

April 3, 2006



Management Dockets, N/A  
Dockets Management Branch  
Food and Drug Administration  
HFA-305, Room 1-23  
12420 Parklawn Dr  
Rockville, MD 20857

**GlaxoSmithKline**  
PO Box 13398  
Five Moore Drive  
Research Triangle Park  
North Carolina 27709-3398  
Tel. 919 483 2100  
[www.gsk.com](http://www.gsk.com)

**Re: NAS 0; Not Product Specific  
General Correspondence: Comments on Draft Guidance for Industry -  
Patient-Reported Outcome Measures: Use in Medical Product Development  
to Support Labeling Claims  
[Docket No. 2006D-0044]**

Dear Sir or Madame:

Enclosed please find comments from GlaxoSmithKline (GSK) on the 'Draft Guidance for Industry – Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims' which was announced in the February 3, 2006 Federal Register.

GSK is a research-based pharmaceutical and biotechnology company. Our company is dedicated to the discovery, development, manufacture, and distribution of medicines and vaccines that enable people to live longer, happier, healthier, and more productive lives.

The FDA draft guidance for the use of patient-reported outcomes (PROs) in medical product development to support labeling claims is a very welcome delineation of the Agency's current thinking on this topic and recognizes the importance of PROs in drug development. It will add much to the use and application of PROs and will form a vital reference point for drug development. GSK congratulates the Agency on its efforts to develop this important guidance for industry.

Much of the guidance describes best practice, scientific and methodological standards. Some of the sections within the guidance describe the ideal, the cutting edge of the science and statistics supporting PRO data analysis. We recognize that this guidance represents the Agency's current thinking on PROs and GSK recommends the Agency be flexible in its application of the guidance and consider all aspects of PRO development from conceptual framework to data analysis on an individual, situation specific basis. All aspects of PROs cannot be specified with certainty, as the state of the science of PROs

differs across instruments and therapy areas and is in a constant state of evolution. We fully support the development of a comprehensive set of guidelines that remain pragmatic and flexible in their application, allowing the best possible use of these important patient-reported outcome measures in drug development.

We would like to thank the FDA staff in the Center for Drug Evaluation and Research, the Center for Biologics Evaluation and Research, and the Center for Devices and Radiological Health for compiling this draft guidance document and for the opportunity for comment by stakeholders. Members of the Global Health Outcomes, Legal, and Regulatory Affairs groups at GSK have reviewed the guidance document and contributed comments.

Our overall comments on the draft guidance are provided first, in summary format by section. Specific comments are included in the attachment that follows this letter, and are organized under the same section headings as used in the draft guidance, with a cross-reference to the specific page and line number.

### **Overall Comments**

#### **1. Application and implementation of PRO Guidance (Sections I-II-III-IV)**

There are a number of areas within the draft guidance that are likely to be dependent on the state of the scientific knowledge at the time of PRO application (e.g. determination of Minimum Important Differences [MIDs] and the practicality and feasibility of implementing the PRO guidance [e.g. translations]). While it is appreciated that the FDA guidance may be striving for an 'ideal' approach to the use of PROs to obtain label claims, a degree of pragmatism as well as open dialogue would be expected between developers, sponsors and users of PROs and the FDA. We would like to see this balanced and open approach more clearly stated in the guidance document, consistent with comments from Agency personnel at the recent Mayo Clinic conference on PROs, held February 23-25, 2006.

We applaud and appreciate FDA willingness to discuss with sponsors the development and use of PROs to support product labeling. We recommend that the guidance provide more discussion and an outline of the process for communicating with FDA on PRO-related topics. It would also be helpful to sponsors if the Agency could provide a general timeline for comments from FDA responding to written requests from sponsors for advisories on PRO-related topics, e.g., 60-90 days after receipt at FDA. This will provide useful information regarding expected timelines for product development teams.

The 'substantial evidence' standard applies to label claims as well as promotional claims that are not included in a product label. GSK seeks clarification on whether the draft PRO guideline would also apply to non-label promotional claims.

The guidance proposes that sponsors provide FDA with extensive information regarding the development and performance of a PRO instrument. GSK would endorse a checklist outlining documentation required for submission and other information the sponsor should collect and maintain for potential submission, should FDA request that information.

During the recent Mayo Clinic conference, several comments by FDA personnel seemed to indicate more comfort with patient assessment of symptoms compared with health-related quality of life (HRQL), giving the impression that FDA considers HRQL claims complex with significant hurdles to overcome to be included in labeling. The overall impression was an FDA preference for symptom measurement rather than HRQL. We believe that HRQL claims should be included in product labeling if such claims are based on data from HRQL measures with established reliability and validity, as outlined in the draft guidance.

## **2. Flexibility in standards for well accepted measures vs. newly developed measures (Sections I-II-III-IV)**

Throughout the draft guidance reference is made to the ideal scenario where PROs are used to measure the claimed treatment benefit and specific to the intended population and characteristics of the condition or disease treated. It often assumes that there will be full documentation available delineating an instrument's development, consistent with Figure 1, and measurement properties will always be known prior to use in trials used for label claims. Because of the current 'state of the art' of PRO instrument development, we recommend that a flexible, case-by-case approach be taken. This will ensure that many established but possibly not fully documented instruments already being used in drug development are not excluded without due consideration of their heritage and wide clinical acceptance. This also applies to the adaptation, modification and use of PROs in alternative populations, conditions and treatment types.

## **3. The definition of Patient Reported Outcomes – Proxies and Patient Caregiver HRQL. (Sections I-II-III-IV)**

In Section I, Introduction, the Guidance specifically defines PROs as a measurement of any aspect of a patient's health status that comes directly from the patient. This is of course the expected norm. However, it is also indicated in Section IV.E., Development of PRO Instruments for Specific Populations, that when patients become too ill to

complete a questionnaire or respond to an interviewer they may be assessed using proxy reports. It will be important to make this point clearer both in the Introduction and in the Glossary so as not to exclude this possibility.

We would also like to see clarification regarding the use of instruments used to assess the impact of treatments on the HRQL of patient caregivers (for example caregiver of patients with Alzheimer's disease). Would results obtained from such measures used within the context of clinical trials be considered for inclusion in label or marketing claims, and subject to the guidance for PROs?

#### **4. Overall comments on the development of Conceptual framework and Creation of the PRO Instrument (Sections IV-A and IV-B)**

Some sections of the guidance, and particularly Section IV, imply an unrealistically high threshold. However, recent presentations by FDA on the draft guidance suggest there is more flexibility than the guidance would indicate. An introductory statement that captures the spirit in which these guidelines will be applied would provide perspective. It is possible that the use of terms such as 'generally' or 'usually' in some sentences are intended to convey that there is some latitude in what is considered acceptable. Unfortunately, when used to qualify specific recommendations, and with no additional explanation, these terms only raise additional questions and leave the sponsor without a clear understanding of what is acceptable. If additional considerations will be taken into account, or there are known exceptions to a general rule, it would be helpful to explain this.

We believe it is important for instruments to be evaluated in terms of their psychometric properties rather than by comparison with a list of preferred or recommended characteristics such as the duration of the recall period (see below) or the specific terms used as response options.

Assessing outcomes over an extended time period is often necessary since asking patients to report only their current experiences may not provide a representative sampling. It should also be acknowledged that the strength of memories is likely to vary for different health/life events. Psychometric adequacy should be considered the ultimate indication of whether or not recall is sufficiently accurate since memory failures or biases would be revealed as a lack of validity. Therefore, FDA's position that PRO instruments that require patients to recall over a period of time or to average their response over a period of time may threaten the accuracy of the PRO data is too broad and not based on scientific evidence.

The guidance should clearly distinguish statements describing criteria that the Agency will apply in deciding whether or not data (and the instrument used to generate those data) are adequate to support product labeling from statements that simply provide helpful suggestions.

#### **5. Requirements to revalidate a modified instrument (Section IV-D)**

The draft guidance takes the stance that validation must be established on the final modified instrument prior to Phase III and confirmatory analyses needs to be performed on Phase III data. Modification includes application to a new population or condition. All changes and in some cases even some superficial changes (e.g. such as modification to accommodate adaptation to Case Report Form [CRF] format) appear to require comprehensive revalidation. There is a need to adopt a reasonable and pragmatic approach in context to clinical trials and the modifications being made. Minor modifications and indeed some of the modifications outlined in this section of the draft guidance may not be important enough to warrant a full revalidation study. This is extremely important in certain therapeutic areas, where the development cycle moves from Phase I to Phase III directly. In this instance, revalidation of instruments undergoing minor modification may not be necessary.

The current recommendation in the guidance could potentially lead to an endless revalidation cycle with marginal benefit in measurement properties. This could be detrimental to the value of PRO research, as very stringent criteria will be rarely met and important PRO data could be excluded from product label.

#### **6. Translation of instruments (Section IV-D)**

The current recommendation in the guidance document is to consider generally accepted methods for translations and cultural adaptations. Though there is a general understanding among psychometricians about translational and cultural adaptation methods used, there is no standardized method prescribed. Examples in the guidance on what processes would be considered acceptable and what would be required to ensure that the validity of the responses are not affected would be useful.

#### **7. Study Design (Section V)**

FDA requests a hierarchy of endpoints is provided but this could penalize explorative research with PROs and may not always be possible. We recommend the Agency accept the possibility that sponsors may identify unanticipated, material patient benefits through post hoc and ad hoc analyses of trial evidence. We suggest that that exploration is not equivalent to 'fishing' as trial results cannot always be anticipated. We recommend that uncovered, statistically and clinically valid, evidence of patient benefit should be

reviewed by the Agency for inclusion in product label rather than summarily rejected because it arose from post hoc analyses. We acknowledge that the Agency is concerned with minimization or elimination of Type I error, i.e. identifying false positives. However, there will be occasions where sponsors may be willing to bear some risk of Type I error in order to explore or determine alternative endpoints of patient benefit, thus accepting some risk of Type II error (lack of power and/or false negatives).

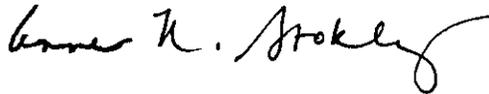
The request for an *a priori* definition of MID may not always be possible but discussion with the FDA at the study design stage would ensure rigor is maintained in all situations.

Clarification regarding the reference as to the concern about bias when unblinding occurs is required to determine whether the FDA is referring to the 'learning effect'.

Again, we thank you for the opportunity to provide comments. The submission is provided in electronic format according to the instructions provided at <http://accessdata.fda.gov/scripts/oc/dockets/commentdocket.cgm?AGENCY=FDA>.

Please contact me at (919) 483-6405 or my colleague Maria Watson at (919) 483-4181, if you require clarification or have any questions about these comments. Thank you.

Sincerely,



Anne N. Stokley, M.S.P.H.  
Director, Policy, Intelligence & Education  
US Regulatory Affairs

Trade secret and/or confidential commercial information contained in this submission is exempt from public disclosure to the full extent provided under law.

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

**Specific Comments from GlaxoSmithKline**

**I. Introduction**

**Line 22:** The draft guidance indicates that PROs are used to measure ‘effectiveness’ in clinical trials. The term ‘effectiveness’ is typically used to describe outcomes observed in clinical practice and naturalistic trials. Endpoints in Phase III clinical trials are usually referred to as ‘efficacy’ outcomes and we suggest this term is substituted for ‘effectiveness’ throughout the guidance.

**Lines 23-24:** The draft guidance indicates that it describes FDA current thinking regarding the use of PROs for product labeling. Clarification is required about the applicability of the guidance for use of PROs in promotional claims that are not included in approved labeling.

**Lines 31 – 32:** FDA’s draft guidance specifically defines PROs as a measurement of any aspect of a patient’s health status that comes directly from the patient. It is, however, indicated in Section IV. E., lines 694 -699, that PROs developed for use in patients that may become too ill to complete a questionnaire or respond to an interviewer can be assessed using proxy reports. The guidance should consistently reflect that when PROs are defined, proxies are a potential alternative method for establishing PRO in special circumstances.

**III. Patient Reported Outcomes – Regulatory Perspective**

**Lines 153 – 156:** The draft guidance indicates that PROs that are used in a clinical trial to support effectiveness claims should measure the adverse consequences of treatment separately from the effectiveness of treatment. It may not always be possible for patients to make this attribution correctly. Moreover, in some cases it could be valuable to quantify the net health impact of a treatment and to include such information in product labeling. The potential value of measuring the net health effects of treatment benefits and adverse events should be included in the guidance.

#### **IV. Evaluating PRO Instruments**

**Lines 178 – 181: 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.**

It is not clear what constitutes a sufficient modification to warrant being considered as new and what appropriate level of measurement properties assessment would be required. Improvements and modifications are usually made throughout the life of an instrument and considerable cost, effort and time may be spent on assessing the impact of marginal changes. We would recommend that the FDA take a pragmatic case-by-case approach that can be agreed on, on the basis of developer recommendation, *a priori* experience, and empirical evidence or other available information or evidence (either published or based on expert consensus).

#### **IV - A. Development of the Conceptual Framework and Identification of the Intended Application**

##### **Lines 199-200**

Existing instruments will not necessarily have been created using all of the steps illustrated in Figure 1. Detailed documentation may be unavailable and a conceptual framework may not have been explicitly described by the developer(s). Additionally, without modification, an existing instrument might also be unlikely to fulfill other specific recommendations delineated in this section. Although it is sometimes possible to work with the instrument developer to modify an existing questionnaire, this is not always feasible. Established questionnaires have a number of advantages, often including a history of use and characterization that provides a rich context for the interpretation of results. It is important that these instruments not be dismissed as suitable candidates for measuring PROs. To this end, the use of established instruments, and how they will be evaluated, should be addressed in the guidance, separately from the process of new instrument development.

##### **Lines 206–208**

The guidance should specify the criteria against which a conceptual framework will be judged as appropriate and under what circumstances, if any, a conceptual framework is unnecessary.

**Lines 212–225**

This section of the draft guidance appears to be contradictory: it seems to state both that a single-item question about a general concept is bound to miss aspects of that concept and that a low correlation between that question and a multi-item measure could indicate that the multi-item measure may not be sufficiently comprehensive. Please clarify the role of a single-item question in supporting the construct validity of a multi-item instrument and, if appropriate, information about the magnitude of the correlation that would provide evidence that the multi-item measure is sufficiently comprehensive.

**Lines 231-237**

If a claim for a general concept is sought based on a summary score calculated from multiple domains, we do not believe it is necessary to demonstrate statistical significance on all of the relevant domains. We recommend that the FDA take a pragmatic, case-by-case approach to determine when a general claim can be made based on the findings for individual domains contributing to the overall score.

**Lines 252–256**

Although the relationships between items and domains and among domains can be hypothesized at the initial stage of instrument development, it is likely that these relationships, which constitute the conceptual framework, will be modified during the validation process. It would be helpful to state this explicitly in the guidance, even though it is implied in Figure 1. It would also be useful to note that modifications in the domain structure will not always necessitate collecting additional validation data. Lastly, we suggest changing ‘expectations’ in line 256 to ‘expected relationships’ or ‘relationships’ to align with language used earlier in the paragraph.

**Lines 269-271**

It is unclear whether the phrase “including excessive severity” is an example of a trial entry criterion or if another meaning is intended.

**Lines 275-278**

When an instrument is developed and validated in patient samples that are appropriately representative of any relevant demographic or disease subgroups, that instrument should be considered appropriate for use in those subgroups.

Please specify how FDA will determine whether the population used for instrument development is sufficiently similar to the patient population.

#### **IV - B. Creation of the PRO Instrument**

##### **Lines 288-300**

Please indicate how FDA will evaluate whether or not the number of patients involved in item generation was adequate.

**Lines 302-303, 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.**

We understand this to mean "*Items that require patients to respond hypothetically may cause patients to respond on the basis of their desired condition rather than on their actual condition and are therefore not recommended.*" We would suggest adoption of this wording to remove ambiguity. Further elaboration on this point and additional examples would be beneficial. For example, it would be helpful to note that elsewhere this distinction has been referred to as one between hypothetical and actual (enacted) performance.

##### **Lines 304-308**

We agree that assessing what patients actually do is preferable to assessing what they think they can do, however it is unlikely that a questionnaire that uses a fixed set of items would match perfectly the activities performed by each of the subjects in a clinical trial. Allowance for this possibility should be made clear in the guidance.

##### **Lines 322-324**

While it is conceivable that multiple modes of administration could lead to systematic differences in results, if the instrument's psychometric properties are acceptable under both modes of administration, and if they are balanced across the groups being compared, then the data should be considered acceptable for evaluating treatment effects. This should be noted in the guidance.

##### **Lines 332-334**

Additional steps to ensure patient understanding should not be necessary if comprehension of the recall period has been evaluated in cognitive testing. The recommendation to evaluate comprehension of the recall period should be added to Section IV – B.5.

##### **Lines 335-337**

Please provide examples of the types of measures that should be taken to ensure that patients make entries according to the study design.

**Lines 339-343**

Assessing outcomes over an extended time period is often necessary since asking patients to report only their current experiences may not provide a representative sampling. It should also be acknowledged that the strength of memories is likely to vary for different health/life events. Psychometric adequacy should be considered the ultimate indication of whether or not recall is sufficiently accurate since memory failures or biases would be revealed as a lack of validity.

**Lines 351 (Table 2)**

The information provided about the various types of response options is inconsistent. In particular, the description of visual analogue scales includes a limitation of these types of scales while limitations of other response options are not included. We recommend that the guidance either include limitations of all response option types or remove this specific limitation.

**Line 363**

The guidance states that it will be important that response options appear to represent equal intervals; however the recommended method for assessing this (distribution of item responses) is not mentioned until Line 414. If other methods, such as cognitive debriefing, could be used instead, it would be helpful to note this here.

**Lines 373-407**

Because the instructions and format of an instrument are developed prior to evaluating patient understanding, the order of Section 5, Evaluation of Patient Understanding, and Section 6, Development of Format, Instructions, and Training, should be reversed.

**Lines 386-398**

We believe the critical question is whether formatting changes could affect the conclusions drawn from the study, not if they can affect patients' responses. Although it is reasonable to encourage consistent formatting, when modifications are required they should be acceptable as long as they are unlikely to affect the conclusions regarding treatment effects.

**Lines 405-407**

Strategies for minimizing missing data and poor data quality are likely to be similar across instruments. Developing a separate User Manual for each questionnaire does not seem to be necessary to meet these objectives and we request that this recommendation be deleted.

**Lines 411-430**

Most established questionnaires use equally weighted items and they have been shown to possess adequate psychometric properties when used in this way. Additionally, research indicates that weighting items makes little difference. It is requested the Agency reconsider this section of the guidance and that it also provide additional detail on the preferred method for assigning item weights, if item weighting is recommended.

**IV - C. Assessment of Measurement Properties**

**Line 481: Table 4, 6th block in 'Test' section, Ability to predict future outcomes (also known as predictive validity).**

The Agency's expectation of data to demonstrate that a PRO can predict future events may be impractical, as it could unnecessarily prolong clinical trial programs. We recommend that "predictive validity" is included in the guidance as an optional test depending on the disease, the purpose and the intended use of the PRO.

**Line 483: Table 4, last 'Measurement Property', Interpretability.**

We request that the Agency specify the Phase of development when it is expected that the MID and responder definition are to be established.

**Line 483: Table 4, FDA Review Consideration for 'Ability to detect change', "Has ability to detect change been demonstrated in a comparative trial setting, comparing mean group scores."**

We would like the Agency to specify if data from Phase II or Phase III data would be required to support the ability to detect change.

**Line 483: Table 4, FDA Review Consideration for 'Interpretability' & Lines 571-577, "The FDA is specifically requesting comment on appropriate review of derivation and application of responder definitions when used in clinical trials."**

Defining responders may be treatment/disease/population specific and it is not clear if appropriate methods to establish responders are well established. Even if the definition is known there is still the need to define what constitutes a meaningful difference between treatment groups in terms of responder rates (e.g. 5%, 10%, 20%, etc.). There is a need to conduct further research to establish best practice and standards for determining 'responders' and meaningful differences in responder rates. Until then a case-by-case, pragmatic approach should be taken to implementing any research strategy aimed at

determining this outcome in clinical trials. This needs to be clearly stated in the guidance.

**Line 483: Table 4, 'Ability to detect change', What is Assessed, "Ability to detect change is always specific to a time interval."**

This statement implies that responsiveness must be evaluated over the same interval as that of the clinical trial. Many Phase III clinical trials are conducted over 52 weeks or more and requiring responsiveness to be demonstrated over this interval prior to including the instrument in a clinical trial seems impractical. In particular, if preceding Phase II studies are over a shorter duration it may not be possible to establish longer term responsiveness. We recommend the wording be changed to *"has ability to detect a change been assessed over a time period likely to indicate it would be suitable for measurement over an appropriate study period."*

In addition, data that supports a claim regarding a treatment benefit is, in itself, evidence that the instrument used to generate those data was responsive. When PRO data support the benefits of a treatment over time and provide evidence that the instrument used was responsive, additional evidence of responsiveness should not be required.

**Line 483, Table 4, 'Interpretability', What is Assessed, "Difference in mean score between treatment groups that provides convincing evidence of a treatment benefit. Can be based on experience with the measure using a distribution-based approach, a clinical or nonclinical anchor, an empirical rule, or a combination of approaches."**

**and Section IV. C. 4, Lines 566-567, "If an 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."**

Line 483, Table 4 suggests that the use of a singular methodology is acceptable, yet in Lines 566-567 the use of a variety of methods is requested. We recommend that there is concordance between the guidance within these sections, by changing the wording in Line 483 to *"Difference in mean score between treatment groups that provides convincing evidence of a treatment benefit can be based on experience with the measure using a variety of methods including an empirical rule, distribution, and/or anchor based approaches."*

**Lines 491-493: “Test-retest reliability is the most important type of reliability for PRO instruments used in clinical trials. Test-retest is most informative when the time interval chosen between the test and retest is appropriate for identifying stability in reference to the clinical trial protocol.”**

The ability to meet the requirement for data to support this measurement property assumes that all diseases are stable. However, there are certain circumstances when remitting and relapsing or episodic diseases are being studied, when it may not be possible to adequately measure this. Assessing reliability suffers two possible drawbacks. Firstly, a person may have changed between the first and second measurement. Secondly, balancing the time between test and retest to avoid recall bias (if the retest is too soon) and the time over which the disease remains stable. We request that the Agency acknowledge that it might not always be possible to measure and or demonstrate stability.

**Lines 510-512: “If developers expected the instrument to discriminate between patient groups (e.g., between patients with different levels of severity), the FDA is interested in evidence that shows the instrument meaningfully discriminates.”**

Using the term “discriminate” in this context is misleading because of reference in the guidance elsewhere to discriminant validity. We request that it be replaced with “differentiate”.

**Lines 542 – 543: “...and a change in an individual that would be considered important, perhaps leading to a definition of a responder.”**

We request that the Agency consider removing the word “perhaps” as it suggests that the Agency is unsure about how to define a “responder” in this instance, yet it is well described later in Lines 569-577.

**Lines 543-547: “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.”**

This statement suggests that there are existing PROs for which it is unnecessary to determine an MID, yet, it is necessary for new PROs. We would like the Agency to clarify this point.

**Footnote # 4, Lines 566-567: “The FDA is specifically asking for comment on the need for, and appropriate standards for, MID definitions applied to PRO instruments used in clinical studies.”**

We believe that there is a need for MID definitions to be applied to PRO instruments used in clinical studies, and importantly, for research to establish best practice and standards for determining MIDs. However, until best practice has been established, we request that the Agency specifies that it is willing to apply a case-by-case, pragmatic approach when reviewing the methodologies used to demonstrate this particular measurement property.

#### **IV - D. Modification of an Existing Instrument**

**Lines 585-589:** The draft guidance indicates that if a modified 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 will not perform adequately in a Phase III study. It is not clear why a randomized study is needed – unless to assess treatment effects, MIDs and responsiveness. If psychometrics are generally unaffected by modification then these parameters may not be affected. A more flexible approach, whereby a separate study should not be required unless characteristics of intended population and the population originally used in PRO development are totally different, should be included in the guidance. Consequently, revalidation as a part of Phase III should be allowed and evaluated on a case-by-case basis.

**Lines 599-670:** The draft guidance takes the stance that validation must be established on the final modified instrument prior to Phase III and confirmatory analyses needs to be performed on Phase III data. Modification includes application to a new population or condition. All changes and in some cases even some superficial changes appear to require comprehensive revalidation. Though this would be ideal, there is a need to adopt a reasonable and pragmatic approach in context to clinical trials. For instance, a well developed generic instrument measuring a common concept (or concepts) is valid when applied across a wide continuum of patient populations and health status. Hence, if it has been adequately validated, and its measurement properties have been established, it is likely to be inappropriate to require revalidation for each and every new patient population and or condition. This is also expected to be the case when minor modifications are applied to enable use in trial settings (e.g. such as modification to accommodate adaptation to CRF format). In these instances confirmatory validation in a Phase III trial should be adequate. Similar logic could work for the use of a single domain from a multiple PRO administered without other domains. Additionally, it is possible that if there is a change in measurement properties it would affect all the groups specified and should not affect the treatment benefit measured.

Also, consider the case of a disease-specific instrument. If results from a Phase II study identify a minute problem which, when corrected would provide an adequate instrument, then confirmatory validation in Phase III should be sufficient to support a claim. It is not possible within the development cycle to run a small randomized study between Phases II and III in most instances for revalidation of such modified versions.

This is extremely important in certain therapeutic areas, where the development cycle moves from Phase I to Phase III directly. In this instance, revalidation of instruments undergoing minor modification would become impractical.

We would recommend that FDA take/accept a pragmatic, case by case approach to the scope of changes, requirement of additional validation and the extent of analysis required when instrument modifications are made.

**Lines 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.”**

This will require post adaptation validation prior to implementation. There are no empirically established gold standard translation methodologies. The emphasis is on the sponsor to provide evidence of valid translation processes. FDA should provide explicit guidance on what translation methodologies are acceptable and which ones provide the best chance for establishing valid versions.

**Line 666:** Most instruments were not originally developed for clinical trials, so a lot of retrofitting is implied by this section. It would be helpful if the FDA would identify what are the critical factors affecting the applicability of any instrument to a clinical trial setting.

**Line 659:** Harmonization is usually only done when there are several translations produced – when new individual language versions are translated this line might indicate that harmonization with other previously translated versions is required. A problem will occur when new language versions are required during a clinical trial program (that is often the case). If harmonization suggests changes to the previously deployed versions, this could lead to either significant data loss and/or rework. Clarification is required from the FDA as to what they would expect in terms of harmonization as new translations of existing, validated measures emerge.

#### **IV – E. Development of PRO Instruments for specific populations**

##### **Lines 692 – 699, “2. Patients Cognitively Impaired or Unable to Communicate”**

In critical care and patients with cognitively impaired functioning there is precedence to the use of proxy (usually a caregiver) to assess observer based functioning of patients. The guidance states that in such situations, 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. In critical care the first assessment occurs at enrollment, so the entire data collection program must be based on proxy assessment. Also, when a part of the data is collected from proxy and a part from patient, there is bias introduced, which can be reduced by use of only one form of respondent. The guidance needs to address trials which are assessing drug benefits in such patients.

#### **V – Study Design**

**Line 731-33:** This section appears to focus on bias caused by unblinding and or access to previous responses. The potential for learning effects may also be considered problematic when PROs are administered and this can similarly be overcome by blinding and randomization. It is recommended that reference to this possibility be made in this section.

**Lines 735-738:** This is a subjective statement that could vary by person and institution. It would be helpful to provide examples to clarify some ‘**certain situations**’ for other situations where there is no reasonable control group.

**Lines 748 – 751:** It would be useful to have a statement on the standardization of the order in which PRO / other clinical investigations should be administered. PROs should be administered before other clinical investigations to avoid ‘test bias’.

**Lines 774-776:** The frequency of PRO assessment should not only correspond to the measurement properties of the instrument but also the likely course of the disease/condition and expected treatment impact. The latter should be mentioned in this section.

**Lines 791-798:** It is not always possible and some times improbable to establish hierarchy *a priori*. While study sponsors will specify which parameters of any survey will be affected by treatment, they may not be able to identify the hierarchy of endpoints with certainty. Inherently, this requirement penalizes exploration of the patient (PRO) perspective. We suggest that the Agency review the level of evidence for all endpoints

and consider accepting greater risk of Type I error, i.e. false positives, in order to identify additional or non pre-specified patient benefits. We acknowledge that the Agency is concerned with minimization or elimination of Type I error, i.e. identifying false positives, and this is a position we support. However, there will be occasions where sponsors may be willing to bear some risk of Type I error in order to explore or determine alternative endpoints of patient benefit, thus accepting some risk of Type II error (lack of power and/or false negatives). We believe this may provide a more complete efficacy picture, supporting the voice of the patient (PRO).

**Lines 802-811:** In Lines 537-577 the FDA have already indicated that establishing MIDs / responder definitions may be problematic. These parameters need to be discussed (as suggested) and **agreed to** prior to study design and implementation. For many instruments, in particular new ones, this may be problematic where practical experience is minimal. In these situations a pragmatic approach to interpretation may be required. It should be made clear that in these situations agreement between the FDA and the sponsor is reached on definition of MID / responder prior to study design. When new instruments are being used a pragmatic approach to determining MID/responders is recommended.

**Lines 815-857:** This section regarding specific concerns when using electronic PRO instruments focuses heavily on records maintenance. While we agree that this is a vital issue, the length of this section may confer relatively more weight, in terms of importance, versus other equally important sections of the guidance. Much of this section is redundant as it duplicates the guidance given in the “Draft Guidance for Industry: Computerized Systems Used in Clinical Trials” and the “Guidance for Industry, Part 11, Electronic Records; Electronic Signature – Scope and Application”. We recommend reducing the length of this section, keeping the introductory text and the footnotes/references to other FDA guidances.

**Lines 844-845:** The guidance should define the terms ‘cybersecurity’ and ‘network attack’.

**Lines 847-857:** The bulleted list is stated as a negative in terms of ‘plan to avoid’. We suggest that the negative statements are changed to a list of positive actions that are encouraged by FDA, or delete the bulleted items, as they are already covered by the “Draft Guidance for Industry: Computerized Systems Used in Clinical Trials” and the “Guidance for Industry, Part 11, Electronic Records; Electronic Signature – Scope and Application”.

## **VI – Data Analysis**

**Lines 871-872, Section VI, Data Analysis, A. General Statistical Considerations, “Unplanned or post hoc statistical analyses are usually viewed as exploratory and, therefore, unable to serve as the basis of a claim of effectiveness.”**

There are situations where post hoc statistical analyses are appropriate in clarifying or further elucidating results, underlying relationships, and causation. We suggest, if there are valid theoretical or evidentiary reasons to perform such analyses, as long as the statistical analyses are appropriately designed, that the Agency not limit the use of these analyses in label claims.

While it is preferable to pre-specify outcomes of interest, the general ‘investigative’ nature of PRO research needs to be taken into account when establishing label claims. Clear patient benefits may not be fully communicated to prescribers unless a more comprehensive approach is taken to including reference to post hoc findings within the label.

### **Lines 876-922, Section VI, Data Analysis, B, Statistical Considerations for Using Multiple Endpoints**

Further discussion is warranted regarding statistical considerations for using multiple endpoints. Sponsors do not usually power clinical studies for secondary endpoints, and in many cases knowledge about clinically meaningful differences and effect sizes can be incomplete at the time of the Phase III trials. We support the guidance regarding HRQL or PRO primary endpoints; however the language is too restrictive and prescriptive to ensure the most appropriate handling of secondary endpoints, instruments, and statistical analyses.

We recommend that the Agency provide qualifying language as a preamble to this section, specifying review and comment consistent with current statistical and endpoint knowledge in a manner which balances Agency and sponsor risk of false positive results. We agree with *a priori* specification of hypotheses and some hierarchy of endpoints and comparisons, however, sponsors may not be able to fully specify the hierarchy. We request that the most important or relevant points in these recommendations be identified in the guidance.

**Lines 884-885:** Further discussion is warranted regarding the association or linkage of clinician-reported outcomes (CROs) and endpoints with patient-reported endpoints, as PROs and CROs are conceptually different, executed differently, and require differing degrees of judgment and subjectivity.

The language in the draft Guidance fails to acknowledge the imperfect correlation among measures: CROs, PROs, symptom severity, adverse events (AEs)/serious adverse events (SAEs). It would be inappropriate to hypothesize relationships or causality among the measures based on 'language' similarities, i.e. multiple measures related to single symptom but assessed by different tools and different personnel or self reported.

We would like the Agency to clarify their reasoning for drawing linkages, the objectives of linking such endpoints and evidence, and to identify what is gained or the utility of such linkage. We recommend not forcing statistical relationship on independent concepts because 'one can'; conceptual theory should precede statistics.

**Lines 892-894, "Because each PRO item or domain often can represent an endpoint that could imply a distinct claim on its own, we recommend careful planning to avoid substantial increases in Type 1 error from multiple endpoints."**

This statement seems potentially contradictory to Lines 946-949, which recognizes that individual component results need to be evaluated in the context of total composite and other individual component results and not solely on their own merit event after controlling for multiplicity. We request clarification of this apparent contradiction.

**Lines 924 and 926-927 – Statistical Considerations for Composite measures**

We would like to see confirmation of definitions used for 'composite endpoints' including index, profile or battery, in the Glossary. Although Table 1 contains this information, when sections are read in isolation the definitions may not be clear.

**Lines 930-931: The Agency needs to clarify what constitutes "...substantial prior empirical evidence of the value of the chosen components of the composite."**

There are two issues raised by this language and the text preceding it. With this text, the FDA ignores the common situation that patients with disease do not present with similar/equal symptoms or physical/functional impairments (equal impairment of the same HRQL domains or with the same symptoms of equal severity). For example, all patients with breast cancer-derived brain metastases will not present with the same symptoms, the same level of severity among those symptoms which are common or resultant HRQL impairments. Therefore there will be some variability among subjects enrolling in trials inherent in the etiology and natural history of disease. The second issue is recognition of variance in composite endpoint values derived from variation in presenting symptoms or the individual components of any composite endpoint. Therefore, some of the recommendations underpinning composite endpoint creation will not achieve the scientific and evidentiary objectives of the Agency.

We request clarification of the evidence required to support composite endpoints. We would like to know if the evidence depends upon the reason why a sponsor is proposing the use of a composite endpoint, the composition (individual metrics) of the composite endpoint, or both. We would like the Agency to clarify what constitutes “substantial prior empirical evidence of the value of the chosen components of the composite”. The guidance needs to specify a forum for discussion of composite endpoints and review of exceptions to the evidentiary support required.

**Lines 940-942: “...it is critical to ensure that patients enrolled in a clinical study are impaired in all domains...”**

The recommendations in the draft guidance add a level of specificity in selecting patients for clinical trial participation which may not be appropriate, i.e. limiting enrollment only to the most severe and those with an equivalent set of disabilities/ symptoms.

We are concerned that this perspective does not represent patients who generally enroll in clinical trials. Such a requirement may limit who enrolls in trials, slow the speed of study enrollment, and unnecessarily restrict qualifying patients (via narrowing the definition of patients who might benefit).

While we acknowledge the need to identify patients who might ‘most’ benefit from therapy, we request that the Agency not define those who most benefit as only the most severe, disabled or are otherwise consistently impaired across specific domains.

**Lines 944-954, Multiplicity of endpoints and where a sponsor ‘spends alpha’**

The language of the draft guidance regarding adjustment for multiplicity of endpoints and alpha spend does not ensure that the most appropriate analysis for the endpoints and research hypotheses is applied.

Although adjusting for a multiplicity of endpoints is appropriate where one is certain of detecting statistically significant differences across endpoints, statistical significance is not the same as clinical meaningfulness. Multiplicity adjustments represent a most conservative approach to inference and furthermore there is more than one way to adjust for multiplicity of endpoints. In many diseases and in clinical trials, patients may not be as sensitive to the impacted HRQL or PRO measure (amelioration of symptoms or reductions in disabilities) as a sponsor hypothesizes. Analyses, models, and statistical approaches form a continuum and the best approach should be identified, rather than the approach specified in the draft guidance, as there are risks/benefit tradeoffs of all approaches.

We support doing a statistical analysis commensurate with the endpoints, hypotheses, and objectives of the study or trial. We recommend that the agency review analysis plans and statistical adjustments according to what is most appropriate to the metric, the instrument, or the trial design rather than specifying adjustment for multiplicity. We recommend that the Statistical Analysis Plan (SAP) recommend an approach and discusses how such an approach addresses the objectives of the Agency (the concern of false positive results), discuss differing objectives, risks and benefits.

**Lines 956-1017: D. Statistical Considerations for Patient-Level Missing Data - Handling of Missing Domains, Items, and Entire Measurements**

Recommendations for handling missing values and missing instruments warrants further discussion as any one approach will not mitigate the risks of every situation of drawing false conclusions or quantifying benefits. In addition, the science is evolving for handling these issues.

Sometimes pre-specifying rules for handling missing data within a domain is inappropriate as appropriate treatment of data requires analysis of the ways in which the data are missing.

A prescriptive approach to handling missing values and failure to recognize differences between statistics and clinical meaningfulness is also inappropriate.

Rules will vary by instrument used – depending on developer’s rules and previous accumulated experience, as well as experience in the context of the specific trial.

We request that the Agency remain open to alternate approaches to missing data and allow the sponsor to identify the most unbiased, scientifically and statistically appropriate, and practical ways forward in handling it—consistent with data, instruments, extent of missing data, and new theoretical (academic literature) recommendations for handling missing data.

**1009-1017: Strategies for handling missing data**

**“When there are few patients with missing measurements and the frequency of missing data or proportion of patients with missing data is comparable across treatment groups, most approaches yield similar results. When a higher proportion of patients have missing data, the FDA recommends the use of several different imputation methods (including a worst-case scenario in which missing data are assumed to be unfavorable for those on the investigational treatment and favorable for those in the control group) and an assessment of the consistency of the study results using each method.”**

The Agency is cautious about the methodological approaches that might be used to address missing data – no single method is preferred. While methods may be pre-specified, it is only once the nature of missing data is known that suitable methods can be applied.

It is recommended that while one or two specific methods for handling missing data can be made in the SAP, the Agency should allow alternative methods (with sensitivity analysis) to be applied once patterns of missing data are known.

**Lines 1019-1035, E. Interpretation of Study Results**

Further discussion is warranted regarding evidence and timing of establishing and documenting MID.

We accept and support the need for MID information, and believe that what is important to patients and clinicians is an MID which is statistically significant as well as clinically meaningful. However, we think it will be impossible and impractical to retrofit all instruments developed and used in trials, or to de novo establish all relevant parameters of MID to an equal level of precision.

Although we accept the definition of MID employed in the draft guidance, we would like the Guidance to discuss degree of retrofitting required across all instruments, measures, surveys, indices, or a discussion of necessary and sufficient evidence given the development histories of various instruments.

**Lines 1031-1035: “When only a modest fraction of people respond to a treatment, that fraction may experience meaningful change in the face of a mean effect that is very small. When defining a meaningful change on an individual patient basis (i.e., a responder), that definition is generally larger than the minimum important difference for application to group mean comparisons.”**

Interpretation of these scenarios is unclear. We request clarification of the ultimate objective of the Agency and the primacy of either the MID or some other measure of meaningfulness or clinical benefit.

**Glossary:** consider adding a definition of ‘recall period’ and ‘composite endpoints’.