



Comments on Guidance for industry – Patient reported Outcome Measures: Use in Medical Product Development to Support Labelling Claims

Lyon, April 3rd, 2006

Dear FDA PRO Working Group Members,

We would like to thank you for providing us with the opportunity to comment on the “Draft Guidance for Industry or proposed rule: Guidance for industry – Patient reported Outcome Measures: Use in Medical Product Development to Support Labelling Claims (Docket # 2006D-0044).

The ERIQA Group members (European Regulatory Issues on Quality of Life Assessment – see detailed information on page 2 of this cover letter) have reviewed the document both individually and as a group. We have synthesized our comments and suggestions in two categories: (1) Major issues; and (2) Issues needing clarification/Minor issues.

We hope that our comments will prove useful in generating the definitive version of the document.

Yours Sincerely,

The ERIQA Group

Mission Statement

"Establishing principles and practices for the integration of Health-related Quality of Life outcomes in the drug regulatory process "

Objectives

1. To provide European regulatory authorities with guidance on how to assess the quality of HRQL studies in clinical trials, and on how to evaluate the validity of HRQL claims, for appropriate decision-making;
2. To establish the relevance of HRQL as a key outcome, i.e. a credible criterion of evaluation of medicines.

ERIQA Group – List of Members

Neil Aaronson, PhD

Head, Division of Psychosocial Research & Epidemiology
The Netherlands Cancer Institute
Amsterdam, The Netherlands

Catherine Acquadro, MD

Scientific Advisor
Mapi Research Trust
Lyon, France

Giovanni Apolone, MD

Laboratorio di Ricerca Clinica in Oncologia
Dipartimento d'Oncologia
Mario Negri Institute
Milano, Italy

Paulo Carita, MD

Director STATISTICS / PRO/
Knowledge Management
Global Health Outcomes & Market Access
SANOFI-AVENTIS
Bagneux, France

Olivier Chassany, MD, PhD

Direction de la Politique Médicale
Délégation Régionale à la Recherche Clinique
Hôpital Saint-Louis
Paris, France

Katrin Conway, MA

Managing Director
Mapi Research Institute
Lyon, France

Dominique Dubois, MD

PGSM Health Economics
Johnson & Johnson
Pharmaceutical Services
Division of Janssen
Pharmaceutica.N.V.
Beerse, Belgium

Birgit Gradl, PhD

SOLVAY PHARMACEUTICALS
Hannover, Germany

Asha Hareendran, PhD

Director, Outcomes Research,
Pfizer Group of Pharmaceuticals
Sandwich, - UK

Bernard Jambon

CEO MAPI Group
Lyon, France

Fred Jost, MSc

Head of Outcomes Research & Scientific Experts
Pharmaceuticals Division
F. Hoffmann-La Roche Ltd
Basel, Switzerland

Jeff Kirsch, PhD

GlaxoSmithKline
Director, Strategic Health Outcomes
Global Health Outcomes
Greenford, Middlesex, UK

Paul Kind

Outcomes Research Group
Centre for Health Economics
University of York
York, England

Patrick Marquis, MD

Executive Director
MAPI Values USA LLC
Boston, USA

Pauline McNulty, PhD

Vice President, Health Economics & Pricing
Pharmaceuticals Group Strategic Marketing
Johnson & Johnson
Raritan, NJ - USA

Carolyn Miltenburger, PhD

SCHERING AG
Head, Corporate Outcomes Research
BERLIN, Germany

Annoesjka Novak, PhD

Global Health Economics
NV Organon
OSS, The Netherlands

Elisabeth Piault, Pharm D

Senior Research Associate
MAPI Values USA LLC
Boston, MA - USA

Margaret Rothman, PhD

Executive Director
HE&P, PGSM
Johnson & Johnson
Raritan, NJ - USA

Pierre Philippe Sagnier, MD

Health Economics and Outcomes Research
BAYER PLC, Uxbridge, UK

Marianne Sullivan, PhD

Health Care Research Unit
Sahlgrenska University Hospital
Göteborg University
Göteborg, Sweden

Maria Watson, PhD

GlaxoSmithKline
Global Health Outcomes
Research Triangle Park
NC - USA

Ingela Wiklund, PhD

Health Outcomes and QOL
AstraZeneca R&D
Moelndal, Sweden

York Zoëllner, PhD

Health Economics Manager
SOLVAY PHARMACEUTICALS
Hannover, Germany

General comments

This is a generally well-written, well thought out, and balanced document. It carries view surprises, and generally reflects the state-of-the-art in PRO assessment.

The scope of guidance document is made explicit – the use of PROs as an efficacy end-point in clinical trials to support claims in the approved labelling if the claims are derived from adequate and controlled investigations using PROs that reliably and validly measure “the specific concepts at issue”. The guidance document also explains how FDA evaluates such instruments for their usefulness in measuring and characterizing the benefit of medicinal product treatment. The requirements in other terms is said to be the same as for other labelling claims.

The FDA will assess the ability of the PRO to measure the claimed treatment benefit, if it is specific to the intended population and if it is specific to the characteristics of the condition or disease. Broader claims must show how a specific symptom benefit translates into other specific end-points such as daily activities or psychological state.

It is appreciated that the FDA guidance document states that PROs measuring efficacy should separately measure potential adverse consequences.

It is very clear that the TPC structure is the driving force behind selection of PROs and their ability to capture the relevant claims.

Specific comments

Major Issues

IV. Evaluating PRO Instruments

Lines 178-179

Who will arbitrate the importance of a change? In other words, what would constitute the kind of evidence to decide what is an important change leading to updating measurement properties?

This is uncharted territory, and we believe that the guidance would benefit from clarifications about the criteria enabling to characterize the magnitude of a change that would trigger an update of the measurement properties.

Importantly, some changes are necessary to make to update the language or cultural context and should be regarded as improvements rather than a change.

See related comments on Lines 579 to 670.

Lines 231-234

Does that mean that when a score defined a priori as the primary PRO endpoint (with MID a priori defined as well) turns out to be statistically and clinically significant, it could still be rejected from inclusion in the label on the basis of a review of its sub-components?

Line 255

Why need to specify domain aggregation in advance?

Clarify if this statement is to be interpreted as “in general” or prior to analysing the data. Depending of the specific condition under study and the specifics of the treatment effect it might be appropriate to aggregate but in other cases not. So, fair to specify prior to analysis of phase III data but not necessarily before that.

Lines 275-279

“Ethnic identity” or “race” may not be relevant “categories” in Europe. And moreover, randomization will account for potential biases introduced.

We would like to have more information about the purpose of the FDA review on comparability.

Lines 302-308

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.

Lines 317-324

It is fair enough if the FDA intends to review the comparability of data obtained using multiple modes of administration. But are investigators’ toxicity ratings reviewed on the basis of who completes the forms (e.g., clinician versus nurse practitioner)? What’s good for the goose is good for the gander.

We would like to know how the FDA intends to review the comparability of data obtained when using multiple modes of administration. What will be the criteria used to determine whether pooling of results from the multiple modes is appropriate?

Lines 334-337

It would be very helpful if the FDA could advise the sponsors on the measures to ensure that patients make entries according to the study design. We acknowledge that this is an important issue, but extremely difficult to assess with certainty.

Line 363

To request a full a priori testing of item response scaling/distance could preclude the inclusion of newly developed instrument in pivotal trials as it usually requires larger sample sizes than needed in a traditional psychometric testing, and the model-dependent Rasch approach is still debated.

Lines 405-407

The creation of a manual at this stage is desirable but could be interpreted as an obstacle to the development of a new instrument.

Lines 464-467

It seems important to be able to revise/modify the conceptual model once the empirical validation data have been collected and analysed.

Line 478

The use of the term “sociodemographic” in this context might have strong implications.

For instance, does it imply the re-validation of the instrument according to the age, gender, etc. of the CT sample?

Line 483, Table 4, Column Interpretability

It is important to acknowledge that MIDs are related to baseline severity, how well the treatment works and if an active or a placebo comparator is used to mention a couple of issues.

Line 504

For newly developed instruments, it is unrealistic to request predictive validity data as it requires longitudinal data that are usually not there at that early stage. In line with table 4, what matters for new instruments is to have a pre-specified difference stated in the protocol as it implies that the instrument will be responsive to treatment-induced changes. To expect predictive validity on top of responsiveness could be unfairly punishing for new instruments.

Line 528

See comments line 275-79. In the European context nationality might be more appropriate.

Lines 579-670

General Comment: this section about the modification of an existing instrument raised a lot of concerns. Who will arbitrate the importance of a change? In other words, what would constitute the kind of evidence to decide what is an important change leading to updating measurement properties?

This is uncharted territory, and we believe that the guidance would benefit from clarifications about the criteria enabling to characterize the magnitude of a change that would trigger an update of the measurement properties.

Specific comments

- **Lines 581-588:** We appreciate the practical approach developed in this paragraph.
- **Lines 590-593:** In line with the paragraph above, we would appreciate a similar pragmatic approach. We consider that recommending additional validation to support the development of a modified PRO instrument when only one minor modification occurs is far too restrictive (e.g. wording or placement of instructions).
- **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.
- **Line 659:** Please clarify the term harmonization. Is it harmonization as in “International Harmonization” or should it be understood as “reconciliation” when the forward versions are reconciled at the beginning of the linguistic validation process?
- **Line 660:** In lieu of comparable measurement properties the evidence of the conceptual equivalence of the translations can be preferably provided, especially in Europe. The organization of patients cognitive debriefing in the target countries is one of the means to ensure that conceptual equivalence between the source and the target versions be retained and should be regarded as the way of “bridging” between languages and cultures.
For new instruments, it is now almost standard practice to collect data in more than one country as part of the initial validation plan. Together with a well conducted cultural adaptation process, and a test of between-country heterogeneity as a pre-requisite to the pooling of the trial data, it forms a solid (and achievable) basis for the proper interpretation of international trials results.
- **Lines 662-670:** This part is not clear. Does this mean that the “older generation” of instruments cannot be used? Many are currently used in clinical trials and a more pragmatic approach seems viable. And the risk rests with the sponsor.

V. Study Design

Lines 717-718

Please rephrase (“rarely credible” is a rather strong statement). There are simply too many situations when double blind is not possible or even desirable. It would not be appropriate to recommend against the collection of PRO data in studies that are not blinded.

In addition, the interpretation of non blinded data is not a PRO-specific question.

VI. Data Analysis

Lines 914-917

See 231-234 above. If a composite score was a priori validated and defined as primary PRO endpoint, would an a posteriori review of its subcomponents invalidate the overall result? This would lead to post-hoc multiple testing of subgroups rather than fewer more global tests, with implications for sample size and data interpretability.

Issues needing clarification/Minor Issues

I. Introduction

Line 36

We suggest to drop “extremely”, and to replace it by “very”.

III. Patient-Reported Outcomes – Regulatory perspective

Lines 135-137

PRO instruments are not typically validated by comparing patients’ responses to those provided by “expert assessors.”

Lines 153-156

The use of the term effectiveness in the context of efficacy does not seem appropriate.

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.

Line 164 – Table 1

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).

Line 166

The definition of a composite score is not clear, and we would appreciate some clarification about the definition used in the guidance.

Our understanding is that a composite score is defined as the combination of objective measures of disease activity with functional outcomes. A well-known example is the ACR20, a disease index used in rheumatology, and defined as 20% improvement in tender and swollen joint scores plus 20% improvement in three out of the following five parameters: patient’s global assessment, physician’s global assessment, level of pain on a visual analogue scale, health assessment questionnaire measuring function, and sedimentation rate.

Another more recent example is the combination of healing of esophagitis, symptom relief and HRQL end-points into a composite score labelled “true healing”.

IV. Evaluating PRO Instruments

Lines 218-221

It is not clear. We: do not see how a single item that is part of a multi-item instrument could inform the completeness of that instrument.

Line 222

“*This may be evidence...*” There is an a priori view here that a single general question has more validity than a multi-item questionnaire. In many cases, it could very well be that the single general item has poorer psychometric properties than the multiple one.

Line 225

It is an important component of that determination but not the only one.

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”)?

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.

Line 270

It is not clear what excessive severity means. Please clarify.

Line 271

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?

Line 299: What is an “adequate” number of patients? Please, clarify.

Lines 339-343

We would question that assertion and conclusion. There are instances where averaging over a period of time will be more accurate than using point estimates.

- **Line 343**

What is an appropriate recall period? For example, it is very common to ask patients to report their symptom experience of 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.

Lines 351-352 – Table 2, Column Description

Table 2 would benefit from the deletion of any types of judgments in the column Description. For instance, the sentence “*These scales often produce a false sense of precision*” should be deleted.

Line 362: How does the FDA expect the investigator to justify the number of response choices?

Lines 367-69

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).

Lines 388-394

The examples of changes outlined in this section are not specific or linked to the development of the instrument per se, but to its application as stated in page 21. These lines should be deleted.

Lines 413-414

The review of the FDA is going too far on the issue of appropriate intervals between response choices. In many cases the same response options are used across a range of available PRO instruments and often “copied” from the SF 36.

Lines 424-430

The FDA is going too far in its requirements. The most common problems reported are detected in the focus groups and make it into the final questionnaire representing items where a change can be observed if the new compound works and can in this sense be regarded as important as well. The last sentence (429/430) should be modified.

In addition, 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.

Lines 474-475

It should be possible to confirm the responsiveness of the PRO instrument during the Phase III trial, if it was not done during Phase II. The risk rests with the sponsor.

Lines 529-530

This sentence is a contradiction in this context, and should be deleted (i.e. because of not enough power to assess the results). This issue is rather related to the design of the study, inclusion and exclusion criteria. Randomization is an effective means of taking this into account and if not done the risk rests with the sponsor.

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?

V. Study Design

Line 712: Please develop acronyms (NDA/BLA/PMA).

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.

Lines 783-784

If this is requested how can one relate the PROs to other objective clinical endpoints? Please clarify.

Lines 791-98

I would 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 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.

VI. Data Analysis

Lines 924-954

Please clarify the use of “composite”. It might be more appropriate to refer to a global score in this context rather than to a composite score.

- **Lines 940-942:** When using composite endpoints based on a combination of clinical/ radiological/biological criteria, there is no expectation that all patients will be impaired for all criteria at baseline. In the context of PRO measures, the statement is difficult to interpret in the absence of a more precise definition (in Table 1) of what “composite” means.
- **Lines 951-954:** See 231-234 and 914-917. We have strong reservations with that statement in the context of a priori defined expected differences for primary PRO endpoints.

Lines 1000-01

Please reconsider the phrasing. There are cases when LOCF is acceptable. Common causes of discontinuation are lack of efficacy or side effects or both. Unless the PRO outcomes are captured at the time of discontinuation and used in a LOFC the true difference might be overlooked.

Lines 1028-31: See comments made on Lines 951-9954.