



Guidance for Industry
Patient-Reported Outcomes Measures: Use in Medical Product Development
to Support Labeling Claims
DRAFT GUIDANCE

Docket Number: 2006D - 0044

Comments and Suggestions regarding the draft document as submitted by
Healthcare Technology Systems, Madison, WI.

April 3, 2006

We applaud your commitment to setting a scientifically rigorous standard for patient-reported outcome (PRO) measures used in clinical trials to support product-labeling claims. Given the high-stakes decisions being made from PROs in clinical trials, we agree that reliability and validity evidence should be gathered to support their intended use. In fact, we believe the instrument development and validation practices outlined in the Guidance document should apply to all measures used in clinical trials, and encourage the expectation of similar rigor for all instruments used to support labeling claims, including clinician-administered assessments.

Our review of the document brought forth several questions and issues needing clarification.

Each is outlined in the following document.

Submitted by:

James Mundt, PhD
Vice President, Research and Development
Healthcare Technology Systems

2006D-0044

CS

P3, line 83-89

For example, PRO-based evidence of improved symptoms alone generally is not sufficient to substantiate a claim related to improvement in a patient's ability to function or the patient's psychological state. Rather, to substantiate such a general claim, a sponsor should develop evidence to show not only a change in symptoms, but also how that change translates into other specific endpoints such as ability to perform activities of daily living, or improved psychological state. Accordingly, many PRO instruments are specifically designed to assess both symptoms and other possible consequences of treatment.

Comment: Clarification is needed regarding the extent to which BOTH symptoms and consequences must be incorporated into a SINGLE PRO, versus the use of separate PROs – for example one to assess changes in symptom presentation/severity and a separate one to assess impact on functional ability or psychological state. If an established PRO that measures *generalized functional improvement* (e.g., SF-36, WSAS, Q-LES-Q) related to symptomatic improvement across multiple disorders (e.g., depression, anxiety disorders, OCD, alcohol abuse/dependence, etc.), can that functional improvement PRO be combined with a *validated PRO assessment of symptoms* related to a different disorder to support a generalized claim of functional improvement related to treatment efficacy if BOTH the symptom and functional PRO assessments support significant improvement relative to placebo or other appropriate control condition?

Also see comments to P22, line 667-668 below.

P3, line 99-101

Systematic assessment of the patient's perspective may provide valuable information that can be lost when that perspective is filtered through a clinician's evaluation of the patient's response to clinical interview questions.

Comment: Recognition of patients' perspectives as separate and distinct from clinician's perspectives is welcomed. It should be pointed out that motivations for participating in clinical research and vested interests in the outcomes of said research also differ between patients and clinicians, and may influence the validity of measurements obtained from each. While patients have a primary motivation to alleviate personal suffering, clinicians are offered monetarily incentives to complete studies quickly and to maximize separation sensitivity between treatments. Such incentives can influence clinical judgment about inclusion severity criteria, as well as interpretation of patient-reported symptoms after treatment due to functional unblinding of clinicians knowledgeable about the typical side-effect profiles of different classes of drugs. The relative naiveté of patients providing self-reported personal outcomes without regard or interest for the overall success of a particular trial may provide a more objective scientific measure.

Also see comments to P23 line 725-737 and P25-26 line 824-835 below.

P4, line 127-132

Self-completed questionnaires that are given directly to patients without the intervention of clinicians are often preferable to the clinician-administered interview and rating. Self-completed questionnaires capture directly the patient's perceived response to treatment, without a third party's interpretation, and may be more reliable than observer-reported measures because they are not affected by interobserver variability.

Comment 1: We agree with the statement, and would like the statement to acknowledge that inter-observer variability cannot EVER be eliminated, and that extensive training has been shown to have minimal impact of rater reliability (Demitrack MA, Fries D, Herrera JM, et al.: The problem of measurement error in multisite clinical trials. Psychopharmacol Bull 1998; 34:19-24).

Comment 2: It should further be pointed out that when PROs are developed to replace existing rater-based assessments (e.g., HAMD, QIDS, MADRS), if convergent validity is to be established – see below – then the inter-rater reliability of the instrument provides the upper bound for the validity coefficient.

P4, line 134-137

Despite these concerns, well-developed and adequately validated PRO instruments have been shown to give answers that match the results obtained by the most expert assessors (indeed, that is the usual way their validity is assessed), and they appear to be particularly suitable in studies involving many investigators.

Comment: When PRO instruments are developed to be equivalent replacements for existing assessment, then the guidance should make it clear that patient participation in item generation is inappropriate. Patient debriefings of item understanding, and their ability to express their feelings fully can still be obtained, but assessment instruments with existing, known items and scoring conventions should not be re-engineered (adding or subtracting items, changing the item score ranges) based on open-ended patient focus groups as suggested in Section IV.B.1.

P5, Table 1

Under “Modes of data collection” Interviewer-administered is identified as a type of PRO.

Comment: To the extent that modifications can and are being made to the ways in which clinical interviews are being administered (e.g., use of telecommunication technology to conduct remote assessments with centralized raters), will the same validation criteria be applied to these methods as those applied to a “new instrument” (lines 178-179)? Additionally, clarification of the implications for investigator site responsibilities for controlling and keeping source data (P 25, lines 815 – 836) should be clarified.

P6, line 176-179

A new PRO instrument can be developed or an existing instrument can be modified if sponsors determine that none is available, adequate, or applicable to their product development program. When considering an instrument that has been modified from the original, the FDA generally plans to evaluate the modified instrument just as it would a new one.

P20, line 581-583

When a PRO instrument is modified, additional validation studies may be needed to confirm the adequacy of the modified instrument's measurement properties. The extent of additional validation recommended depends on the type of modification made.

P20, line 590-591

The FDA intends to consider a modified instrument as a different instrument from the original and will consider measurement properties to be version-specific.

Comment 1: Greater clarification is needed regarding the extent to which “modification” of an existing instrument creates a new instrument. Under strict interpretation of this guidance, the correction of a typographical error, change in a typeface or interviewer voice, or application of an adaptive computer interface that dynamically adjusts input/output based on user responses would become a “new instrument” with each administration. If “additional/new” validation studies are required under every possible scenario that an instrument might be applied to, new instrument development will become excessively burdensome and methodological innovation will cease. New instrument development will be caught in a never-ending loop of evaluation and revision, as suggested by the circular path of Figure 1, without ever settling upon a PRO assessment useful in clinical trials. Such rigorous attention to the minutest of details, if applied to the use of clinical raters taking into consideration each personal characteristic that makes one rater distinctly different than any other rater, would require a complete validation study for EVERY rater to be used in a clinical trial and that raters remain perfectly constant in the conduct of ratings, another practical impossibility.

Comment 2: To the extent that a modality or context of item presentation may influence the reliability and validity of the data obtained, if the EXACT same questions and response options result in systematically different reports obtained from the users, wouldn't the validity of BOTH the existing (old instrument) and the new implementation (modification) be questionable?

More concretely, if a questionnaire asks the question

“Can you walk up one flight of stairs without stopping or becoming fatigued? Mark your response here [] yes [] no”

and the responses to this question were systematically different than responses to the question

“Can you walk up ONE FLIGHT of stairs WITHOUT stopping or becoming fatigued? Mark your response here [] YES [] NO” Wouldn't the validity of this question in either form be questionable for assessing the functional ability to walk up

one flight of stairs? – Regardless of which form of the question was originally developed?

P8, line 223-225

Evidence from the patient cognitive debriefing studies (i.e., the interview schedule, transcript, and listing of all concepts elicited by a single item) can be used to determine when a concept is adequately captured by a single item.

Comment 1: What criteria will be applied to cognitive debriefing transcripts to determine whether a single item adequately measures a complete concept that can be used to support a labeling claim?

Comment 2: What criteria will be used to determine the adequacy and breadth of the cognitive debriefing?

Comment 3: Can the same modality of the PRO be used to obtain cognitive debriefing data? For example, following delivery of a PRO instrument on the web or via IVRS, would data obtained immediately following the assessment using the same medium provide adequate cognitive debriefing validation?

For example, could patient answers to questions like “Were you able to understand the questions easily?” “Were there enough options for providing your answer to allow you to completely describe how you were feeling?” “Are there other aspects of your experience that are important but we did not ask about?” in the same modality as the initial PRO be used as adequate cognitive debriefing data? Open-ended questions and responses could be typed in by the patients or recorded over the telephone and provided in transcript form. This comment also relates to the guidance provided on P10, line 332-334 below.

P10, line 295-297

PRO instrument item generation is incomplete without patient involvement. Item generation generally incorporates the input of a wide range of patients with the condition of interest to represent appropriate variations in severity and in population characteristics such as age or sex.

P16, Table 4. FDA Review Considerations of Validity

Have patients similar to those participating in the clinical trial confirmed the completeness and relevance of all items?

Comment 1: If PRO items were adapted from an existing instrument with established items, such as the HAMD or QIDS, it wouldn't be inappropriate to change the number of items that comprise the instrument. Patients' understanding of items can be established and perceptual comparability between modalities for instrument delivery examined, but the guidance should clarify that this section does not suggest that existing, established items that define current instruments should be modified based on patient input.

Comment 2: If a PRO instrument is intended to capture symptoms that are directly related to diagnostic criteria defined by DSM, patients' clinical understanding of disorders may differ from diagnostic reference use by clinicians. Shouldn't established

clinical nosology be weighted more heavily for item selection and weighting than patients' nonclinical perspectives? See also Comment 4 below.

Comment 3: What criteria will be used to determine the adequacy of disease severity and population demographics representation?

Comment 4: Content validity is often established by review by a panel of experts to ensure the instrument content is appropriate and representative (APA, AERA, NCME. Standards for Educational and Psychological Testing; 1999; AERA: Washington, DC). Particularly for instruments assessing DSM-IV-related symptoms, should not the establishment content validity include a review of items by a panel of experts in that particular disorder to ensure adequate coverage of symptoms?

P10, line 322-325

The FDA intends to review the comparability of data obtained when using multiple modes of administration to determine whether pooling of results from the multiple modes is appropriate.

Comment 1: On what basis will multimodal comparability be assessed?

Comment 2: If each new mode is considered a "new instrument" requiring independent validation (presumably showing convergent validity with established measures), is this a separate validation criterion?

P10, line 332-334

When evaluating PRO-based claims, the FDA intends to review the study protocol to determine what steps were taken to ensure that patients understand the appropriate recall period.

P11, line 339-343

PRO instruments that require patients to rely on memory, especially if they must recall over a period of time, or to average their response over a period of time may threaten the accuracy of the PRO data. It is usually better to construct items that ask patients to describe their current state than to ask them to compare their current state with an earlier period or to attempt to average their experiences over a period of time.

Comment 1: Prior research has established that retrospective recall indicating clinical change can be superior to evaluating differences between serial point-specific severity measures (Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. JAMA 1999; 282:1157-1162). When problems arise it is often due to the adequacy of the memory for the prior clinical state. Procedures are currently being developed to specifically aid patients' memories of pretreatment status in order to promote more accurate and sensitive retrospective ratings. Initial results suggest these methods are superior to prior retrospective PRO instruments (Mundt JC, DeBrotta DJ, Moore HK, & Greist JH. Memory Enhanced Retrospective Evaluation of Treatment (MERET): Anchoring Patients' Perceptions of Clinical Change in the past. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 45th Annual Meeting, Boca Raton, FL. June, 2005).

Other “state-of-the-art” retrospective PRO instruments such as timeline follow back (TLFB) procedures use memory aides such as calendars to promote accurate retrospective data reports. FDA guidance should reflect that methods and techniques used to enhance the reliability and validity of retrospectively reported data exist, and the use of such techniques should be encouraged or mandated when retrospective data are to be collected.

Comment 2: Clarification is needed for validation of existing clinical instruments, such as the HAMA, HAMD, QIDS, that ask clinicians to assess patients experiences over a period of time (typically one week). Again, while recent data suggests the clinical raters are not effective at doing so (Mundt JC, Moore HK, DeBrotta DJ, & Greist JH. Recency Effects in Standard Depression Measures Using Daily Telephone Assessment Ratings. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 45th Annual Meeting, Boca Raton, FL. June, 2005), if ePRO instruments are to be validated against existing standards, the instructions for both methods should remain invariant. The more fundamental question is: To what degree will validation guidelines being established for PROs also be applied to assessments obtained by clinical raters?

P10-11, line 334-337

If a patient diary or some other form of unsupervised data entry is used, the FDA plans to review the protocol to determine what measures are taken to ensure that patients make entries according to the study design and not, for example, just before a clinic visit when their reports will be collected.

Comment 1: Does this guidance imply that paper-pencil methods for obtaining diary data will require an external device, such as an electronic pen, photoelectric eye, postmarked envelopes, or other device, to verify the date/time of data registration by patients?

Comment 2: This guidance implicitly acknowledges that some data can ONLY be obtained in an “unsupervised” environment outside of clinic visits. The implication of this is that for ANY unsupervised data collection (via paper or other means) there is always a period of time for which the site investigators cannot meet the regulatory requirements for storage and verification of record accuracy of source data (see P. 25-26, line 824-832). Under this acknowledgement, isn’t the difference between unsupervised completion of paper document returned to the investigator (if not lost by the patient) and providing of electronic PRO records to investigators by 3rd party technology vendors equivalent?

Comment 3: When investigators send x-ray, EKG, EEG, or blood data out for scoring to obtain clinical reports, isn’t this equivalent to assessment reports being provided by technology vendors based on direct patient interaction with web, palm, or telephonic data collection systems?

P12, line 373-378

Sponsors are encouraged to examine the procedures used with patients to determine readability and understanding of the items included in the PRO instrument. The FDA’s evaluation of these

procedures is likely to include a review of a cognitive debriefing report containing the readability test used, the script used in patient cognitive debriefing interviews, the transcript of the interviews, the analysis of the interview results, and the actions taken to delete or modify an item in response to the cognitive debriefing interview or pilot test results.

Comments are essentially the same as those expressed previously to P8, line 223-225:
What criteria will be applied to cognitive debriefing transcripts? What criteria will be used to determine the adequacy and breadth of the cognitive debriefing? Can the same modality of the PRO be used to obtain cognitive debriefing data?

P12-13, line 382 - 403

PRO study results can vary according to the instructions to patients or the training given to the interviewer or persons supervising PRO data collection. Sponsors should consider all PRO instrument instructions and procedures contained in publications and user manuals provided by developers, including procedures for reviewing completed questionnaires and re-administration to avoid missing data or clarify responses. Other important considerations include the format of the questionnaire, the final wording of PRO instruments as implemented in clinical trials, and any potentially important changes in presentation or format. Examples of changes that can alter the way that patients respond to the same set of questions include:

- Changing an instrument from paper to electronic format
- Changing the timing of or procedures for PRO instrument administration within the clinic visit
- Changing the order of items or deleting portions of a questionnaire
- Changing the instructions or the placement of instructions within the PRO instrument

It is important that the PRO instrument format used in the clinical trial be consistent with the format that is used in the instrument validation process. *Format* refers to the exact appearance of the instrument. Instrument format is specific to the mode of administration, including paper and pencil, interviewer-administered or supervised, or electronic data collection. The FDA plans to review the PRO instrument in the format used in the clinical trial case report forms, including the order and numbering of items, the presentation of response options in single response or grid formats, the grouping of items, patterns for skipping questions that are not applicable, and all instructions to patients in the interview schedule or on the questionnaire.

Comment 1: Concerns are similar to those expressed above in reference to P6, line 176-179. Clarification is needed to define what changes are “important” versus those that are essentially cosmetic/inconsequential. Under strict interpretation of this section any change to any aspect of any instrument could cascade into an endless loop of instrument validation that will result in a stagnation of instrument development. Taken literally, Figure 1 on page 7 would suggest that no PRO is ever completed and usable, but a never-ending reiterative developmental process. Additionally, the extent to which a modality or context of item presentation influences the reliability and validity of the data obtained with the EXACT same questions and response options undermines the conceptual

validity of the item presented under either modality or context regardless of which was developed initially.

Comment 2: The number of *Format* variations for presentation of a PRO by interviewers is likely equal to the number of interviewers used in a study, and such format likely changes from one interviewee to the next. An acknowledgement of different standards for procedural reliability of PRO data collection methods compared to those applied to clinical ratings should be made. In terms of reliability of question stimulus, electronic presentation of questions is dramatically more reliable than presentation by human interviewers.

P19, line 543 – 564

“For many widely used measures (pain, treadmill distance, HamD), the ability to show *any* difference between treatment groups has been considered evidence of a relevant treatment effect. If PRO instruments are to be considered more sensitive than past measures, it can be useful to specify a minimum important difference (MID) as a benchmark for interpreting mean differences. An MID is usually specific to the population under study.

The FDA has reviewed MIDs derived in many ways. Examples include:

- Mapping changes in PRO scores to clinically relevant and important changes in non-PRO measures of treatment outcome in the condition of interest (e.g., when PRO measures of asthma or COPD are mapped to spirometry scores).
- Mapping changes in PRO scores to other PRO scores to arrive at an MID that is appreciable to patients (e.g., when multi-item PROs are mapped to a single question asking the patient to rate his or her global impression of change since the start of treatment). A problem with this approach is that it uses individual rates to reach a conclusion about mean effects. It may be more useful to look at the distribution of individual effects in treatment and control groups.
- Using a distribution-based approach (e.g., defining the MID as 0.5 times the standard deviation). This, of course, may bear no relation to the patient’s assessment and is usually inadequate in isolation.
- Using an empirical rule (e.g., 8 percent of the theoretical range of scores). Again, this arbitrary approach does not take into account patient preferences or assessment.

If a MID is to be applied to clinical study results, it is generally helpful to use a variety of methods to discover whether concordance among methods confirms the choice of an MID.

The FDA is specifically asking for comment on the need for, and appropriate standards for minimum important difference(MID) definitions applied to PRO instruments used in clinical studies.”

“The FDA is specifically requesting comment on appropriate review of derivation and application of an MID in the clinical trial setting.”

“The FDA is specifically asking for comment on the appropriate review standards for the definition of a responder when applied to PRO instruments used in clinical studies to support medical product “

Comment: These concerns appear to be deferential to non-PRO measures (e.g., clinical raters), setting up different standards for evaluating data collected by different methods. Given the suggested requirements to establish reliability/validity of PRO measures and equivalence between alternate forms, it is not clear why additional evaluation criteria are being specifically applied to PROs that are not applied to non-PRO measures.

P 21, line 646-648

The FDA recommends that sponsors provide evidence that the methods and results of the translation process were adequate to ensure that the validity of the responses is not affected.

P 22, line 660

The evidence that measurement properties for translated versions are comparable

Comment 1: Does every language translation need to be independently validated in separate studies? Clarification is needed to specify the criteria for evaluating the evidence that measurement properties are ‘comparable.’

Comment 2: How do these requirements for PRO translations comport with those made of translations employed by human interviewers?

P22, line 667-668

A PRO instrument developed and previously used as a stand-alone assessment is included as a part of a battery of measures

Comment 1: It is impractical to validate every PRO instrument for every possible combination of other assessments that a given PRO might be included with.

Comment 2: If “stand-alone” PRO measures of current psychological state or functional abilities are interdependent upon concurrent PRO measures, can such a “stand-alone” instrument EVER be a reliable and valid measure of a stable construct? If data from a PRO were dependent upon questions or assessments completed just before the PRO, the scientific evidence would indicate that the PRO is an assessment of a manipulable state that cannot be a stable (no test-retest reliability) index of treatment efficacy.

Comment 3: See comments to guidance statements on P3, line 83-89 above.

P 22-23, line 694-699

Over the course of some clinical trials, it can be anticipated that patients may become too ill to complete a questionnaire or to respond to an interviewer. In such cases, proxy reporting may help to prevent missing data. When this situation is anticipated, the FDA encourages the inclusion of proxy reports in parallel with patient self-report from the beginning of the study (i.e., even before the patient is no longer able to answer independently) so that the relationship between the patient reports and the proxy reports can be assessed.

Comment: If convergent validity is established between concurrent PRO and proxy report, will proxy reports be acceptable as primary outcome or be useable to impute missing PRO data, should patient become too ill to respond or effectively use PRO technology?

P23, line 725-732

The importance of blinding can be determined, in part, by the characteristics of the PRO instrument used. For example, questions that ask how patients' current status compares to baseline seem likely to be more influenced by unblinding (optimism can readily be expressed as a favorable comparison) than questions that ask about current status (which requires a current assessment, not a statement about duration). Questions that ask for current status, or PRO instruments that ask many questions, are harder to answer in a biased way when previous answers are not available. For the same reasons, allowing patients access to previous responses can bias results when unblinding is a possibility.

P23, line 735-737

There are certain situations, particularly in the development of medical devices, where blinding is not feasible and other situations where there is no reasonable control group (and therefore no randomization).

Comment 1: The importance of treatment blinding is critical for obtaining unbiased measures, however we are not aware of data suggesting that asking patients for retrospective judgments relative to baseline compromises the treatment blind. Clinicians are often asked to provide ratings of global impressions of improvement and typically have access to patient data and records collected previously. Clinicians are also more likely to see through treatment blinds if the compound under investigation belongs to a class of drugs with known, recognizable side effect profiles. Clinical raters have a vested interest in "optimistic" outcomes, and multiple studies have found clinical raters consistently provide data with greater change from baseline FOR ALL TREATMENT GROUPS, INCLUDING PLACEBO than PROs. Such results have led some researchers to suggest that the difference between PROs and clinical rater data is actually an index of rater bias (Petkova E, Quitkin FM, McGrath PJ, Stewart JW, & Klein DF. A method to quantify rater bias in antidepressant trials. *Neuropsychopharmacology*. 2000; 22(6):559-565).

Comment 2: Referenced above in the comments to P10 line 332-334 and P11 line 339-343, techniques for anchoring patients' recollections of baseline states are being developed to enhance the precision of retrospective ratings. The evidence gathered to date suggests that these techniques do improve patients' perceptions of improvement *when given active treatments with established efficacy, but that patients' perceptions of improvement in placebo conditions remain unaffected*. Such data would contradict the predictions inferred by the guidance commentary above.

Comment 3: The suggestion that questions asking about current status are less likely to produce biased responses than when previous answers (and experiential anchors

underlying those answers) should be supported empirically. Experiential habituation and physiological adaptation to repeated exposure to environments and circumstances are known to have powerful effects on the subjective experiences of patients. For example, a depressed patient may rate a day that they were able to get out of bed and not have multiple crying spells as a “pretty good” day prior to treatment; after receiving effective treatment for a period of time, the very same experiences may be at the “pretty bad” end of recent experiences. Such habituation and adaptation may produce a relative insensitivity for PROs to establish the effectiveness of treatments that may emerge over extended periods of time. Allowing patients to better recall baseline and pretreatment experiences would facilitate more accurate judgments as data reflecting clinical change, rather than promote optimistic or “wished for” responses.

P25-26, line 824-835

The principal record keeping requirements for clinical investigators include the preparation and maintenance of adequate and accurate case histories (including the case report forms and supporting data), record retention, and provision for the FDA to access, copy, and verify records (i.e., source data verification). The investigator’s responsibility to control, access, and source documentation can be satisfied easily when paper PRO instruments are used, because the subject usually returns the diary to the investigator who either retains the original or a certified copy as part of the case history. The use of electronic PRO instruments, however, may pose a problem if direct control over source data is maintained by the sponsor or the contract research organization and not by the clinical investigator.

Comment 1: Isn’t ANY unsupervised data collection, via paper or electronic means outside of the investigator office, outside investigators’ ability to maintain control, access, or verification of the source data until delivered to them? Certified copies of electronic records, verifiable through the use of appropriate audit trails and other control processes, can be provided by third party vendors who have the same legal and professional relationship with sponsors.

Comment 2: Clinical investigators have vested financial interests in study outcomes that may bias source data. Data demonstrating screening and baseline inflation to facilitate patient enrollment has been demonstrated in multiple studies. Investigator expectations and motivations to separate drug from placebo (sites are aware that sponsors compare this performance between sites and are more likely to engage those that do in future studies) may also influence treatment response measures. Patients have less knowledge about inclusion/exclusion criteria and/or commonly expected side-effect profiles, which may influence clinicians’ interpretation of symptoms reported by patients. Use of independent PRO assessments with separate, independent record keeping promotes scientific integrity and objectivity.

Comment 3: The time lag between patient entry of data into a PRO and access by investigators is the same problem that exists with paper diaries being brought in by patients at site visits.