



April 4, 2006

Division of Dockets Management (HFA-305)  
Food and Drug Administration  
5630 Fishers Lane, Room 1061  
Rockville, MD 20852

**Re: Guidance for Industry – Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [Docket No. 2006D-0044]**

Dear Captain Burke:

As an organization, the Center for Health Outcomes Research at United BioSource is pleased to see the release of the draft PRO guidance document. This document reflects the extensive work of health outcomes researchers and the thoughtful input of FDA staff. Below we document some specific comments members of our organization have made in response to this document. Our goal is to assist with the finalization of a PRO guidance document of maximal use to health outcomes researchers involved with regulatory submission in the US.

1. Section I, p. 2, line 45: definition of instrument – It is extremely helpful to have the definition stated here and to appropriately include not just the text of items, instructions, and response scale but also the scoring instructions and administration.
2. Section II could be enhanced by setting context for this document and recasting the definitional lines drawn here. The relationship of patient derived data to the multiple methods of obtaining those data deserve attention, and the distinction between objective laboratory measurements and non-laboratory, potentially subjective measurements would add clarity. The latter includes clinician report, other proxy report, and interviewer-based measures (e.g., HAM-D). Patient performance measures are distinct (e.g., neuropsychological measures) yet can be entirely patient reported. Stated recognition of the close relationship of PRO measurement to other forms of measurement (like structured interviews) would be helpful.
3. Section II, p. 2, lines 63 etc and Section IVB.1, p.10, lines 305 etc.: Perception of performance capacity vs. self-report of performance may both be useful ways to address valid concepts and recommending against the former is too restrictive. Recognition should be made that both types of measurement are distinct from performance-based measurement. To extend on the point made above, some PROs may incorporate performance-based measurement as well.
4. Section IIIA.3, p. 4, line 120 etc.: Reducing variability through training in multiple investigator studies is an important point.

5. Section IIIB, p. 4, line 150: The text mentions generic, condition-specific, treatment-specific – it would be helpful to have the text and Table 1 be consistent (Table 1 references generic, condition-specific, and population-specific). It may be more useful to categorize measures as generic and condition-specific only, with “treatment-specific” a possible subcategory for condition-specific and “population-specific” a subcategory for both generic and condition-specific.

6. Section IIIB, p. 5, line 155: The document makes the point that AEs should be measured separately from effectiveness. While it is important to separate these concepts for regulatory purposes, patient global ratings that are based on multiple treatment features (e.g., perceived effectiveness, side effect profile) are a valuable source of information and can be useful as PROs. Similarly, aggregate scores that combine across relevant features may be appropriate for some measurement applications. In addition, for some domains patients may not be able to make attributions as to whether the positive or negative effects of treatment are influencing the domains of interest. For example, in chronic hepatitis C both the disease and the currently available treatments are associated with increased fatigue.

7. Table 1, p. 5, line 164 etc.: Timing of administration can also be at irregular intervals – this should be included here (e.g., ecological momentary assessment).

8. Section IV, p. 6, line 179: Evaluation of instrument modifications as new instruments is not always appropriate and certainly not always practical. There are many minor changes to PRO instruments that are not viewed as instrument modifications, and these ‘new’ instruments should not require full psychometric evaluation. It is advisable to complete pre-testing and cognitive interviewing if there is evidence that the changes to the PRO measures might change the content or responses. Rewording should be considered here to provide additional context as to when modifications may be acceptable without re-validation.

9. Section IVA, p. 7, lines 194 etc.: The section on development of the conceptual framework is extremely important and an important contribution for the document. See also Glossary, line 1049 – confirm vs. support – as it is always empirical and non-observable, the definition should state that the validation process supports rather than confirms the conceptual framework.

10. Section IVA.1, p. 8, line 213: One-item PRO instrument: The document states that a one-item PRO instrument may be a reasonable measure to support a claim concerning a concept of interest, if documentation exists to support that this one item is a reliable and valid measure of that concept. The problem with this statement is that an instrument with only one item is generally considered to be an unreliable measure. Specifically, the usual reliability coefficients cannot be computed, such as coefficient of equivalence, split-half reliability, and coefficient of internal consistency. In addition, it is very difficult to demonstrate that a single item is the complete and comprehensive measure of a domain or concept, i.e., that it has content validity. In short, one-item

PRO measurement should generally be discouraged. However, there may be situations (say when measuring pain intensity) where there is sufficient research evidence to support a one-item measure.

11. Section IVA.1, p. 8, lines 227 etc.: We do not fully agree with the FDA position on mixing constructs within an instrument. The acceptable match between claim and measure used to support it is an important topic and specificity is always desirable. However, some PRO-relevant concepts like social functioning or psychological well-being may be appropriately measured by a multidomain instrument and the relative importance of component domains may vary across individuals.

12. Section IVB. Creation of the PRO Instrument, p. 9-12, lines 282 etc.: It is very encouraging to see that the document emphasizes the rigorous and thorough process of developing the PRO instrument from generation of items, instrument formatting, scoring, to finalization of the instrument (pp. 9-14). One important point worth noting is that although items can be generated from many sources including literature review, focus groups, interview with patients, clinicians, and researchers, **item generation is “incomplete” without patient involvement (page 10, line 295)**. The document emphasizes the readability and understanding of the items by the patients again on page 12. We appreciate inclusion of this point as we often see poorly constructed items with words with ambiguous meaning in some PRO instruments.

13. Section IVB.3, p. 10, line 329: “it is important to consider patients’ ability to accurately recall the information requested...” According to the document, recall to the beginning of the study is usually inappropriate but in some cases the concept of perception of change since an earlier timepoint could be valid. This point should be mentioned to avoid inappropriate exclusion of alternate recall periods in PRO measures.

14. Section IVB.4, p. 12, line 378: We agree that evaluation of patient understanding is important and would recommend inclusion of modifications to directions and response scale as possible outcomes of cognitive debriefing as well (not just item modification).

15. Section IVB.6, p. 12, line 382; page 20, lines 579 etc. and Section IVD. Modification of an Existing Instrument: In many places the document emphasizes that an instrument has to be validated when it is first created and whenever the instrument has been modified. Research supports the idea that any amount of change can alter the way that patients respond to the instrument. We generally agree that once an instrument is validated, its use should be standardized in all aspects, including format, instructions, and scaling.

16. Section IVB.7, p. 13, line 416, Equal Weighted scores for items: The statement “Equally weighted scores for each item are appropriate only when responses to the items are relatively uncorrelated,” is puzzling. For a set of items to measure the same concept they have to be correlated. For example, in a set of items that are intended to measure the patients’ view of their pain intensity, all items have to ask about “pain

intensity.” It is inconceivable that two items can be relatively uncorrelated but still measure the same concept. On the other hand, if the two items are independent then they probably measure two different concepts. If they are to form a single score, that score is usually called a “composite.” It is very common in PRO instrument development to create equally weighted score of multiple items that measure the same concept, such as the mean or sum of the individual item scores. It has the advantages of being easy to understand and less likely to be incorrectly calculated. We recommend the FDA clarify whether this “equal-weight” statement is intended for a set of items measuring the same concept forming a single score or sets of items measuring different concepts forming a composite score.

On the other hand, this statement can also be interpreted as the “response to one item is independent from the response to another item.” If this is the case, this characteristic is often referred as the “local independence” assumption. When two items are locally dependent, their collected information regarding that patient is less than two locally independent items will provide. It then is reasonable to say that they are over-weighted when they are treated as two equally weighted items. The extreme scenario will be to ask the exact same item in the same instrument twice. This can only be counted as one item. We would like to see clarification of the meaning behind this statement.

17. Section IVB.7, p. 13: We appreciate that the FDA will consider if response choices represent appropriate intervals but basing this evaluation on review of item distributions is not entirely logical. Sample characteristics may inform distribution of responses and may not reflect true interval appropriateness. Additionally, the question of how patients perceive the distance between response options is asked in cognitive debriefing interviews to ascertain the “qualitative” difference among response options. While the majority of patients will understand the response options and the qualitative difference among the response options, they may not to use the entire response option range.

18. Section IVB.7, p. 13, line 429: We recommend expanding discussion of the appropriate application of population-specific preference weights. Also, definitions of population “equivalence” or sufficient equivalence for acceptance of psychometric data requires additional attention. It would be helpful to have specific recommendations on psychometric data that are required from a new population when a PRO is used on a different population than the one(s) on which it was developed and validated. It may not always be necessary to match inclusion/exclusion criteria to the development sample and in some cases careful cognitive interviewing may suffice.

19. Table 4, p. 17, line 483 etc. and Section IVC.4a, line 537 etc.: The FDA is looking for comment on MID and responder definitions. Given the current literature on MID it would be helpful to add language to the guidance indicating that confidence in a specific MID value evolves over time and is confirmed by additional research evidence, including clinical trial evidence. In addition, it should be noted that responsiveness and MID may vary by population and contextual characteristics, and there may not be a single MID value for a PRO instrument across all applications and patient samples.

There is likely a range in MID estimates that vary across patient population and clinical study context. The main point should be emphasized that evidence is required to support the psychometric characteristics of the PRO instrument such that there is confidence that changes in scores over time with the application of treatments with some efficacy can be detected and that the measurement error (or noise) is not so large that it is problematic to observe meaningful changes in patient health status.

Addition of language could serve as a reminder that the key to interpretation is determining the appropriate “*decision threshold*.” There are many methods for establishing that threshold and MID is just one.

The definition of MID in the glossary (line 1075) is “*The amount of difference or change observed in a PRO measure between groups in a clinical trial....*” and does not reflect the minimal important difference that a *patient* perceives. We recommend clarification of the applicability of the MID criterion to within-subject vs. between-group comparisons with consideration of noted placebo effects.

20. Section IVC.1, p. 18, line 495 etc.: The document would benefit from recognition of the distinct purposes of internal consistency reliability data and test-retest reliability data. The document states that test-retest reliability is the most important type of reliability for PRO instruments used in clinical trials, whereas internal consistency reliability does not generally constitute sufficient evidence of reliability. We disagree with the implied evidence hierarchy. We agree that test-retest reliability is one of the important indicators of the reliability, and should be included when it is appropriate and feasible. However, we also believe that there are circumstances when to test-retest reliability may not be suitable or applicable. For example, there may be a practice effect or improvement by the medication over time that influence the results of the test-retest of the PRO instrument. On the other hand, there may be situations where retest of the instrument is not feasible. We would like to argue that when test-retest reliability is not appropriate, internal consistency reliability should be used as sufficient evidence of reliability. Internal consistency reliability has been well studied and used and is a good and conservative approximation of true reliability.

21. Section IVC.4.a, p. 19, line 545: "If PRO instruments are to be considered more sensitive than past measures..." – that's a high bar and although increased sensitivity is often a motivator for creation of PROs, it's not always. Sometimes the goal of a PRO is to obtain the patient perspective, and optimizing all psychometric performance is always part of good instrument creation. Rewording would mitigate concern about an “unlevel playing field.”

22. Section IVD.5, p. 22, line 654: The document acknowledges the need to follow “accepted standards” for translation and cultural adaptation and to provide support for the accuracy of translated PRO measures and the validity of the resulting data. However, the document does not specifically state what the “accepted standards” are, though it mentions the need for experienced translators to carry out the translations, an

adequate translation/adaptation methodology, harmonization of the various versions, and evidence that the measurement properties are similar across versions. This last point could be interpreted to suggest that a psychometric validation study would be required for each new version. This is not in line with current practice for translation and cultural adaptation and would represent a heavy additional burden for sponsors. The document would benefit from a clear statement as to what type of “evidence” would be required or considered sufficient to show that measurement properties of various versions are comparable. Also, it would be important to state whether existing translations which have been used widely but have not been psychometrically validated, would need to undergo this type of evaluation to meet FDA Guidelines.

23. Section IVE.2 references cognitive impairment of the respondent and suggests inclusion of proxy reports in addition to patient report to address expected decrement in respondent ability to accurately self-report over time. There are several other issues related to collection of PRO data from individuals with cognitive impairment, and the document could benefit from referencing some of these issues. First, validity and reliability of self-report warrant special consideration for individuals with cognitive impairment. Presence of cognitive impairment can be compatible with collection of meaningful patient report but the expectations for PRO measurement must be stated clearly in advance. Correspondence with proxy report may be moderate to low for a variety of reasons unrelated to actual measure validity and reliability, and therefore proxy report must be viewed as a separate form of measurement. Finally, the document does not include proxy or caregiver report under the rubric of PRO. For cognitive impairment in particular, this may be problematic. The standards expressed for patient self report apply to proxy report for a patient and consideration should be given to including non-clinician proxy report as part of PRO measurement.

24. Section VA.2, p. 24: Clinical trial QC and standardized instructions are extremely important and mention in the guidance is helpful.

25. Section VF, p. 26, lines 830-836: Electronic Capture of PRO Data: The document states that it is problematic if the direct control over source data is maintained by the sponsor or contract researchers and not by the clinical investigator, and that the sponsors should avoid direct PRO data transmission from collection device to the sponsor. Removal of investigator from this chain is a concern in terms of accountability for confirming the accuracy of the data. We suggest that FDA clarify the definition of clinical investigator and clarify acceptable methods of data capture and transmission to encompass usual practice.

26. Section VI, p. 27, lines 859 etc.: We agree with the instruction to integrate PRO documentation with rest of documentation (protocols, statistical analysis plans) and with the statement that PRO validation within a trial can be described in a separate section of the SAP.

27. Section VID, lines 956 etc., Missing Data: For statistical strategies to handle missing PRO data, the document advises against using “completer only” or “last observation carried forward” methods. The concern is that the missing data may be treatment-related. The FDA recommends the use of several different imputation methods and an assessment of the consistency of the results using each method. This is reasonable and sound advice for conducting clinical trials. However, the amount of missing data is generally unknown prior to data collection so the option endorsed seems to be inclusion of multiple potential “imputation strategies” for dealing with missing data when writing statistical analysis plans. In addition, the requirement for conducting multiple imputation methods and assessing the consistency is both time-consuming and costly.

28. Glossary, line 1097: We recommend wording to include lack of decrement as appropriate to allow for a more comprehensive definition of benefit.

In general, this document provides an excellent guidance to those seeking to submit PRO data to the FDA. One final point is that the FDA should consider creation of a separate entity, perhaps formed from a coalition of industry, academia, and FDA staff, who can aid with determination of what constitutes "good enough" evidence for submission.

Respectfully submitted,



Dennis Revicki, PhD and Lori Frank, PhD, on behalf of the  
Center for Health Outcomes Research  
United BioSource Corporation