

**EORTC**  
*European Organisation for Research and Treatment of Cancer*  
AISBL-IVZW  
Avenue E. Mounierlaan, 83/11  
Brussel 1200 Bruxelles  
België - Belgique

Division of Dockets Management (HFA-305),  
Food and Drug Administration,  
5630 Fishers Lane, rm. 1061,  
Rockville, MD 20852.  
U.S.A

Wednesday, 29th March, 2006

Dear Laurie,

***Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims: Initial comments and observations from the EORTC***

We are pleased that the FDA has given the scientific community at large the opportunity to comment on the recent PRO guidance. On behalf of the EORTC Data Center and Quality of Life Group we would like to provide our comments and observations on the recent PRO guidance.

Overall, we are impressed and very pleased that the FDA has produced this important document. This text now helps set out the importance of PRO data in registration studies. It is very well written and will be a considerable benefit to researchers and those planning future clinical trials with a PRO component.

With that said, after a detailed review and subsequent discussion within the EORTC, we felt that we should make a number of both observations and also recommendations that may lead to greater clarity in this document. We have noted out comments below as major or minor.

## Major comments

### General comment

The FDA is a little inconsistent with the terms *efficacy*, *effectiveness*, and *efficiency* (most of the time the term *effectiveness* is used). However, these terms have distinct meanings. Generally, it seems now accepted that efficacy denotes that a treatment works under optimal, controlled conditions (e.g., RCT) - effectiveness means that a treatment works under routine conditions (cohort/observational studies, clinical routine) and - efficiency relates the benefits of a treatment to the costs involved. The FDA should make ideally be clear how these terms are used in the context of the document.

### Specific comments

On page 6: *When considering an instrument that has been modified from the original, the FDA generally plans to evaluate the modified instrument just as it would be a new one.* This basically implies that if researchers want to use a modified version of a tool or module that has been fully validated (e.g. to leave out a sexual functioning subscales from a multidimensional questionnaire because of lack of relevance in a particular population or to add a few study-specific items via an item-bank), they would need to validate this modified version via a full field study. This could be problematic and in some cases unnecessary.

On page 11: *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.* We would take exception to the last part of this statement. All EORTC and many other tools, such as the FACIT questionnaires rely on recalling experiences over a certain time period. The advantage in doing so is that the likelihood that patient-reported data will be influenced by an unduly “good” or “bad” day is minimized. In symptom management trials, it may be important to assess symptoms frequently (e.g., on a daily basis) rather than asking patients to average their symptom experience over a number of days. However, this may be much less relevant and appropriate in clinical trials that run over a longer period of time (in some cases many years) and where there the focus is not on acute symptom experience. Also on page 11, lines 339-340 the documents states that *PRO instruments that rely on memory lack accuracy.* In some sense all questionnaires rely on memory, most instruments refer to the "past seven days" as noted above. In this context, the issue of "accuracy" is an almost philosophical one. What is of interest is how patients evaluate the reference period in the respective domain -- not how accurate they recall all the single events that lead to their judgement.

On page 20, Paragraph D, it is suggested that modifications of an instrument, including translation into other languages, will require a full validation of the modified instrument. While in the ideal world this would be desirable, in practice it would be extremely difficult if not impossible. For example, the EORTC QLQ-C30 is translated into some 68 languages and many of the condition-specific questionnaire modules into nearly as many. New language versions are continually being developed, as clinical trial activities are expanding into other parts of the world. In the long run, we expect that data will become available by which we can examine any deviations in psychometric properties between the parent instrument and the many language versions. However, it is simply not feasible to conduct independent psychometric field studies to examine this question for each language. This holds for other measures as well (e.g., the FACT-G, the SF-36) that are used extensively in clinical trials. We hope that the FDA will reconsider its position on this point.

On page 23 the FDA note: *Every effort should be made to assure that patients are masked to treatment assignment throughout the trial.* In some fields of medicine, including oncology, double-blind trials may often not be feasible and may in some cases be ethically unacceptable. The suggestion that PRO data are less valuable in non-blinded studies is not backed up by a strong evidence base (i.e., the literature is mixed on this point). In any case, the alternative – i.e., not to collect PRO data in non-blinded trials – would seem to be a case of “throwing the baby out with the bathwater.” We hope that the FDA will reconsider this issue and that the final document will be less dogmatic on this point.

On page 24 section B: *Frequency of measurements:* A statement should be added regarding the comparability of the arms. The timing of the PRO assessments should allow direct comparison between the two treatments.

### **Minor comments and observations**

On page 1: The document states *A PRO is a measurement of any aspect of a patient’s health status:* one can argue that PROs are not limited to health status only. Indeed, in previous FDA discussions, it was noted that PROs were used as a broad umbrella term, to also denote things such as satisfaction with care/treatment and adherence to treatment.

On page 3, between lines 82 to 89 this is a rather confusing section on the use and misuse of simple versus more complex PRO instruments. The considerations that are reflected in this section are exactly the reason why a multidimensional quality of life concept has been proposed and is so well accepted in the scientific and clinical community.

On page 4 it is not clear what *expert assessors* means or how this is defined. Also PRO instruments are not typically validated by comparing patients’ responses to those provided by *expert assessors*” Also on page 4, between lines 120 to 137 it is important that self-administered questionnaires are filled in by patients in a standardized situation and this argument should be made more explicit in this section. Furthermore, “informal interview” should not be mixed up with a scientifically sound qualitative research approach (numerous papers have been published on this in recent years often in the BMJ). There are situations

where most can be achieved when a formal quantitative assessment is combined with a qualitative approach (of course after the questionnaire assessment is finished to avoid any carry over or context effects).

On page 5 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.

On page 5 in table 1, under the section *Intended measurement populations and condition* should 'treatment specific' not be added this box with generic, condition and population specific box? In addition, 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). Also it would be helpful in this section to detailed points regarding the issue of PRO assessment at baseline and how that should be included

On page 8, the third paragraph, 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.

On page 9, the document states *A PRO instrument can be developed for a variety of roles, including defining trial entry criteria, excluding excessive severity, evaluating treatment benefit, or monitoring adverse events.* Is it really the intent of the FDA to allow PRO data to be used for purposes of defining and identifying adverse events in an RCT?

Still on page 11, table 2: This is an informative table, but missing the crucial difference between "intensity" and "frequency"

On page 12: The document states *responses choices are generally considered appropriate when the number of response options is justified.* How does the FDA expect the investigator to justify the number of response choices?

On page 13: *Equally weighted scores for each item are appropriate only when the responses to the items are relatively uncorrelated.* This whole section on combining items via weighted scores is somewhat confusing. First, it advocates the grouping together of uncorrelated items (which is in contrast with the principles underlying internal consistency reliability). Second, it gives no alternative on how to choose appropriate weighting coefficients. It seems only logical that if all items of a scale are highly intercorrelated and of equal importance, equal averaging would produce a single meaningful scale.

Also in the subsequent paragraph the document states, *When empirically determined patient preference ratings are used to weight items or domains, the FDA also intends to review the composition of samples and the process used to determine the preference*

*weights*. However, 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.

On page 14, lines 452 to 458 notes about *the burden* related to the filling in of questionnaires -- compared to all the medical test a patient has to undergo throughout the treatment that's nothing! Therefore, it would be helpful to give a more balanced account here.

On page 20, 597: the document states *An instrument that is developed and validated to measure one concept is used to measure a different concept. For example: A single domain from a multiple domain PRO is administered without the other domain* . This point (administering a single domain from a multiple domain PRO without the other domains) should not be considered a major measurement violation necessitating revalidation.

On page 24 section C: *Duration: Generally, duration of follow-up with a PRO assessment should be at least as long as for other measures of effectiveness*. This seems a little unusual. What other measures are meant?. The duration of the PRO assessments should be based on the PRO research questions being posed, independent of the other endpoints.

On page 24 the documents states: *The frequency of PRO assessment depends on the natural history of the disease and the nature of the treatment. Some diseases, conditions, or study designs may necessitate more than one baseline assessment and several PRO assessments during treatment. The frequency of PRO assessment should correspond with the demonstrated measurement properties of the instrument and with the planned data analysis*. 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.

Section D on page 25 presents is a rather general section. We wonder if it could be relevant to provide some recommendations about circumstances where it is justified to choose a PRO outcome as primary endpoint. Further, given that it is acknowledged that PRO outcomes may be adequate primary endpoints, it might be useful to add some considerations about sample size evaluation (possibly with section VI E) and to raise the issue of the definition of a minimally important difference. Also when including multiple endpoints (as is often the case with PRO measures), investigators should be required to define 1 or 2 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.

On page 27, section VI B (beginning with line 892): We are somewhat confused by the statement in the previous paragraph saying that (line 885) control for multiplicity is generally not a concern when all endpoints are shown to be superior to those of the comparison group. This needs clarification as aspects of this section are also rather vague.

On page 28, Section VI C: This section is a little confusing. On the one hand, it appears to note that researchers are discouraged from using a composite endpoint which indeed reduces the multidimensionality of the concepts measured, but the document also specifies that claims related to only one component are generally not adequate. It is of course subject of debate, but it could be considered adequate provided that the multiplicity problems were adequately addressed and we feel that the paragraph is somewhat contradictory being both against the use of composite endpoints and against conclusions derived from results on individual components. More specifically, how does the FDA propose to deal with "*composite endpoints*"? It would be helpful if a definition of composite endpoints specific for PROs is added. Is a scale composed of several items considered a composite endpoint? This seems to be suggested by the 1st paragraph of section VI C. If so, it seems inappropriate to analyze the individual items of a scale. Also the statement: *In general, if analysis of scores for the individual component endpoints of a composite shows the improvement is driven primarily by a single domain (e.g. performance of a specific activity), the findings for the composite score would not support a general claim (e.g. psychological or emotional benefit, or even general physical state if all that is shown is symptom improvement).* is not clear. It can be indeed that a single item drives the total change for a certain scale. This does not invalidate the claim that the scale is invariant.

On page 29 Section VI D: 2 Missing measurements. One important reason in oncology for missing data is patient's death, and it would be helpful to have some statements related to that aspect (e.g. endpoints like time to deterioration in PRO outcomes might be useful).

On page 30: for the interpretation of the study results, when clinical trials show small mean effect sizes, it is advised to examine the possibility that the mean improvement reflects very different responses in subsets of patients. Written in such a way, it seems to open the door for multiple subgroup analyses. It should be reminded that the subgroup analyses which will be performed in order to support the claims should be described in advance in the statistical analysis plan.

Overall, while the PRO guidance document is very detailed, some possible important areas to include in future versions of these guidelines could also be:

- Data management: It would be helpful to provide prior specification of rules of how missing/unreadable/ambiguous data are coded.
- Compliance is not greatly detailed. We think it may be helpful to note the need for adequate calculation and also the reporting of compliance figures should be more stressed. Also monitoring compliance can be used as a quality control measure.
- There was also somewhat limited mention of the frequent requirement to investigate the missingness mechanism seen in PRO data: the relation between missing data and biomedical factors can provide interesting information.

Finally, we should reiterate our view that the FDA PRO guidance document is an extremely important contribution and one that can provide invaluable guidelines for the use of PRO data in clinical trials. The EORTC hope that our comments are helpful in assisting you in revising the document and we would be more than happy to provide additional comments upon receipt of any further versions.

**Andrew Bottomley**  
Head, EORTC Quality of Life Unit  
On behalf of the  
EORTC Data Center,

**Neil Aaronson**  
Chairman, EORTC Quality of Life Group  
On behalf of the  
EORTC Quality of Life Group,

#### **List of scientific contributors to the review (in alphabetical order)**

Corneel Coens (EORTC Data Center)  
Laurence Collette (EORTC Data Center)  
Michael Koller (EORTC QLG Translation Committee)  
Marianne Paesmans (EORTC Data Center)  
Teresa Young (EORTC QLG Treasurer)

#### **List of scientific reviewers**

Françoise Meunier (EORTC Director General)  
Patrick Therasse (EORTC Data Center)  
Richard Sylvester (EORTC Data Center)