I believe this commentary attempts to cover too much territory and, as a result, the message is compromised. From my perspective, the authors and readers would be better served by a commentary addressing instrument development or a commentary addressing translation and cultural adaptation. The authors acknowledge that parallel instrument development in the languages/cultures required rarely happens. As a result, the translation/adaptation process is most often a separate undertaking. As currently written, some issues involved in the development of PRO instruments are insufficiently described and substantiated. There are assertions made that are not adequately supported by the existing literature. In moving beyond the critical issues involved in instrument development, the authors select just one (i.e., translation and cultural adaptation) of the other factors that must be considered when selecting and using a developed measure. For instance, a PRO instrument could be chosen for use in the culture/language in which it was developed. In this case, the method of instrument administration (e.g., paper-based, hand-held computer, smart phone, interactive voice response system) and the equivalence of the data obtained on the alternative methods versus the original administration format is an important issue; cross-cultural adaptation isn’t relevant.

2. Table 1: This table needs to be explained. There are entries in the table that are not self-evident. Why is the construct being assessed under Functioning labeled “Disability”? Are levels of patient-reported functioning only seen as negative? Why not “Ability”? What is the rationale for equating “Health Status” and “HRQL”? The entries in the “Utility” row are puzzling and I believe this characterization of utility as a “Type of PRO” is problematic from a number of perspectives.

3. Page 2, para 4: The three sentences in this paragraph (that start with “To
summarize”) reflect an over-generalization. For someone who is experiencing excruciating pain, the relief of that symptom is likely to be that person’s primary concern/objective. In addition, a poorly developed (e.g., “expert”-based versus patient-based) “QoL scale” might assess outcomes that are not the patient’s primary concern.

4. Page 2, para 5: What evidence supports the assertion that “The most widely applied QoL model is concerned with the extent to which disease and its treatment prevents an individual from meeting his or her needs”?

5. Figure 2 is more bewildering than illuminating. Unless substantive information is provided to explain it, this figure should be dropped.

6. Page 2, para 6: Why are treatment satisfaction and utility assessment addressed in the same paragraph? They are very different constructs. In most uses of the EQ-5D and HUI self-report instruments, the population-based “utilities” for the health states have already been derived. The respondents are just reporting their health status based on each of the instrument’s descriptive system. Hence, most uses of the EQ-5D and HUI do not involve “utility assessment.” (Please note that it is the “Health Utilities Index” not “Health Utility Index.”) The last sentence of this paragraph states that “more accurate measurement” of utilities can be obtained via responses to “disease-specific QoL measures.” It would help if this very strong assertion was supported by empirical evidence other than that generated by the authors of this manuscript.

7. Page 3, para 2: The authors assert that “it is now generally accepted that such comparisons are scientifically flawed and invalid,….” I am unsure how such a provocative assertion can be made without more substantiation than a cited commentary co-authored by one of this manuscript’s authors. There is scientific debate around this issue but is far from settled.

8. Page 3, para 4: I am unsure who is acknowledging “that generic measures are of little value in the measurement of health.” Again, this is a very strong assertion without empirical evidentiary support. In addition, do the authors have a citation for the development of the SF-36 in the “early 1970s”?

9. Page 3, para 5: To state that when using a disease-specific questionnaire, “patients are only asked questions that are relevant, meaningful and acceptable to them” is, again, an over-generalization. A disease or condition can manifest itself in multiple ways within the population affected by it; although there may be many commonalities, all patients do not necessarily experience all the same symptoms or functional impairments. Hence, some items on the questionnaire may not be relevant and/or meaningful to all patients with the same condition or disease.

10. Page 4: The section heading labeled “Use of PROs within the medical sector” is confusing. I am puzzled as to what is meant by the “medical sector” since this section doesn’t provide the reader with a clear context in which the PRO measures will be used. It appears to move back and forth from clinical decision
making at the physician level to regulatory decision making in the context of clinical trials. The authors spend a considerable amount of text on “QoL” when neither the FDA nor EMA will accept “QoL” as a primary or key secondary endpoint in clinical efficacy trials. This whole section (pages 3-4) needs to more effectively define and focus on the specific context of use. The term “medical sector” does not provide sufficient clarity.

11. Figure 3: The diagram does not effectively communicate the types of PRO measures used and the relationships among them.

12. Page 4, para 3: The statement that “QoL is the primary outcome of relevance and importance to patients” is not universally true. There is a significant body of evidence showing that patients are willing to accept decrements in QoL (e.g., due to chemotherapeutic regimens) if there is a possibility of increasing length of life. The second sentence in this paragraph acknowledges this fact somewhat. Therefore, the first sentence should be written in a less absolute manner. In addition, I don’t believe the last sentence of this paragraph reflects the reality of many predominantly asymptomatic conditions. For instance, if the objective of an intervention is to decrease blood pressure since we know that high blood pressure has been shown to increase morbidity and mortality, then QoL is not likely to be an effective means of evaluating treatment efficacy.

13. Page 4, para 4: The last half of this paragraph ignores the fact that the needs-based approach to QoL assessment is not likely to be the most effective approach to measuring the most salient outcomes in efficacy trials. Related to the authors’ example of sexual functioning, if an erectile dysfunction drug is being tested in a clinical trial, patients for whom having an erection is irrelevant will not be enrolled in the trial. Hence the trial endpoint will focus on erections, not on the ways the patient has adapted to his inability to have erections.

14. Page 5, para 1 and 2: These two paragraph are not helpful. I do not understand the relevance for these needs-based QoL measures in the context of a registration trial submitted to the FDA or EMA. In addition, the authors use this opportunity to criticize the NHP, SIP, and SF-36 again, but these “HRQL” instruments are not relevant when discussing clinical trials.

15. Page 6, para 1: There are many cases in which PRO measures serve as the primary endpoint in clinical trials submitted to the FDA. There are no identified biomarkers for many conditions (e.g., pain, IBS) and patient report provides the most appropriate efficacy endpoint. Willke and colleagues (2004) reviewed the efficacy endpoints reported in product labeling for new molecular entities approved by the FDA from 1997 through 2002. They found that PRO measures were the only endpoints reported for 23 of the 215 products reviewed. (Willke RJ, Burke LB, Erickson P. Measuring treatment impact: A review of patient-reported outcomes and other efficacy end points in approved product labels. Control Clin Trials 2004;25: 535–552.)

16. Page 6, para 1, last sentence: References 21 and 22 are not sufficient to support the statement that the SF-36 “has never proved to be a valuable
instrument for showing differences between treatments.” This is an overly strong and, I believe, demonstrably incorrect statement.

17. Page 6, para 5: In a registration trial conducted for submission to the FDA and/or EMA, I am unclear how and why comparing trial data to outside data (“healthy populations or other disease groups”) would be done. In a clinical trial, the comparator is either a placebo control arm or an active control arm. Comparison to data not collected in the trial would not be acceptable in this regulatory environment. Again, this lack of clarity points out the need for the authors to focus on a specific PRO assessment context. The lack of focus and reliance on broad generalizations is problematic. Although there are basic scientific principles that apply across PRO measurement in general, the PRO measurement context is critical since it introduces other factors (e.g., regulatory requirements) that cannot be ignored.

18. Page 7: In regard to the section titled “Development and validation of PRO measures,” the measurement context remains unclear. This is an important clarification since a considerable amount of the text in this section addresses measuring of QoL, which is not a relevant measurement target for an efficacy endpoint in a regulatory trial. The authors should decide on the intent of the PRO assessment and then use relevant examples.

19. Page 7, para 1 under “Generation of questionnaire content”: The sentence stating that “Such interviews are not intended to explore issues identified from the literature or clinical experts” needs further clarification. At times it may be important to confirm or debunk issues identified from the literature or clinical experts with patients during the concept elicitation process. I am not intending to be overly critical, but this unqualified statement is an example of the cursory nature of this manuscript. Concept elicitation and item generation is an incredibly important step in the instrument development process. Instrument developers need to avail themselves of all opportunities to inform the item generation process by exploring the overlap or lack of overlap among the sources (i.e., literature, experts, patients,) of item content.

20. I will not take the time to document other parts of page 7-9 that are too superficial and/or insufficiently supported by references to the broader scientific literature, but this overview of the “Development and validation of PRO measures” is emblematic of my concern that the authors are taking on too many issues in this one manuscript. As a result, few of the issues are adequately addressed or substantiated. I will provide one more example of this. In Figure 4, the authors list “Reliability (reproducibility)” and state that the “ideal” test-retest time interval is two weeks. In reality, the “ideal” retest interval is dictated by the characteristics of the construct being measured in the target condition/disease. Without further qualification, providing rules of thumb such as “2 weeks ideal” for the test-retest interval diserves the reader and perpetuates unsubstantiated measurement myths.

21. Second text box: The second point in this text box is that “HRQL consists of symptoms and function.” This ignores a critical of element of HRQL—the
patient’s values, beliefs and judgments about the impact of symptoms and
dysfunction on his or her life. HRQL is not just an inventory of symptoms and
function (or dysfunction).

22. Page 7, para 6: The statement that “Well developed measures are now
generally of a better quality and are more sensitive than clinical outcome
measures, even those considered to be objective” is another overly broad
assertion. I wholeheartedly agree with the first part of the sentence, but I
disagree with the last part since it is not that absolute. The truth to the later part
of the sentence depends on the condition/disease and the purpose of the
intervention. For example, in the early stages of polycystic kidney disease (PKD),
there are no manifestations of the disease that are apparent to the patient; the
disease does not affect the patient in ways that he or she can perceive. However,
if an effective treatment of early stage PKD were tested, radiological evidence of
reduction or stabilization of total kidney volume would be the best efficacy
endpoint. Patient-reported endpoints would provide no evidence of treatment
benefit.

Minor Essential Revisions

23. In the first sentence of the Abstract, the word “patient’s” is singular
possessive and should be followed by “his or her” rather than “their.”

24. I recommend that the authors be much more precise with the use of “PRO”
as an abbreviation. I firmly believe that “PRO” should be used when referring to
“patient-reported outcomes” not patient-reported outcome measures or
instruments. All too often the authors write “PRO” or “PROs” when they should
be writing “PRO measure” or “PRO instruments,” respectively. It is an inapt
shorthand that blurs the very important distinction between the target of
measurement (i.e., PRO construct) versus the measurement tool (i.e., PRO
instrument).

25. Figure 1: In the fourth text box in the right column, shouldn’t the fourth bullet
point end with “responsiveness” rather than “reproducibility”?

26. Page 2, para 1: The first sentence states that the article highlights “current
standards in the development and testing of PRO measures” but the authors
address issues for which there do not appear to be consensus standards. If fact,
some of the issues are rather contentious. Hence, it would be more appropriate
for the authors to indicate that lack of consensus may exist on some issues and
that the relevant text represents their perspective on what the standards should
be. In addition, this manuscript goes well beyond the realm of development and
testing of PRO measures.

27. Page 2, para 3: A citation should be provided to support the authors
statement that the “term PRO was coined in 2000.” Was it truly “coined” in 2000
or was that the year it was adopted as the umbrella term for patient-reported
clinical trial endpoints within the US Food and Drug Administration’s (FDA’s)
regulatory context?


Discretionary Revisions

30. It is my understanding that “patient-reported” in the context of patient-reported outcomes is a compound adjective and should be hyphenated.

31. In the fourth sentence, the authors refer to this manuscript as a “review” but I believe that is a misnomer. This is an opinion piece and should be labeled accordingly.

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.