



UCB, Inc. – 1950 Lake Park Drive – Smyrna, Georgia 30080

3 April 2006

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

Subject: Docket No. 2006D-0044
Draft Guidance for Industry on Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims

Dear Sir/Madam:

Thank you for the opportunity to comment on the “Draft Guidance for Industry on Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims” published in the *Federal Register* on February 3, 2006, page 5862. UCB welcomes the draft guidance and believes that it is a positive first step toward a collaborative context to utilize patient-reported outcomes in studies to support product claims in approved labeling. However, we have some concerns about this proposed draft guidance document. Below are UCB’s comments for your consideration.

I. Introduction

Line 31: Clarify whether or not caregiver-reported assessments are included within this guidance document. If caregiver-reported assessments are not included in this guidance, which FDA guidance document addresses caregiver-reported measures?

Lines 38-40: “*For this data to be meaningful, however, there should be evidence that the PRO instrument effectively measures the particular concept that is studied.*” Clarify the extent of evidence required for: 1) well established measures used repeatedly in clinical trials for the same treatment indication and when a PRO claim is already achieved by competitor (e.g., in Crohn’s disease, IBDQ results are used in successful labeling claim for Remicade®); 2) PROs used in numerous clinical trials and studies providing validation results (published results) but no label claim approved by FDA; and, 3) newly developed PROs.

2006D-0044

C4

II. Background

Lines 82-83: “*PRO instruments that measure a simple concept may not be adequate to substantiate a more complex claim*”. If a disease-specific instrument does not cover all domains of HRQOL, can it be combined with domains of a generic instrument to substantiate a general claim on HRQOL assuming that all appropriate domains of HRQOL are covered by the these two measures?

III. Patient-Reported Outcomes – Regulatory Perspective

B. A Taxonomy of PRO Instruments

Line 151: As PRO concepts can also be population specific (children, menopausal women, etc.) you may want to consider adding this to the proposed list.

Line 164: Taxonomy table. This table is not exhaustive. For example, it does not include productivity, health-related quality of life, and daily symptom charts. Please consider including a footnote that the table is not exhaustive.

IV. Evaluating PRO Instruments

Line 179: Edit at the end of the sentence: “would **be** a new one”

A. Development of the Conceptual Framework and Identification of the Intended Application

2) Identification of the Intended Application of the PRO Instrument

Lines 264-267: “*the development and demonstrated measurement properties of a PRO instrument [...] is best established before the study commences, but would in any case be part of the FDA’s application review*”. What documents is the FDA expecting in the briefing package before the start of phase III studies?

3) Identification of the Intended Population

Lines 275-278: “*The FDA plans to compare the patient population used in the PRO instrument development process to the study populations enrolled in clinical trials to determine whether the instrument is appropriate to that population with respect to patient age, sex, ethnic identity, and cognitive ability.*” Please clarify the wording to include “ethnicity” and “cognitive ability” if available. Is the FDA suggesting that this comparison be consistently provided for all PRO measures along with other instrument details as an Appendix to a submission?

Line 277: It is mentioned here that the FDA plans to compare populations of the clinical trial and of the questionnaire development. However, some of the characteristics listed here are not always documented in the development documentation (scoring manual or publications). Also, some of the suggested patient's characteristics may, for some conditions, not be important or relevant. Therefore we understand that the Sponsor should ensure that the most important characteristics are similar (for example if we are seeking to work in children, that the age categories are the same as the questionnaire targeted age groups) and referring to the user manual or the original development publication is sufficient.

B. Creation of the PRO Instrument

1) Generation of Items

Lines 298-300: *"The FDA plans to review instrument development (e.g., results from patient interviews or focus groups) to determine whether adequate numbers of patients have supported the opinion that the specific items in the instrument are adequate and appropriate to measure the concept"*. Would this also apply to well-known instruments that were developed a long time ago and for which the validity has been demonstrated in numerous circumstances (e.g., SF-36)?

Line 299: What would be an adequate number of patients for FDA and sponsors to be confident in setting up the conceptual framework and generating items?

2) Choice of the Data Collection Method

Lines 315-324: *Choice of data collection method*. The details of this section address issues of data collection and all procedures and protocols of clinical trials. However, this section is contained within the overall section of "Creation of a PRO Instrument". Should this information be included in the "Study Design" section?

3) Choice of the Recall Period

Lines 332-334: *"When evaluating PRO-based claims, the FDA intends to review the study protocol to determine what steps were taken to ensure that patients understand the appropriate recall period."* As the comment in this section is "when evaluating PRO-based claims", the information here appears to be referring to a clinical trial and not the development of a PRO instrument. Yet, this section "Choice of the Recall Period" is contained within the overall section of "Creation of a PRO Instrument". Please clarify, if the FDA recommends that as part of a clinical trial, the patient's understanding of the recall period be verified. Or if indeed, this responsibility of verifying the patient understanding of recall period is only referring the process of developing a PRO instrument.

Lines 334-337: *"If a patient diary or some other form of unsupervised data entry is used, the FDA plans to review the protocol to determine what measures are taken to ensure that patients make entries according to the study design and not, for example, just before a clinic visit when their reports will be collected."* Does the FDA have a recommendation of sufficient measures to ensure patients make entries according to study design?

6) Development of Format, Instructions, and Training

Line 394: *"Changing the instructions or the placement of instructions within the PRO instrument"*. This phrase is from an earlier section of the Guidance Document, but it refers to modifications of existing instruments. Some questionnaires (although widely used already in the field) may have no instructions or confusing inadequate instructions. Therefore, it is sometimes appropriate to make modifications to the instructions. It seems inappropriate to perform full testing on a measure for such a slight change. Would the FDA consider cognitive debriefing in a small sample of patients (N=5) for their understanding the instructions sufficient to validate this change in a questionnaire?

7) Identification of Preliminary Scoring of Items and Domains

Lines 413-414: This sentence may need to be changed as the expression "appropriate intervals" refers to the distance between answer choices which can not be evaluated by looking at the distribution of responses.

Line 417: *"Equally weighted scores...relatively uncorrelated"*. This statement could be misleading. We understand the author's concern is to avoid overly correlated items reflecting redundancy, but we do expect items to be correlated significantly if they pertain to the same dimension. Moreover, in some cases, having two different items measuring concepts that are related (and therefore highly correlated) and that are both important to the patients (as identified through patients interviews) may be appropriate.

Lines 424-430: Does this mean that the FDA does not accept widely used utility scores such as the EQ-5D or HUI or that the FDA would not consider a global HRQOL claim based on such instruments?

Lines 426-429: *"Because preference weights are often developed for use in resource allocation (e.g., as in cost-effectiveness analysis that may use predetermined community weights), it is tempting to use those same weights in the clinical trial setting to demonstrate treatment benefit"*. Please consider adding *"in HRQOL / health status"* at the end of this sentence.

C. Assessment of Measurement Properties

4) Choice of Methods for Interpretation

a) Defining a minimum important difference

Line 543-545: "*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*". As pain is reported directly by patients, shouldn't this be considered a PRO?

Line 545: "*If PRO instruments are to be considered more sensitive than past measures....*" Why should PRO instruments be considered more sensitive than past measures? PRO measures are providing a different perspective and adding a new vector in the condition space allowing a more complete picture of the disease and the treatment impact on patients' health status and patients' lives. In our opinion, PRO measures may in some cases be more sensitive but in most cases, they are not expected to be more sensitive.

Lines 546-547: "*...it can be useful to specify a minimum important difference (MID) as a benchmark for interpreting mean differences*". Please note that definitions of the "Minimal Important Difference" found in the literature refer to a 'within patient change' (Jaeschke, Guyatt...) and not to a 'between group difference'.¹ The confusion often seen in published work between this initial MID definition provided by Jaeschke et al and other papers is probably due to the ambiguous terminology used. What Jaeschke et al refer to is a 'within patient change' and therefore, could be called a minimal important (patient) change leaving the expression MID for between group differences. The methodology proposed by these authors suggests using the MID threshold to define responders rather than using the threshold as a minimal 'between group difference'.

As mentioned later in the text, for many clinical endpoints, the ability to show *any* difference between treatment groups has been considered evidence of a relevant treatment effect. If we consider an endpoint as secondary, the sample size will not be calculated based on that endpoint. However, after obtaining positive results on the primary endpoint, consideration will be given to the secondary endpoints and if the given sample size allows to demonstrate a statistically significant difference (after accounting for multiplicity), why should there be a special consideration in the case of PRO data to define a Minimal Between Group Difference? If the PRO endpoint is to be considered primary, then as for any primary endpoint, the minimum difference

¹ Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989 Dec;10(4):407-15

between the treatment groups that is judged to be clinically important will need to be discussed based on previous work.

b) Definition of responders

Line 577: The major risk in describing the percentage change is that it generates highly skewed data and provides higher weights on changes occurring at the bottom of the scale (changing from 0 to 10 translates into a percentage change of infinity, whereas changing from 50 to 60 translates into only a 20% increase. These highly skewed data require transformation of the data or non parametric testing.

D. Modification of an Existing Instrument

5) Changed Culture or Language of Application

Line 660: "*The evidence that measurement properties for translated versions are comparable*" From this statement, we understand that if the first two points are addressed adequately, (i.e., ensure qualified and highly experienced individuals are involved in the translation/adaptation and a rigorous translation/adaptation methodology is used) then psychometric testing need NOT be performed to evaluate the measurement properties of each language version to the original. Is this correct? If no, please specify in which cases, full psychometric testing must be performed for each language version. And to what extent?

6) Other Changes

Lines 667-668: What is the rationale for requiring additional validation work in the case where a questionnaire is being included as part of a battery (except for burden consideration, that can be assessed by describing the return rates and quality of completion), and what kind of data would be needed?

V. Study Design

A. General Protocol Considerations

1) Blinding and Randomization

Lines 717-718: "*Because responses to PRO measures are subjective, representing a patient's impression, open-label studies, where patients and investigators are aware of assigned therapy, are rarely credible*". What is the FDA's opinion on open-label extensions as follow-up of double-blind studies: when a treatment benefit is demonstrated in the double-blind study, can the open-label extension be used to substantiate that the improved state is maintained on the long term?

D. Design Considerations for Multiple Endpoints

Lines 793-795: We understand that this statement means that secondary end points are to be considered only after the primary or co-primary endpoint(s) results have proven positive. This may need to be more explicit to avoid confusion.

VI. Data Analysis

D. Statistical Considerations for Patient-Level Missing Data

1) Missing Items within Domains

Lines 969-971: *“For example, the SAP can specify that a domain will be treated as missing if more than 25 percent of the items are missing; if less than 25 percent of the items are missing, the domain score can be taken to be the average of the nonmissing items”*. Most questionnaire developers recommend a 50% missing item rule for imputation, would this be acceptable?

Glossary

Lines 1059-1061: *“An HRQL measure captures, at a minimum, physical, psychological (including emotional and cognitive), and social functioning.”* Please consider removing “cognitive” or making a note that the assessment of cognitive function is dependent on whether or not there is evidence that disease or treatment may impact cognitive functioning.

Line 1075: The definition of MID is not clear and it seems that different sections of the document refer to different definitions of the MID. Are we talking about a minimal between group difference (on a continuous variable at a given time point; or on a change over time...) or a within patient change? To avoid confusion, it may be better to have two different expressions to cover those two different things:

1. The “minimum difference between groups that is judged clinically important” (Minimum Difference) which is the concept common to all clinical trial endpoints and that is used to estimate sample sizes.
2. The “Minimal important change overtime” (MIC) which corresponds to the definition provided by Jaeschke and Guyatt for the MID expression.

Please include definitions for “Index”, “battery”, and “profile”.

UCB appreciates the opportunity to comment on this draft guidance. Please contact me (770-970-8584) should you have any questions regarding this letter.

Cordially,

A handwritten signature in cursive script, appearing to read "Robert A. Paarlberg".

Robert A Paarlberg
Director
Global Regulatory Policy & Intelligence