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]

Executive Summary

The following document represents the commentary and response of the International Society for Quality of Life Research (ISOQOL), an organization which has as a central part of its mission the scientific development and evaluation of patient-reported outcome instruments.

The FDA is to be commended for completing the draft guidance for which the field of outcomes research has been waiting. In general, the guidance is well-done, and will be a useful resource for researchers, sponsors, and the FDA in evaluating data based on PROs. The danger inherent in the development of any guidance, and this one is no exception, lies in being overly prescriptive about the scientific underpinnings of research in this area. As in any field, it is the consensus of experts as applied to a particular measurement context that should be ultimately the most important consideration in evaluating the validity and interpretability of data. There are many solid recommendations contained within this guidance; the FDA has done a very good job in putting this together. However, there are also instances, which we have outlined below, where bold assertions are made as to how something should be done which: (a) either are lacking in empirical support, (b) inconsistent with established practice in the field, or (c) may have some scientific merit yet would be prohibitively impractical for sponsors of research to implement. It is our fervent hope that in such instances, the guidance will not supersede the counsel of independent experts in the field. It is equally our hope that the fruitful dialogue between the FDA and experts in the measurement of patient reported outcomes, which this guidance has promoted, will continue into the foreseeable future.

These comments were assembled from the ISOQOL Task Force on PROs and Regulatory Issues. Dennis Revicki serves as chair of this task force, with members including Neil Aaronson, Jakob Bjorner, David Cella, Diane Fairclough, Peter Fayers, Ron Hays, and William Lenderking.

RESPONSE/COMMENTS AREAS FOR FDA GUIDANCE

General Areas for Commentary

1. The treatment of content/face validity versus other psychometric characteristics (existing, well researched measures versus recently developed measures) is an issue. There is an implication that qualitative research (focus groups, cognitive interviewing, etc.) is more important than quantitative research and evidence supporting reliability, construct validity and responsiveness, which is not supported by data or practice standards.
2. There is an uncompromising tone to Section IV compared to more open tone of Sections V – VI. Section IV presumes evidence base not currently in existence and renders policy from unsubstantiated opinions. We suggest that you encourage more gathering of evidence and flexibility in evaluation of existing approaches that have no known superior alternative.
3. We understand that the FDA's focus is directly on the label and any statements within the label relating to PRO endpoints. While this focus is important, it appears that this may limit content on what is important to the patient. For example, if patients produce a list of issues/items that do not fit into a framework fit for labeling (based on FDA review), then this information may not be included in the label, and issues outside acceptable label claim must be removed. This makes it difficult to keep it "about the patient" and to provide all the PRO information that may be meaningful to patients, their families and their clinicians.
4. We have commented on methods for determining responsiveness and MID (see #1 above). See attached document.
5. PRO statistical analysis issues are described below (i.e., multiplicity, missing data, sensitivity analyses, etc.).
6. Open-label (un-blinded studies) and PRO endpoints (issues of bias, etc.) are discussed below.
7. We provide some observations about validation studies (i.e., generalizable populations (specific versus general), large versus small studies, new versus established PRO measures, etc.).
8. We discuss recall period-and related issues. There is evidence that patient reports of a common domain (e.g., pain intensity) vary by recall period, but there is little evidence that this recall bias impacts on detecting treatment differences (for a treatment with evidence of effectiveness). The recall bias may add additional measurement error and noise to the assessment, but this measurement error may not be large enough to mask detection of treatment differences. Furthermore, it would be to the sponsor's disadvantage to use an instrument that has this type of recall bias, making it more difficult to detect treatment differences.

Specific Major Areas for Commentary

1. Glossary: The glossary is very useful. However, some crucial terms relating to FDA recommendations are not defined in the document (e.g., side effects, item generation, etc.). They should be defined in the glossary.

2. Lines 339-343: Standard questionnaires versus diaries: Some remarks suggest a recommendation of diaries over standard questionnaires. Is there sufficient empirical data to make such a recommendation? While diary data may be sufficient for assessing

domains or symptoms that are more transitory (i.e., acute pain, sleep disturbance), standard questionnaires are more relevant for assessing domains (i.e., physical function, social activities, functional outcomes, etc.) that require integration of experience and events over longer periods of time.

3. Context effects: In numerous places, the guidelines assume that context effects have a substantial impact on questionnaire responses and thus revalidation is required with any change in context. This is not supported by data and is an impractical recommendation.

4. Generalizability across populations: The guidelines appear to assume that it is never possible to generalize results from one population or context to another. Generic health concepts are more tied to the general human condition than to the specific condition a given human has, and are generalizable across conditions. There are also distinct advantages to working with cross-cutting generic concepts when one wishes to compare relative effectiveness across treatments and diseases on a level playing field. We consider cultural validation to be essential, however, there is good evidence for domains included in generic instruments to be cross- culturally relevant. Even the WHOQOL instrument, which had a simultaneous cross-cultural approach in the construction, identified common domains.

5. Lines 153-156: *"Some PRO instruments (e.g., health-related quality of life instruments) attempt to measure both the effectiveness and the side effects of treatment. PRO instruments that are used in clinical trials to support effectiveness claims should measure the adverse consequences of treatment separately from the effectiveness of treatment."* This is not possible in practice for many conditions, since treatment side effects often overlap with condition-specific symptoms (e.g., fatigue in chronic hepatitis C). Separating them is artificial and misleading…an example of the problem one inherits when buying into the hegemony of the label. This needs to be clarified or revised. Side effects may ultimately affect the same domains as the treatment effect. Do they mean that researchers should always make sure that the instruments cover potential side effects? Should the assessment be able to separate the treatment effect from the unintended effect? (difficult, our instruments do not evaluate causality – neither can a clinician in many cases). It is not always possible to assess adverse effects of therapy separately from effectiveness of treatment. This requires attributions by patients which may not be possible. Rather, this is a question of study design. For example, patients may not be able to distinguish between treatment-induced and disease-related fatigue, but the trial design can help clarify this.

Suggested revision:
"Some PRO instruments (e.g., health-related quality of life instruments) attempt to measure both the effectiveness and the side effects of treatment. Clinical trials to support effectiveness claims should be designed to provide information about the adverse consequences of treatment separately from the effectiveness of treatment, when it is possible to separate them."

Separation of side effects from treatment effects is an important topic that extends beyond a particular PRO instrument. Steps to separate these effects should be taken in various parts of the study design (e.g. frequency of assessment of side effects, efficient reporting mechanism). Further, the original sentence could be misunderstood to mean that the separation of effectiveness from side effects could only be achieved by a PRO instrument measuring these separately and not by combining an instrument aimed at evaluating effectiveness with an instrument aimed at evaluating side effects.

6. Lines 178-181: *"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. Therefore, in such instances, we encourage sponsors to document the original development processes, all modifications made, and updated assessments of its measurement properties."*

Suggested revision:
"When considering an instrument that has been modified from the original, the requirements for documentation will depend on the extent and nature of the modification and the level of documentation of the original instrument. Therefore, in such instances, we encourage sponsors to document the original development processes, all modifications made, and updated assessments of its measurement properties."

The original statement appears to contradict section D line 582-583: *The extent of additional validation recommended depends on the type of modification made.* The extensive definition of modification in section D also renders excessive the plan to evaluate modified instruments just as new ones.

7. Lines 214-223: *"If the concept of interest is general (e.g., physical function), a single-item PRO instrument is usually unable to provide a complete understanding of the treatment's effect because a single item cannot capture all the domains of the general concept. For this reason, single-item questions about general concepts that imply multiple domains rarely provide sufficient evidence to support claims about that general concept. However, single-item questions about general concepts can be useful to help interpret multi-item measures of the same concept and to determine whether important items or domains of a general concept are missing (e.g., when results using single general questions do not correlate with results using a multi-item questionnaire, this may be evidence that the questionnaire is not capturing all the important domains of the concept contained in the claim)."*

Suggested revision:
"If the concept of interest is general (e.g., physical function), a single-item PRO instrument is usually unable to provide a complete understanding of the treatment's effect because a single item cannot capture all the domains of the general concept.
For this reason, single-item questions about general concepts that imply multiple domains rarely provide sufficient evidence to support claims about that general concept. However, single-item questions about general concepts can be valid measures of the

patient's own overall assessment of the concept. Thus, care should be taken in the labeling of the concept."

The original text could be confusing: if a single-item measure does not capture all domains of a concept, it is hard to understand why a low correlation with results from a multi-item questionnaire should be taken as evidence that the questionnaire is not capturing all domains. In the instance of a low correlation between single and multi-item measures of a concept, in the absence of other evidence to support the validity of either, it is difficult to say whether it is the single or multi-item instrument that is lacking. The part of the multi-item questionnaire that shows low correlation with the single item, could be the very part this is relevant but not captured by the single item. On the other hand, global single items can be valid measures of the patient's global assessment of a health domain. One example of this is the strong body of evidence supporting the validity of single items on general health perception (Fayers & Sprangers, 2002)

8. Lines 227-237: The text here is rather dense. Isn't this really an issue of having measures that allow one to aggregate and disaggregate the data (capturing on one level, what the FDA terms "domains"; and also allowing one to aggregate scores to capture at another level what the FDA terms "concepts"? Good HRQL measures capture both the positive and negative impacts of interventions on patient functioning and well-being.

9. Lines 462-467: Confirmation of conceptual framework: What evidence is required to support the conceptual framework of the final instrument? The distinction between "framework" and "model" is not clear. "Framework" is used throughout the text in ways completely consistent with "model."

10. Line 481: Table 4: Who decides whether content validity is demonstrated? It appears that there will be a lot of subjective judgment possible here. The FDA guidance document needs to recognize that while qualitative research (i.e., cognitive interviewing and debriefing, patient focus groups, etc.) is useful for developing an understanding of symptoms, problems and domains from the perspective of the patients, and to examine respondent understanding and comprehension of question stems and response scales, instructions, etc., there is a lot of subjective judgment that goes into reviewing and interpreting the data from qualitative studies.

11. Ability to detect change is an aspect of validity (page 17). It is not clear how "effect size and standard error of measurement" are used as a test of ability to detect change. Ability to detect change requires independent information about true change. It is also unclear why the sentence, "Whether PRO scores are stable when there is no change in the patient" is included in the ability to detect change row when the former refers to reliability. The statement about "based on experience with the measure using a distribution-based approach" is unclear. Distribution-based indices are simply ways to express change in a standardized metric (e.g., effect size).

The responder definition needs to be clarified in the table (a little better job is done in the text). Responders can be defined by statistical significance of individual change--see,

e.g., Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K. (2005). Evaluating the statistical significance of health-related quality of life change in individual patients. Evaluation and the Health Professions, 28, 160-171.

12. Lines 288-300: *"PRO instrument item generation is incomplete without patient involvement."*

Suggested revision:
Item generation includes establishing the content to be covered by the items, generating item wording, evaluating the completeness of item coverage, performing initial assessment of clarity and readability. PRO instrument item generation is potentially incomplete without patient involvement.

The problem is that item generation is not clearly defined. Is item generation: 1) Generation of wording (only); 2) Establishing content; 3) Evaluating completeness of coverage; 4) Evaluation of the measurement properties of each item? We agree that patient input during item generation is important to ensure that what is being asked represents the range of issues important to the patient group (content validity), and that input about item clarity and consistency to a shared concept. A clear definition of the term *item generation* would be helpful in clarifying this section and such definition should also be put in the glossary. The text above is an attempt to provide such a definition. The original text could be interpreted to mean that patients should generate the item wording. However, if item wording comes directly from patient statements, considerable editing is often necessary to make the wording clear and unambiguous in a questionnaire context. Thus patient involvement in evaluating completeness of content coverage and item clarity and readability should be sufficient. It is also the case that many instruments have historically been developed based on expert opinion, without much patient input. Such scales are still in use, because empirical evidence has suggested over time, that they perform well.

13. Lines 302-308: *"Items that ask patients to respond hypothetically or that give patients the opportunity to respond on the basis of their desired condition rather than on their actual condition are not recommended. For example, in assessing the concept performance of daily activities, it is more appropriate to ask whether or not the respondent performs specific activities (and if so, with how much difficulty) than whether or not he or she can perform daily activities (because patients may report they are able to perform a task even when they never do so). Of course, it would be critical to know that each item refers to something that patients actually do."* There is not good evidence for this recommendation. Instruments that ask only what the respondent has done, will have huge ceiling effects as measures of physical functioning. If one takes this viewpoint seriously, then measures such as the physical functioning scale of the SF-36 or of the EORTC QLQ-C30 (and other questionnaire's scales as well) would be disqualified, as both assess what the respondent is able to do (or not do), not what s/he has actually done. Ability is the critical element being measured.

14. 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. "*

Suggested change
"For PRO instruments that require patients to rely on memory, care must be taken to select an appropriate recall period. Long recall periods for everyday symptoms may threaten the accuracy of the PRO data. It is usually better to construct items that ask patients to describe their current state than ask them to compare their current state with an earlier period."

Again, this is not clear and evidence in support of the position is inconsistent and often absent in many clinical areas. Almost any response process relies on cognitive processing and to some extent memory. What is a period of time? Minute? Hour? Day? Week? This could be read as a recommendation of diaries over standard questionnaires. What is an appropriate recall period? For example, it is very common to ask patients to report their symptom experience or level of functioning during the previous week. For many patients, this may be interpreted as an "average" effect; for some, however, patients may provide a response based on their worst (or best) period during the week. Recall periods must fit the domain of interest and no one recall period can fit all patient reported outcome domains. It is more important to provide a clear rationale for selecting a particular recall period, and it is likely that even with some recall bias (which adds measurement error), this bias may not be large enough to mask treatment effects (for a treatment with evidence of clinical efficacy). Is it better to ask people to recall the more salient "worst" or "best" of a symptom, as opposed to the more challenging task of remembering the "average" if the recall period is greater that one day? There is some empirical evidence supporting that respondents are better at remembering "worst" than "average. However there is some research that finds that this is not always true. These statements about recall period represent one of several opinions that is at risk for becoming fiat based on guidance wording. Therefore, we recommend that section IV be revised to read more like sections V-VI, offering general recommendations but not prescribing the right and wrong ways to assess patient reported outcomes.

The implicit assumption in the original text that cognitive processing will hamper validity and accuracy ignores the fact that any response process relies on cognitive processing and to some extent on memory. Even if the item refers to the "current state", patients have to define the time period that is referred to, and those definitions will vary. The original wording is too general; it does not clarify the term *"period of time"*, and can be read as a general recommendation of diaries over other standard PRO questionnaires. Both approaches have well described advantages and disadvantages.

15. Line 367-369: *"Response options do not bias the direction of responses (e.g., offering one negative choice, one neutral choice, and two or more positive choices on a scale makes it more likely for patients to respond that they feel or function better). "*

Suggested change:
"Response options do not bias the direction of responses (in a post-intervention evaluation e.g., offering one negative choice, one neutral choice, and two or more positive choices on a scale makes it more likely for patients to respond that they feel or function better). "

This is not correct. Not all response scales can be bidirectional and balanced (e.g., pain rating scales). How does the FDA expect the investigator to justify the number of response choices? In some types of PRO measures – for example, patient satisfaction questionnaires – it may be quite appropriate to use an asymmetrical scale in order to compensate for the tendency to score on the positive end of a scale (i.e., to elicit subtler degrees of dissatisfaction). In a design with baseline and follow-up assessment the issue of balanced response options is not an issue since real outcome is the difference between baseline and follow-up. Thus, the issue of balanced response scales is only relevant in designs that rely solely on post-intervention assessment, which is clearly a less advantageous method for assessing the effects of an intervention anyway, since it is retrospective.

16. Line 393: *"Changing the order of items or deleting portions of a questionnaire"*
The available evidence does not support this. Also, IRT provides a framework for appreciating the value of this approach and addressing concerns about context and local independence…this may be the place to introduce IRT.

17. Lines 397-398:
"Format refers to the exact appearance of the instrument."

Suggested revision:
"Format refers to the exact appearance of the questionnaire (or survey)."

Because the Glossary defines an "instrument" as being a means to capture data <u>plus</u> information that supports its use (e.g., documentation on scoring, analysis and interpretation of results), the word "instrument" may be too broad here, since materials such as user manuals and scoring documents generally are not seen by patients. Thus, the word "questionnaire" might be substituted. However, "questionnaire" as defined in the Glossary is limited to questions "shown to a respondent", which may be too narrow in the context of this paragraph. Another term (for example, "survey") might be added to the Glossary, to describe any means (e.g., questionnaire, diary, interview script) that is used to collect PRO data.

18. Line 416-422: *"A scoring algorithm creates a single score from multiple items. Equally weighted scores for each item are appropriate only when the responses to the items are relatively uncorrelated. Otherwise, the assignment of equal weights will overweight correlated items and underweight independent items. Even when items are*

uncorrelated, assigning equal weights to each item may overweight certain items if the number of response options or the values associated with response options varies by item. The same weighting concerns apply with added complexity when combining domain scores into a single overall score."

Suggested wording:
"When a scoring algorithm is used to create a single score from multiple items, care should be taken to ensure that psychometric requirements for multi-item scoring are fulfilled. Item selection and weighting should ensure that the score adequately represent the concept of interest. Equally weighted scores for each item is a widespread approach with considerable robustness. Care should be taken to document the validity of the instrument scoring method. Inclusion of items that are conceptually too similar can lead scores to be overly influenced by a single domain. Care should also be taken when creating scores from items where the number of response options or the values associated with response options varies by item. The same weighting concerns apply with added complexity when combining domain scores into a single overall score. Item response theory can be a useful tool in this context to establish the consistency of measurement across the range of the concept, and whether or not one end of the range is more or less heavily represented by multiple items."

The intent of the original statement is probably to avoid the conclusion that some domains are overly weighted in global indices by the inclusion of many similar items from the same domain. However, taken at face value the text is not correct according to psychometric research. Within item response theory, equal weighting of items is justified when items confirm to the so-called Rasch model. This is also true when items differ in their number of response choices (regardless of the sometimes heated debate on Rasch models versus other IRT models every IRT expert would agree with the previous statements). An implication of this is that items can be combined without weighting when they have similar item discrimination and when they fulfill requirements of unidimensionality and local independence. Even if items are strongly correlated an unweighted score is valid if these requirements are fulfilled. Further, a large body of research shows that equal weighting is robust to deviations from the requirement of equal item discrimination. Within classical psychometrics, equal weighting is justified when items have roughly equal variances and item-total correlations. However, here also items can be highly correlated. This revision is important since the original wording would erroneously indicate that the scoring of highly valid and reliable PROs was wrong.

19. Lines 417-418: *"Equally weighted scores for each item are appropriate only when the responses to the items are relatively uncorrelated."* This is not correct as stated. The FDA responses to questions about this at the Mayo meeting indicated that what they really meant is that redundant items should not be included in a domain because that would overweight the contribution of the content of that item to the scale score. Local independence is one of the assumptions in item response theory that can be tested to provide an indication of item redundancy. Advice on psychometric approaches to scaling should either be substantially revised or dropped. There is a large body of literature that suggests that weighting of items within scales does not add significantly to measurement

precision beyond an equal weighting strategy, or that it contributes significantly to improving the validity of measures. In addition, some questionnaires intentionally overweight content with items tapping one sub-concept over another. This can approximate a weighting scheme, when items are then scored equally in classical scoring mode. Why is this a problem? This should actually be encouraged if one wants content validity.

20. Table 4: *"Have patients similar to those participating in the clinical trial confirmed the completeness and relevance of all items?"*

Suggested revision
Is there empirical evidence supporting that the measurement range and item content are relevant for the patient group participating in the clinical trial and that the items cover all aspects of the domain in question?

This original demand seems excessive for generic instruments. If completeness and relevance have been demonstrated for many patient groups for generic concepts, reexamination for each patient group are not likely to add much. In particular, this will be a problem for an item bank based assessment, because the total of "all items" will be large. The original text seems to enforce particular approaches to assessing the completeness and relevance of all items with qualitative research, e.g. focus groups. While focus groups can be valuable in test development, the requirement that a similar group of patients should confirm the completeness and relevance of items every time a trial is launched seems counter-productive. For well-validated scales with a strong conceptual model, e.g. a scale of depression, should some items be dropped if one focus group questions their relevance? This could have dramatic consequences for the ability to compare results between studies and thus spoil the ability to interpret and generalize the results.

21. Lines 495-496: *"Internal consistency reliability, in the absence of test-retest reliability, does not generally constitute sufficient evidence of reliability for clinical trial purposes."*

Suggested revision:
"Internal consistency reliability for multi-item scale, in the absence of test-retest reliability, is sufficient evidence of reliability for clinical trial purposes only if the scoring assumptions have been carefully evaluated, and the domain in question does not appear to exhibit large day-to-day variations."

By focusing only on the concept of reliability (and not test precision/measurement error) and only on two ways of assessing reliability this recommendation becomes too narrow. It essentially boils down to a requirement of test-retest reliability for all purposes. Test-retest reliability is not the most important or only form of reliability estimation. Available evidence has shown that internal consistency reliability is a good estimator of reliability if the basic scoring assumptions are fulfilled. Domains with a considerable day-to-day variation will have less power in clinical trials (because of the larger variation in change scores) even if the instruments are reliable according to psychometric

definitions of reliability. This issue pertains to the domain in question, not to the measurement instrument.

22. Lines 547-548: *"An MID is usually specific to the population under study."*

Suggested revision:
An MID may differ between populations. Therefore, the use of an MID in a particular population should be justified. Once several populations have been studied and MIDs established, it may be sufficient to use a pre-existing MID.

This statement is too broad and not empirically justified. This would mean that no general score interpretations are possible. We agree that MIDs are not fixed properties of instruments. However, after an instrument has been studied in several settings and has produced a stable range, we do not think it necessary to keep testing for MIDs in a new population simply because it hasn't been one of the studied samples to date.

While the MID may very by population, much evidence suggests that MID for generic measures are probably fairly population independent. This particularly pertains to results from anchor-based analyses. Results on MID from consensus groups of clinicians have differed between different populations but also for different studies of the same population. The likely explanation is that the clinical consensus method is not very reliable and strongly depends on the information presented to the clinicians. While this topic is under-researched, the original statement is too strong given current evidence.

23. Section D: Modification of an Existing Instrument: We think it would be helpful to precisely define modification. The examples in the following section are very broad. The stance on a modified instrument leans towards overly conservative. And what is the problem with a "single domain from a multiple domain PRO ... administered without the other domains". The first bullet (administering a single domain from a multiple domain PRO without the other domains) should not be considered a major measurement violation necessitating revalidation (line 597). Needs more clarification as to how much modification of an instrument will require validation studies. However, sponsors may need to provide justification for deleting domains of a multiple domain PRO if a case could be made that the deleted domains might reveal adverse effects of a drug. This section needs more clarification as to how much modification of an instrument will require validation studies.

24. Lines 610-622: *"An instrument developed for use in one population or condition is used in a different patient population or condition. For example:*

- *Patients in the proposed trial have a disease, condition, or severity level that is different from that of the patient population used for instrument development and validation*
- *Patients in the proposed trial differ in age, gender, race, or developmental or life stage from those for instrument development and validation"*

This seems to define the idea of generic health status and HRQL instruments.

"An instrument is altered in item content or format. This includes changes in the following:

- *Number of items (more or fewer) used to assess a concept or domain*
- *Wording or placement of instructions*

Change in response options

- *Wording or order of the items"*

This makes computerized adaptive assessment, based on established item banks, impossible. It also makes customized short forms based on item banks unworkable. Given the NIH sponsored PROMIS project has as its primary objective the development and evaluation of item banks and computer adaptive testing for pain, fatigue, physical function, social function and emotional well-being, these statements seem overly restrictive. Properly developed and tested item banks may help ensure that the content validity is adequate for very different kinds of patients within a specific disease and across different diseases.

25. Lines 666-670: *"• The PRO instrument was not developed and validated for use in a clinical trial*

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

• A PRO developed to measure a treatment benefit is subsequently used to measure a decrement as interpreted by a score change in the opposite direction"

Does this mean that PRO tools always need to be developed and evaluated within a clinical trial? Do any new combinations of questionnaires require revalidation? PRO instruments may be developed that have potential for evaluating treatment effects in a clinical trial, and the acceptability of the instrument should be based on the evidence supporting its psychometric qualities, including responsiveness (as aspect of construct validity).

26. Table 3: Define "large segment". Change "missing data points" to "missing data." Response range and Variability sound very similar. It doesn't seem like "differences among patients are not detected when important differences are known" belongs in a row about "variability"--rather it should be in the "ability to detect change" row. That row refers to "item" being non-responsive. Typically we don't evaluate the responsiveness of items per se but rather the scale. As noted above, ability to detect change is a form of validity. The "Item discrimination" row should follow the "inter-item correlation" row.

"validity, ability to detect change" -> validity (including ability to detect change)
"develop or validate" -> develop or evaluate.

27. Lines 717-718: The statement that PRO data from open label studies are "rarely credible" is itself not credible. There are simply too many situations where blinding is simply not possible. It would not be appropriate to recommend against the collection of PRO data in studies that are not blinded.

28. Lines 317-324: Modes of administration: These statements are fair enough if the FDA intends to review the comparability of data obtained using multiple modes of administration. In general, it is recommended that a single mode of PRO administration be used with a clinical trial. If multiple modes of administration are used, then there needs to be evidence that the PRO scores are comparable across modes of administration. But are investigators' toxicity ratings reviewed on the basis of who completes the forms (e.g., clinician versus nurse practitioner)? Laboratories are not required to be comparable to one another, only internally consistent with their reference norms. If laboratories cannot replicate one another, yet the data are used in trial review routinely, why is the use of a SINGLE mode of administration in a clinical trial of much concern? Is this not something which, even if different (and evidence is mixed but favors equivalence in most comparisons), one need not worry about because of randomization and the use of a single mode? Maybe we need to certify PROs through ISOQOL or some other scientific association just like labs get certified. Certified labs get a pass through FDA.

29. Lines 772-775: The frequency of assessment is not only dependent on the natural history of the disease and the nature of the treatment, but on the specific research questions being addressed. For example, if one is interested in the acute (side) effects of treatment, then frequent assessment during treatment may be appropriate. In many cases, however, the acute toxicities are known, and one is more interested in intermediate or long-term effects. In such cases, repeated assessments while on-treatment may not be necessary, but rather assessments over a longer period following completion of treatment would be more appropriate.

30. Lines 791-798: The document should be even more explicit here. When including multiple endpoints (as is often the case with PRO measures), investigators should be required to define 1 or 2 (or a small set of) primary PRO outcomes. These primary outcomes (for the PRO part of the study) should drive sample size estimates and should be the focus of the hypothesis testing. All other PRO endpoints would then be analyzed on a more exploratory basis, and can be used to further support the primary PRO endpoints or to generate hypotheses about additional impact of treatment.

Line 793: Replace 'A PRO instrument' with 'Selected scales from a PRO instrument'
Add after Line 798: Analysis plans should also include power calculations for all proposed hypothesis tests and any additional plans for exploratory analysis of other endpoints.

31. Lines 1059-1060: *An HRQL measure captures, at a minimum, physical, psychological (including emotional and cognitive), and social functioning.*

Suggested revision: *An HRQL measure captures, at a minimum, physical, psychological and social functioning.*

While there are some diseases in which cognitive functioning is expected to be impaired, many, if not most, people do not have noticeable decrements in cognitive functioning. Routine measurement of cognitive functioning in unimpaired populations adds unnecessary respondent burden and cost, and cognitive functioning has not been included

in a number of widely-used generic measures, including the COOP charts, Duke Health Profile (17-item version), EQ-5D, FACT-G, Nottingham Health Profile, and SF-36. The issue of PRO measurement for patients who are cognitively impaired is addressed separately within the document, in Section E.2.

32. VI. D. 2. Lines 1003-1017.

1. The strategies for handling missing data section refers to methods that involve imputation of missing data on a "per-patient basis" but does not mention other approaches such as use of mixture models or joint/shared parameter models. Is this an inadvertent omission or intentional? Given the rapid development and increased application of methods for non-ignorable missing data, allowing inclusion of alternative well-justified approaches as part of the sensitivity analyses would seem appropriate.
2. With respect to imputation on a "per-patient basis", it is suggested that several issues are clarified. The first is whether simple imputation methods would be acceptable, given that these methods generally underestimate the variance and thus overestimate test statistics. The second is whether imputation schemes that only included baseline covariates and the previously observed measures of the outcome of interest are adequate as a sensitivity analysis, given that MLE of the observed data with baseline covariates will result in almost identical results as multiple imputation. Unless ancillary data potentially related to the missing PRO data (e.g. other outcomes, measures of toxicity) are also included in the imputation scheme, imputation will only obscure the problem and not result in a true sensitivity analysis..

Specific Minor Comments

1. Lines 79-80: Focus on patient reported symptoms and function seems limited, given that there are other domains that may be important in understanding the effectiveness of treatment.

2. Lines 135-137: PRO instruments are not typically "validated" by comparing patients' responses to those provided by "expert assessors."

3. Table 1: What is the difference between Overall health status and Health perceptions (in a PRO context)? Technically, timing or frequency of administration is not an attribute of an instrument, but of the study design. Time frame of the questions (recall period) is missing here (it is brought up elsewhere in the document).

4. Table 1: Mode of administration: except for interviewer-administered questionnaires, all of the other modes cited are self-completed (i.e., computer-administered, IVR or web-based).

5. Line164 – Table: Technically, timing or frequency of administration is not an attribute of an instrument, but of the study design. Time frame of the questions (recall period) is missing here (it is brought up elsewhere in the document).

6. Lines 214-233: "If the concept of interest is general (e.g., physical function), a single-item PRO instrument is usually unable to provide a complete understanding of the treatment's effect because a single item cannot capture all the domains of the general concept. Absolutely nothing in existence provides "a complete understanding…" These type statements are completely untenable and not based on the health outcomes research evidence. For this reason, single-item questions about general concepts that imply multiple domains rarely provide sufficient evidence to support claims about that general concept. This seems inconsistent, given the recent history of oncology (and analgesia) approvals based on single items? However, single-item questions about general concepts can be useful to help interpret multi-item measures of the same concept and to determine whether important items or domains of a general concept are missing (e.g., when results using single general questions do not correlate with results using a multi-item questionnaire, this may be evidence that the questionnaire is not capturing all the important domains of the concept contained in the claim)." This is inconsistent, if a single-item measure does not capture all domains of a concept, then a low correlation with a multi-item questionnaire cannot be taken as evidence that the questionnaire is not capturing all domains.

7. Lines 204 and 236: Need some clarification of the differences/similarities between 'concepts' and 'domains'.

8. Line 313: Does the FDA mean 'item response theory analysis' when they mention 'item-response analysis'? Wilson refers to it as item response models. We recommend that item response theory analyses are mentioned in the guidance, since these approaches

to instrument construction and psychometric evaluation have the potential to address many of the measurement problems in PROs.

9. Line 491: *"Test-retest reliability is the most important type of reliability"* There are not different types of reliability, but different ways of estimating or assessing reliability. The advantage of test-retest reliability is that it can be applied to single item measures.

10. Line 520: *When a concept is expected to change, the values for the PRO instrument measuring that concept should change.*

Suggested revision:
When patient experience of a concept is expected to change, the values for the PRO instrument measuring that concept should change.

The concept being measured should not change; rather how the patient experiences the concept might change.

11. Line 1093: "*Item – An individual question, statement or task that is evaluated by the patient to address a particular concept"*

Suggested revision:
"*Item – An individual question, statement or task (and its standardized response options) that is evaluated by the patient to address a particular concept"*

Including response options in the definition of an item is helpful in emphasizing that response choices is an integral part of the item that needs to be carefully evaluated.

12. Pages 21, 28: "Validation" and "validated" do not convey specific information and validity is not a dichotomous attribute.

13. Line 585-588: *On the other hand, if the PRO instrument is to be used in an entirely new population of patients, a small randomized study to ascertain the measurement properties in the new population may minimize the risk that the instrument may not perform adequately in a Phase 3 study.*

Suggested revision:
On the other hand, if the PRO instrument is to be used in an entirely new population of patients, a small study using a representative sample from the new population to ascertain the measurement properties may minimize the risk that the instrument may not perform adequately in a Phase 3 study.

The FDA guidance seems to imply that when a PRO measure is used in a new population, a 'small randomized study' is needed to evaluate measurement properties. While this may represent one of several possible study designs for examining psychometric qualities of a PRO instrument, other study designs (i.e., observational studies) can provide useful and relevant information on measurement qualities.

Most studies to determine the psychometric properties of an instrument are not randomized. We propose that a “representative sample” from the new population be studied, rather than a random sample.

14. Page 13: It wasn't clear what they were arguing for in terms of unequally weighted items. Most scale scores are item sums in classical test theory so item variances weight items differently to some extent. It's not clear to me what they are advocating when they criticize equally weighted scores for each item only if responses are uncorrelated.

15. Page 16: "Validation hypotheses" < what does "validation" mean?

16. Page 19: Interesting that the guidance states that an MID is "usually specific to the population under study" It is necessary to clarify what is meant by "distribution of individual effects". Not clear what is meant by a "distribution-based approach" since all this does is impose a change or difference divided by a SD estimate.

17. Page 20: As noted above, the only sensible way to define responder is based on statistical significance. Statistical significance is necessary but not sufficient. The minimum requirement for the definition of a responder is that the score difference is beyond what can be ascribed to random variation (measurement error).

18. Lines 550-567: It is unclear what is meant here. Why is the aggregation of individual patients’ global ratings of whether a meaningful change has taken place between two assessment points in order to generate mean effects a problem?

19. Line 597: The first bullet (administering a single domain from a multiple domain PRO without the other domains) should not be considered a major measurement violation necessitating revalidation.

20. Lines 645-652: For changed language, why isn’t the situation where a PRO measure is developed in another country and then imported to US addressed.

21. Page 22: Define "harmonization" and what is acceptable level of evidence on translation and cultural adaptation.

22. Lines 926-934: What does the FDA mean by composite measures, something like the SF-36 MCS or PCS, or something like the ACR response criteria? This needs to be clarified in the guidance document. It is possible to create summary scores, based on various methods, or to specify composite measures made of various endpoints. These composite measures should have evidence supporting their measurement qualities (i.e., reliability, validity).

23. Lines 939-942: There will be a problem for studies in requiring that subjects are impaired in all HRQL domains for study entry.

24. Page 31-32: "Scale" -> Response Scale

25. Lines 239-247: The example provided is not clear. If the items assessing dyspnea, in the example, are not valid, then the symptom scale as a whole could not be valid. It is unclear what is meant here. Why is the aggregation of individual patients' global ratings of whether a meaningful change has taken place between two assessment points in order to generate mean effects a problem? (line 550-567)

26. Lines 748-751: It is also important to specify who is allowed to assist a patient with the completion of a questionnaire. For example, it may be difficult for a patient to provide frank answers in the presence of family and close friends.

27. Lines 933 and 934: Substitute 'component' for 'endpoint' in both lines for additional clarity.

28. Lines 1033-1034: Interpretation of study results: This sentence implies that MIDs may be different for examining individual subject change versus examining group mean differences in changes. Therefore, it seems that the criteria for individual change may be different (and higher) than for group change (or differences). This may be reasonable, although there may need to be additional guidance on how this is determined. Regardless, it is recommended that the criteria for interpreting individual change should be stated a priori, with evidence supporting the selection of the MID criteria.

RESPONSIVENESS AND MINIMAL IMPORTANT DIFFERENCES FOR PATIENT REPORTED OUTCOMES

Introduction

Patient reported outcomes (PROs) provide the patient's perspective on the effectiveness of treatment, and for many diseases the patient is really the only source of health outcome endpoint data (Leidy et al., 1999; Revicki et al., 2000; Wilke et al., 2004). For clinical trials evaluating new pharmaceuticals, PROs need to be based on a clear conceptual framework, have evidence supporting content validity (i.e., the instrument content reflects the key characteristics of the construct from the patient's perspective), and must have demonstrated acceptable psychometric qualities (e.g., reliability, validity) (Leidy et al., 1999; Revicki et al., 2000). The PROs must also have evidence documenting responsiveness or sensitivity to changes in clinical status to be most useful as effectiveness endpoints in clinical trials. Without evidence that the PRO can detect meaningful changes in health status, using the PRO in a clinical trial may be risky. Responsiveness is an aspect of construct validity and is determined by evaluating the relationship between changes in clinical and other endpoints and changes in the PRO scores over time, or based on the application of a treatment of known and demonstrated efficacy, in either observational studies or in clinical trials (Guyatt et al., 1987; Revicki et al., 2000; Hays & Revicki, 2005).

Demonstrating responsiveness is necessary, but additional information is needed to determine the minimal important difference (MID) for a PRO measure. Responsiveness represents the instrument's ability to detect changes in health status while MID is used to interpret whether the observed change is important from the patient's or clinician's perspective. Increasingly, in health outcomes research the MID is based primarily on the patient's perspective with the clinician's viewpoint serving to confirm the findings on MID. Responsiveness and MID vary by population and contextual characteristics, and there is no 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 MID has been defined as the smallest change in a PRO measure that is perceived by patients as beneficial or that would result in a change in treatment (Guyatt et al., 1987; Guyatt et al., 2002). There are a number of anchor-based and distribution-based methods that have been used to determine the MID for PRO measures (Guyatt et al., 2002; Crosby et al., 2003; Wyrwich et al., 2005). However, the current situation for determining the MID is fluid and evolving, and there is no clear consensus as to the recommended, best practice approach for determining the MID (Guyatt et al., 2002). The recommended approach is to estimate the MID based on several anchor-based methods, with relevant clinical or patient-based indicators, and to examine various distribution-based estimates (i.e., effect size, standardized response mean, standard error of measurement) as supportive information, and then to triangulate on a single value or small range of values for the MID. Confidence in a specific MID value evolves over time and is confirmed by

additional research evidence, including clinical trial experience. It must be recognized and accepted that aspects of PRO assessment include some measurement error and that no PRO measure is perfect and should not be expected to be perfect in order to be used in clinical trials. There does however need to be evidence that the psychometric characteristics of the PRO instrument are such that there is confidence that changes in scores over time with the application of treatments with some efficacy can be detected (Sprangers et al., 2002) and that the measurement error (or noise) is not so large that it is problematic to observe meaningful changes in patient health status.

Assessing the Responsiveness of PRO Instruments

Longitudinal studies are needed to determine whether a PRO instrument is responsive to changes or differences in health status. These studies may be randomized clinical trials comparing treatments of known efficacy or observational studies where patients are treated with usual medical care and followed over relevant periods of time. To assess responsiveness, some criterion is needed to identify whether patients have changed (either improved or worsened) over time. These criteria, or anchors, may be clinical endpoints (i.e., laboratory measures, physiological measures, clinician ratings), patient rated global improvement or other PROs with established responsiveness, or some combination of clinical and patient based outcomes. The anchor-based approaches use an external indicator, either clinical or patient-based, to assign subjects into several groupings reflecting no change, small positive changes, large positive changes, small negative changes, or large negative changes in clinical or health status. The anchors can be clinical (i.e., hematocrit, ACR response, clinician-rated change, etc.) or patient-based, such as global ratings of change or actual changes in PRO measures that have demonstrated MID in the target patient population. It is highly recommended to use multiple independent anchors and to examine and confirm responsiveness across multiple samples.

Selecting anchors should be based on criteria of relevance for the disease indication, clinical acceptance and validity, and evidence that the anchors have some relationship with the PRO measure. It is recommended that researchers determine the strength of the association of the anchor measure with the MID. An anchor that has a very low or no correlation with the PRO instrument may provide misleading information in determining whether significant change has occurred. There also needs to be an understanding of the trajectory of health outcomes in the target disease to evaluate responsiveness. For example, do most patients improve over time with treatment, as with seasonal allergic rhinitis, or as in many chronic disease (e.g., COPD, arthritis, etc.), is the expected trajectory one of maintenance of health status versus varying levels of deterioration in health status over time, even with treatment?

Once groups of patients are identified as improving, worsening or remaining stable based on several relevant external anchors, several data analyses and indicators can be used to examine responsiveness. First, analysis of variance or covariance procedures can be performed comparing differences in mean baseline to endpoint changes in the PRO scores across the meaningful change groups (i.e., stable versus small improvement, stable

versus moderate improvement, etc.). Second, responsiveness to change is frequently evaluated using different indicators (Hays & Revicki, 2005; Sprangers et al., 2002), such as the effect size (ES) (Kazis et al., 1989), standardized response mean (SRM) (Liang et al., 1990), and the responsiveness statistic (RS) (Guyatt et al., 1987). For these three indices, the numerator is the mean baseline to endpoint change and the denominators are the standard deviation (SD) at baseline (ES), the SD of change for the group (SRM), or the SD of change in patients that remain stable over time (RS). For the ES, Cohen (1988) provided guidance on interpretation of the magnitude, where a 0.20 ES is considered a small change, 0.50 is viewed a moderate change, and 0.80 is viewed as a large change.

Some researchers has suggested that the ½ standard deviation rule (Norman et al., 2003) or that the standard error of measurement (SEM) (Wyrwich et al., 1999; Wyrwich et al., 1999) may represent the MID for PRO instruments. While this magnitude of change is certainly clinically significant and important, since in the case of the ½ SD this represents a moderate effects size (Cohen, 1988), it is not likely to reflect the MID. These differences in PRO scores are just too large to be considered minimally important. While these different distribution-based indicators demonstrate that change has occurred and provide some insight as to whether the change (responsiveness) is small or large, the indices do not necessarily inform as to whether the observed change is MID. To determine MID, it is necessary to get information as to whether the observed change in important from the patient's or clinician's perspective (Osoba, 2005). Based on these methods, MIDs are often in the range of 0.20 to 0.30 ES (or SD units).

Determining the MID for PRO Instruments

For interpreting differences or changes in PRO instruments, information needs to be provided as to whether the changes seen in the scores are important from either the patient's or clinician's perspective. The clinical meaningfulness of the observed change is based on that change being perceived as minimally important and that would be perceived as beneficial from the patient's viewpoint. It is recommended that the patient's perspective be given the most weight, since these are PROs, although the clinician's perspective is considered important as well. The MID is determined based on multiple anchors, that is the same external criteria used to evaluate responsiveness of the PRO measure. However, there are differences in how these data are used and compared to determine MID. Since the focus is on determining the MID, it is necessary to identify the smallest difference or change that is important to the patient.

In many cases, global assessments of change in health or clinical status are used to categorize patients into groups that reflect, based on their own reports, different amounts of change in the construct of interest. For example, based on the Overall Treatment Effect (OTE) scale (Jaeschke et al., 1989), patients can be assigned into groups representing no change (i.e., remaining stable), small improvements, moderate improvements or large improvements, and small amount of worsening, moderate worsening, or large amounts of worsening. As there is often some variation observed even among the stable group, most often the MID is based on the difference in mean baseline to endpoint change scores between the stable group and the small improvement

(or worsening) group. Note that there is evidence that there is asymmetry in worsening and improvement in PROs depending on the specific disease (Cella et al., 2002; Yost et al., 2005). Equally, clinician global assessments of change in clinical status or evaluations of clinical severity, clinical response criteria (i.e., ACR response criteria) or other indicators can be used to determine MID. For these clinical anchors, it will be necessary to identify, based on previous research or clinical consensus, what a small and clinically meaningful effect may be based on these measures. For example in rheumatoid arthritis, the differences between groups of stable patients and those experiencing a 20% ACR response can be used to determine the MID of a PRO score. If multiple anchors are used, there will be several different estimates of MID derived corresponding to these different anchors, and the result will be a range of MID estimates for the targeted PRO instrument.

Finally, the application of multiple methods to determine the MID for a PRO instrument in a specific patient population will result in a range of values for the MID. This is the essence of triangulation, that is, examining multiple values from different approaches and hopefully converging on a small range of values (or one single value). It is recommended that the different MID estimates be graphed to visually depict the range of estimates. To identify a single MID value (or narrow range of MID values), it is recommended that the anchor-based estimates be assigned the most weight and experience from clinical trials be used to further support and perhaps further narrow the range of values. Interpretation of the MID from different anchors should also take into account the proximity of the anchor to the target PRO measure, that is, assign more importance to MIDs generated from more closely linked concepts. A systematic consensus process involving several clinicians and health outcome researchers is recommended and can be completed, based on Delphi methods, to arrive at a single MID value, or at least a narrower range of values. There is no consensus as to how much data are needed as supportive evidence for the MID of a PRO instrument. Clearly, the more data and evidence the better, but a single, generalizable study with multiple patient-based and clinical anchors may be sufficient. As with other aspects of construct validity, responsiveness and the MID value are confirmed based on accumulating evidence from multiple studies and, with additional data, we can be more confident in the MID value. It is uncertain whether a single MID is appropriate for all applications and across all patient populations. For example, the MID derived for an asthma-specific quality of life measure in mild to moderate asthma patients may not be generalizable to clinical trials comparing an add-on treatment for patients with moderate to severe asthma (Niebauer et al., 2006).

Summary and Conclusions

For PRO endpoint data to be accepted as evidence of treatment effectiveness, there must be evidence documenting the instrument's conceptual framework, content validity, and psychometric qualities, including reliability, validity and responsiveness. For responsiveness, it is necessary to demonstrate that the PRO scores are sensitive to actual changes in clinical or health status. While demonstrating responsiveness is a key component to establishing an instrument's construct validity, it is also important to determine the MID to assist in interpreting statistical significant PRO results in clinical

trials. The MID may vary by population and context, and no one MID may be valid for all study applications involving a PRO instrument. Responsiveness and MID must be demonstrated and documented for the particular study population, and these measurement characteristics are needed for PRO labeling and promotional claims.

REFERENCES

Cella D, Hahn EA, Dineen K. Meaningful changes in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207-221.

Cohen J. Statistical Power Analysis for the Behavioral Sciences, Second Edition. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395-407.

Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171-178.

Guyatt G, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clinic Proceed 2002;77:371-383.

Hays R, Revicki DA. Reliability and validity (including responsiveness). In Fayers P, Hays R (eds.). Assessing Quality of Life in Clinical Trials, Second Edition. New York: Oxford University Press, 2005.

Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Contr Clin Trials 1989;10:407-415.

Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27**:** S178-S189.

Liang MJ, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28**:**632-642.

Leidy NK, Revicki DA, Geneste B. Recommendations for evaluating the validity of quality of life claims for labeling and promotion. Value in Health 1999;2:113-127.

Niebauer K, Dewilde S, Fox-Rushby J, Revicki DA. Impact of omalizumab on quality-of-life outcomes in patients with moderate-to-severe allergic asthma. Ann Allergy Asthma Immunol 2006;96;316-326.

Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. Med Care 2003;41:582-592.

Osoba D. The clinical value and meaning of health-related quality-of-life outcomes in oncology. In Lipscomb J, Gotay CC, Snyder C (eds.). Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005.

Revicki DA, Osoba D, Fairclough D, Barofsky I, Berzon R, Leidy NK, Rothman M. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res 2000;9:887-900.

Sprangers MAG, Moinpour CM, Moyniyhan, TJ, Patrick DL, Revicki DA. Assessing meaningful changes in quality of life over time: a user's guide for clinicians. Mayo Clinic Proceed 2002;77:561-571.

Wilke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Contr Clin Trials 2004;25:535-552.

Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T, Sloan JA. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005; 14: 285-295.

Wyrwich KW, Tierney W, Wolinsky F. Further evidence supporting a SEM-based criteria for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999;52:861-873.

Wyrwich KW, Nienaber N, Tierney W, Wolinsky F. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1999;37:469-478.

Yost KJ, Cella D, Chawla A, Holmgren E, Eton T, Ayanian JZ, West DW. Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. J Clin Epidemiol 2005;58:1241-1251.