Volume 10 • Supplement 2 • 2007
VALUE IN HEALTH

Patient-Reported Outcomes to Support Medical Product Labeling Claims: FDA Perspective

Donald L. Patrick, PhD, MSPH[1], Laurie B. Burke, RPh, MPH,[2] John H. Powers, MD, FACP, FIDSA,[3] Jane A. Scott, PhD,[4] Edwin P. Rock, MD, PhD,[5] Sahar Dawisha, MD, FACP, FACR,[6] Robert O'Neill, PhD,[7] Dianne L. Kennedy, RPh, MPH[2]

[1]Special Government Employee, Food and Drug Administration and Professor, University of Washington, Seattle Quality of Life Group, Seattle, WA, USA; [2]Study Endpoints and Label Development Team, Office of New Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA; [3]Scientific Applications International Corporation, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD, USA (formerly with FDA); [4]Mapi Values Ltd, Cheshire, UK (formerly with FDA); [5]GSK Biologicals, Collegeville, PA, USA (formerly with FDA); [6]Division of General, Restorative, and Neurological Devices, Center for Devices and Radiological Health, Food and Drug Administration, Rockville, MD, USA; [7]Office of Biostatistics, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA

ABSTRACT

This article concerns development and use of patient-reported outcomes (PROs) in clinical trials to evaluate medical products. A PRO is any report coming directly from patients, without interpretation by physicians or others, about how they function or feel in relation to a health condition and its therapy. PRO instruments are used to measure these patient reports. PROs provide a unique perspective on medical therapy, because some effects of a health condition and its therapy are known only to patients. Properly developed and evaluated PRO instruments also have the potential to provide more sensitive and specific measurements of the effects of medical therapies, thereby increasing the efficiency of clinical trials that attempt to measure the meaningful treatment benefits of those therapies. Poorly developed and evaluated instruments may provide misleading conclusions or data that cannot be used to support product labeling claims. We review selected major challenges from Food and Drug Administration's perspective in using PRO instruments, measures, and end points to support treatment benefit claims in product labeling. These challenges highlight the need for sponsors to formulate desired labeling claim(s) prospectively, to acquire and document information needed to support these claim(s), and to identify existing instruments or develop new and more appropriate PRO instruments for evaluating treatment benefit in the defined population in which they will seek claims.

Keywords: clinical trials, FDA, patient-reported outcomes, PRO, QOL, statistical analysis.

Why a Guidance on Patient-Reported Outcomes (PROs)?

Frequently, it is not possible to evaluate effectiveness of new medical products (which include drugs as well as medical devices and biological products) without direct input from patients for whom the product is intended. Survival or changes in clinical tests may not be the only outcomes of interest. For example, in evaluating therapy for community acquired pneumonia, mortality is the critical outcome, and a measure of patient symptoms would be an important complementary outcome. In contrast, in evaluating therapies for osteoarthritis, patient symptoms are of central interest, and survival is not usually the most relevant outcome, because osteoarthritis does not directly affect mortality. In some cases, such as functional dyspepsia, patients' perceptions of symptoms (i.e., abdominal pain) and the symptoms' impact on functioning are the only outcomes relevant for evaluating therapies because no clinical test or physical evidence is available.

Clinical trials evaluating medical product effectiveness increasingly incorporate self-reported outcomes from patients, known in the context of clinical trials as PROs. These reports help to determine the magnitude of treatment benefit, that is, improvement in survival and how patients feel and function as a result of medical therapy. PROs are a category of outcomes that one can distinguish from other types of outcomes, including laboratory measures, clinician ratings, and caregiver reports.

A PRO is any report coming directly from patients, without interpretation by physicians or others, about how they function or feel in relation to a health condition and its therapy. PRO instruments (e.g.,

Address correspondence to: Donald L. Patrick, Special Government Employee, Food and Drug Administration and Professor, University of Washington, Seattle Quality of Life Group, Box 358852, Seattle, WA 98103–8652, USA. E-mail: donald@u.washington.edu

10.1111/j.1524-4733.2007.00275.x

 1098-3015/07/S125 S125–S137

questionnaire items, instructions, and guidelines for scoring and interpretation) are used to measure these patient reports.

The term "PRO" is often used to refer to the things being measured (i.e., concepts and domains (discrete concepts within a multidomain concept)), the instrument used to measure the concepts, and the actual end points (i.e., the outcomes as analyzed in a particular clinical trial). We advise, however, that it is critical to distinguish the concept and outcome one is attempting to measure, such as a decrease in pain, from the instrument used to make the measurement, and the end point used in the statistical analyses. Pain intensity, in this example, is the concept, decrease in pain intensity is the outcome, and change over a certain time interval in pain intensity as measured by a 10-centimeter visual analog scale (instrument) is the end point used in the analyses.

Claims are statements or implications of treatment benefit that appear in any section of a medical product's Food and Drug Administration (FDA)-approved labeling. Labeling refers to the medical product description and summary of use, safety, and effectiveness that the FDA must approve.* When sponsors use PRO measures to support labeling claims, the FDA holds them to the same regulatory and scientific standards as other measures used in clinical trials. Before licensing in the United States, a sponsor must provide evidence that a medical product is safe and effective for its intended use. For drugs and biological products, substantial evidence is the standard for making conclusions that a drug will have a claimed effect. It also requires confirmation that adequate and well-controlled investigations provided the basis for determining whether there is substantial evidence to support claims of effectiveness for new drugs. Medical devices for human use require reasonable assurance of safety and effectiveness using valid and scientific evidence. The Federal Food, Drug, and Cosmetic Act (the Act) specifies these requirements and the Code of Federal Regulations (CFR) further clarifies them.

To inform sponsors, clinicians, and researchers of FDA's current thinking on how best to develop and use PRO measures to support potential claims in product labeling, the FDA published the draft guidance for industry "Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims" [1]. The guidance provides FDA's current thinking on the use of PRO measures for sponsors of all types of medical products and provides consistency across three medical-product-specific Centers within the FDA, including the Center for Biological Evaluation and Research (CBER), the Center for Drug Evaluation and Research (CDER), and the Center for Devices and Radiological Health (CDRH). After the FDA addresses public comments, a notice that the final guidance has been released will appear in the Federal Register.

This article reviews selected major challenges in using PRO instruments, measures, and end point analyses to support treatment benefit claims in medical product labeling. Appropriate responses to these challenges highlight the need for sponsors to formulate desired labeling claim(s) prospectively, to acquire and link information needed to support such claim(s) (including an end point model and conceptual framework for the PRO instrument), and to identify, develop, or modify appropriate PRO instruments for evaluating treatment benefit in a defined population.

PROs, Quality of Life, and Health-Related Quality of Life

Quality of life (QOL) and health-related quality of life (HRQL) may be considered to be PROs, but the terms should not be used interchangeably. The term *PROs* addresses the source of the report, and not the concept or content of the report. The organizing term PRO would never be a claim, i.e., a label would not say *Product x improves PROs*. A label claim requires that the concept or the "thing" being measured is defined, for example, pain intensity. The draft FDA guidance describes HRQL as a multidomain concept that represents the patient's overall perception of the impact of an illness and its treatment. HRQL measures capture, at a minimum, physical, psychological (including emotional and cognitive), and social functioning.

Improvement in HRQL is a highly sought-after claim by medical product manufacturers. The draft guidance clearly states that HRQL should not be equated with QOL which is described as a general concept that implies an evaluation of the impact of all

*The term *medical product* includes drugs and biological products as well as medical devices. *Labeling* refers to the medical product description and summary of use safety, and effectiveness that must be approved by the FDA. See 21 CFR 201.56 and 201.57 for regulations pertaining to prescription drug and biological product labeling. For medical device labeling, see 21 CFR 801. **For drugs and biological products**, section 505(d) of the Act establishes *substantial evidence* as the evidence standard for making conclusions that a drug will have a claimed effect and states that reports of *adequate and well-controlled* investigations provide the basis for determining whether there is substantial evidence to support claims of effectiveness for new drugs. See 21 CFR 314.216 for a description of the characteristics of an adequate and well-controlled investigation. Part of these regulations is the requirement for well-defined and reliable outcome measures (21 CFR 314.126.(6)). See the guidance for industry *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products* for considerations concerning the quantity of evidence necessary to meet the substantial evidence standard. (http://www.fda.gov/cder/guidance/1397fnl.pdf) **For medical devices**, the Medical Device Amendments of 1976 to the Act established *reasonable assurance* of safety and effectiveness of medical devices intended for human use. See 21 CFR 860.7 for the types of *valid scientific evidence* used in the determination of reasonable assurance of safety and effectiveness of a medical device.

aspects of life on general well-being. QOL is affected not only by health status but also by other valued aspects of human existence such as a safe environment, adequate housing, guaranteed income and freedom. Therefore, QOL is not an appropriate outcome for evaluating a medical product. HRQL, by contrast, can represent a summation of how patients feel or function as a result of therapy. Adequacy of an HRQL instrument depends on its ability to measure concepts that are relevant to the medical condition, including the important positive and negative concerns of patients undergoing therapy. HRQL end points, like all other end points, must be indicators of clear and interpretable treatment benefit or harms in clinical trials.

Begin with the End in Mind

Wheel and Spokes Diagram

The development, modification, and validation of a PRO instrument usually occur in a nonlinear fashion with a varying sequence of events, simultaneous process, or iterations. The FDA draft guidance summarized this process in four major iterative steps illustrated in a wheel and spokes diagram and described the important considerations for each step. Figure 1 includes five major steps that we believe are a better reflection of the actual developmental process. When contemplating the use of a PRO instrument in a clinical trial, all five steps apply regardless of whether sponsors use an existing instrument, modify an existing instrument, or develop a new instrument. The purpose of the wheel and spokes diagram is to organize the development process and provide the path by which the PRO can lead to a claim, as shown in the hub of the diagram.

The draft FDA guidance emphasizes the need for documentation of PRO instrument development (including modification), assessment of measurement properties, implementation, analysis, and interpretation to support proposed labeling claims. Other articles in this volume also address these issues. Table 1 outlines major types of documentation recommended for PRO instruments at each stage. The FDA has not specified an exact format for submission of this documentation, but the elements of documentation are identified to provide sufficient information for FDA review of PRO instruments intended to support product development.

The Importance of Targeted Claims

FDA advises sponsors wishing to use PROs to first identify the claims they seek in the product labeling. Success in obtaining a labeling claim based on a PRO measurement depends on alignment of product development, PRO instrument choice or development, application, analysis and interpretation in the context of otherwise adequate and well-controlled clinical trials. To help specify potential PRO-based claims for drugs, sponsors may use the Target Product Profile (TPP), a summary of the drug development program described in the context of prescribing information goals [2]. A well-developed TPP can facilitate and improve communication concerning labeling goals and PRO development and use in clinical trials. The TPP embodies the notion of "beginning drug development with the goal in mind." Early in product development, as suggested in Fig. 1, sponsors should link targeted claims to hypothesized concepts. To be used effectively, a sponsor updates a TPP before each discussion with FDA throughout all phases of the Investigational New

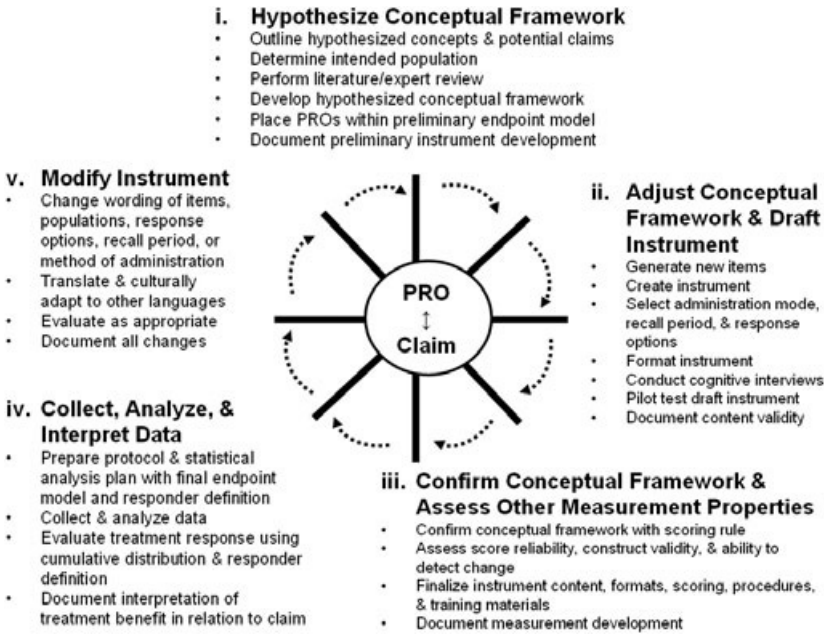

Figure 1 PRO instrument development and application in FDA. FDA, Food and Drug Administration; PRO, patient-reported outcome.

Table 1 Recommended Documentation for PRO Instruments Used in Clinical Trials*

Targeted or potential claims

1. List all targeted or potential treatment benefit claims for the medical product or device (Target Product Profile)
2. Identify potential claims to be supported by PRO instruments and end points
3. Link each targeted PRO claim to a hypothesized concept to be measured by PRO instrument(s).

Describe intended population (diagnosis, age, gender, or other characteristics)

Describe intended rationale and PRO instrument characteristics. Hypothesized PRO conceptual framework

1. For each PRO instrument comprised of single items or more than a single item, provide a diagram of concepts measured by each item, domain, or overall score, if applicable.
2. Provide summary of all literature reviews
3. Provide documentation on development of conceptual framework using literature review and expert input.

Model of hypothesized relationships among end points (end point model)

1. List all measures (PRO and non-PRO) that may be used as study end points in the clinical trial(s) to support claims. (This may include physiologic\lab\physical, caregiver, clinician-reported or patient-reported measures.)
2. Place PROs within this preliminary end point model\hierarchy
3. Describe hypothesized relationships among these measures.

Adjust conceptual framework and document instrument development, item generation, and content validity

1. Chronology of all item development activities
2. Protocols for qualitative interviews and focus groups, cognitive debriefing interviews and any other research used to identify concepts, generate items, or revise an existing instrument, including training of interviewers
3. Development of response options, modes of administration and scoring
4. Size, characteristics, location, and (if requested) transcript of each qualitative interview and focus group
5. Documentation on how saturation was achieved (i.e., no new information was obtained from additional qualitative interviews or focus groups)
6. Description of any pilot test, including cognitive interviewing transcripts (if requested)
7. Versions of the instrument at various milestones of development
8. Item tracking table that list the source of each item in the final instrument, and how it changed during development
9. A summary statement of qualitative research in support of content validity of the PRO instrument, i.e., how does the qualitative research listed above support the conclusion that the PRO instrument measures the concept(s) that it purports to measure and that are reflected in the proposed claims.

Confirmation of conceptual framework and assessment of PRO measurement properties

1. Protocols for PRO instrument development (design, methods, analysis plan)
2. Documentation of psychometric testing for each domain or summary score proposed as support for claims
   a. Confirmation of conceptual framework (concepts, domains, scores)
   b. Reliability
      i. Cronbach's alpha
      ii. Test–retest reliability
   c. Construct validity
      i. Convergent validity
      ii. Discriminant validity
      iii. Known-groups validity
   d. Ability to detect change
3. Descriptive and statistical analysis findings from each study
4. Estimate of patient burden
5. Instrument user manual that includes
   a. Procedures for PRO administration in its final format
   b. Scoring
   c. Final version of instrument in all alternate forms

Trial protocol-related documentation: in addition to usual protocol concerns*

1. The final version of the instrument planned for use in clinical trials.
2. Instrument administration procedures, training and instructions for patients and study personnel
3. Data collection, data storage, and data handling/transmission procedures, including ePROs
4. PROs in Statistical analysis plan
   a. Proposed Responder definition
   b. How between-group differences will be described (e.g., cumulative distribution function)
   c. Plans for confirmation of PRO instrument measurement properties within the clinical trial
   d. Plans to avoid missing data at both the instrument and patient levels
   e. Plans for multiple end point testing

Modifications of existing or new instruments

1. For language translations and cultural adaptation processes, include:
   a. Description of the expertise of the translators
   b. Description of procedures used (forward, back, reconciliation, harmonization, assessment of measurement properties)
   c. Description of patient testing
   d. Results of translation/adaptation including clear description of all translation issues and how they were resolved
2. For content, wording, format, or mode of administration changes, describe results from studies conducted to evaluate modification, or rationale for not conducting studies
3. For use in a new indication or new population, document instrument development and assessment of measurement properties as described above.

Bibliography

1. Provide listing and copies of all relevant published and unpublished documents.

*Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. (1988) Guideline for the format and content of the clinical and statistical sections of an application. See Appendix C, in particular. http://www.fda.gov/cder/guidance/statnda.pdf.
PRO, patient-reported outcome.

Table 2 Linking PRO concepts to claims

Desired claim		PRO concept		PRO instrument(s)
Product X reduces problems with swallowing and speaking to others and improves daily activities	→	Swallowing Speaking to others Daily activities	→	Swallowing diary and retrospective report Conversation diary Daily activities diary

PRO, patient-reported outcome.

Drug application process. As sponsors complete clinical trials, an updated TPP may serve as the basis for an ongoing dialogue on evolving goals. Although the approval process for medical devices does not use the TPP per se, the process of beginning with the desired claims and designing the studies to assess these claims is the same.

Table 2 illustrates the matching of desired product claims first to PRO concepts and then to PRO instruments. In this hypothesized example, the claim suggests that the medical product reduces problems with swallowing and speaking to others and improves daily activities. End point analyses would determine statistical significance and clinical meaningfulness of these changes. The concepts are clearly identified and the instruments are listed that operationalize the concepts and represent possible claims of treatment benefit. The actual measurement strategy relating to these claims and concepts could employ a single instrument with three domains that are specific to the intended measurement concepts. Alternatively, sponsors could use three separate instruments. Regardless of the measurement strategy, sponsors should provide copies of the actual instrument(s) specified and documentation supporting the adequacy of the instrument(s) to FDA reviewers.

Hypothesized Conceptual Framework

The adequacy of a proposed instrument to support a claim depends on the conceptual framework of the instrument, i.e., a diagram of the relationships between the questionnaire items in a PRO instrument and the concepts measured by the instrument and represented as scores [3]. An instrument may create a single score, thereby measuring a single concept, or it may be developed with multiple domain scores each represented by a concept within a more general concept of measurement. Table 3 illustrates a simple conceptual framework for a PRO instrument using three PRO concepts to assess outcomes of treatment for head and neck cancer as an example, namely, swallowing, speaking, and basic activities of daily living.

With a clearly stated target claim, identified concepts, and a hypothesized conceptual framework for the PRO instrument, the sponsor can review the scientific literature to determine whether an existing PRO instrument is adequate or if one needs a new or modified PRO instrument to secure the desired claim. The conceptual framework evolves and is adjusted after patient input and assessment of measurement properties, as shown in Fig. 1. The confirmed conceptual model is sometimes known in the field of outcomes measurement as the measurement model, including how items, domains, and total scores are derived.

Preliminary End Point Model

Early in product development, sponsors are encouraged to identify all measurement concepts—PRO and non-PRO—that may be appropriate for end point definition in clinical trials designed to support claims. To facilitate communication with the FDA, the sponsor may want to prepare a preliminary endpoint model (i.e., a plan for end point measurement) that specifies the hierarchy and hypothesized relationships among all treatment benefit end points intended to support claims. At a minimum, an end point model describes measurable concepts of a specific disease state, including the spectrum of both prominent symptoms and expected clinical course; additional treatment-specific concepts relevant to the patient population are incorporated into the model.

PRO end points are placed within this model within the hierarchy of all end points. In this way sponsors can lay out their hypothesized relationships among all measures—including PRO and non-PRO measurements—that could serve as end points in terms of overall goals of therapy. Articulation of an end point

Table 3 Three concepts of head and neck cancer: simple conceptual framework

Items		PRO concept
Difficulty in • Swallowing saliva • Swallowing liquids • Swallowing solid foods	→	Swallowing
Difficulty in • Speaking loud enough • Being understood by others	→	Speaking
Difficulty in • Eating • Dressing • Bathing • Toileting (using the bathroom) • Transferring (moving back and forth from bed to chair) • Remaining continent	→	Basic activities of daily living

PRO, patient-reported outcome.

Table 4 Hypothetical end point model for head and neck cancer

Progression concept	End point	Hierarchy for consideration of drug approval
		1 = By itself sufficient* 2 = If Overall Survival and Progression-Free Survival not in wrong direction, acceptable as a component of a time-to-event progression metric with swallowing, speaking, and activities
Longer life	Overall Survival (OS)	1
Better life—absence of disease progression	Progression-free survival (PFS) as determined via radiographic and clinical assessment	1
Better life—clinician-reported function	Clinician-rated performance status score	2
Better life—patient-reported function	Swallowing diary score	2
Better life—patient-reported symptoms	Conversation diary score	2
Better life—patient-reported function	Daily activities measure score	2

*If of sufficient magnitude in the context of an acceptable risk-benefit ratio

model ties together disease natural history, treatment goals, and the instrument(s) intended to demonstrate treatment benefit. This preliminary end point model also informs the development of the clinical study protocol(s) and its final depiction, after adjustment through the different phases of clinical research, is reflected in the Statistical Analysis Plan for registration or phase III clinical trials.

FDA, sponsors, and researchers always use an end point model, although they do not always explicitly state their thinking as such. End point definition is a basic process used in clinical trials. Formal articulation of the final model is not required at this point. The preliminary end point model provides a context to show how multiple end points fit together to support the primary hypothesis being addressed in a clinical study supporting a product approval.

Table 4 provides a hypothetical end point model for the head and neck cancer example. Clinical end points would include overall survival and progression-free survival as determined via radiographic and clinical assessment, which if significant statistically and interpreted as a medically important, would be sufficient for a claim. If these two survival end points showed a treatment benefit, a physician-reported performance status assessment, as often included in cancer clinical trials, would be acceptable as part of the analysis. The three PRO instruments are subsequently listed in order of importance as complementary end points that may result in a claim.

End point models should be linked to natural history for the condition and what is known about the relationship of concepts and end points. A clear end point model specifying outcomes that will be the basis for claims is important for all end points but is particularly helpful in elucidating the importance of PRO measures in the end point hierarchy that are intended for labeling claims.

FDA Review of Instrument Development

Food and Drug Administration determines adequacy of a PRO instrument based on review of the instrument's development, modification, adaptation, assessment of measurement properties, implementation in clinical trials, and interpretation of the PRO end points in evaluating treatment benefit. All aspects of use of a PRO instrument are evaluated in relation to the intended labeling claim. This allows FDA to determine whether the instrument provides a valid and reliable measure of the concept(s) implied or stated in labeling claims and provides sufficient evidence to support claims. This evidence, however, does not mean that sponsors must demonstrate the validity of a PRO measure "beyond a reasonable doubt." In any field of science, there is always a degree of uncertainty related to results of an experiment, but the goal is to minimize uncertainty to a degree that is acceptable for public health.

The draft PRO guidance does not set precise criteria for the level of evidence sufficient to demonstrate that a PRO instrument is adequate to measure a particular medical product claim. This allows FDA and sponsors flexibility when developing and reviewing an instrument and allows accommodations for changes and

Table 5 Simple item tracking matrix with swallowing measure as example

Long list of items	Item source	Final decision
How much difficulty do you have swallowing?	Qualitative interviews/focus groups	Dropped—high frequency, low severity, too vague in meaning, did not discriminate between severity levels. Patients reported difficulty in averaging over time in retrospective report
How much difficulty do you have swallowing liquids?	Qualitative interviews/focus groups	Retained—high frequency of report, high importance to patients, worked well in cognitive interviews and discriminated between severity levels. 24-h diary recall preferred by patients
How much difficulty do you have swallowing soft foods?	Qualitative interviews/focus groups	Dropped—highly correlated with swallowing solid foods and will be covered by that item
How much difficulty do you have swallowing solid foods?	Existing instrument—head and neck cancer swallowing diary	Retained—worked well in cognitive interviews and discriminated between severity levels 24-h recall period preferred by patients
How often do you need to spit?	Qualitative interviews/focus groups	Dropped—highly correlated with swallowing saliva—related to other conditions
How much difficulty do you have swallowing your saliva?	Qualitative interviews/focus groups	Retained—worked well in cognitive interviews—high importance to patients 24-h recall period preferred by patients

advances in the science of patient-reported measurements. Sponsors should be aware, however, that prior acceptance of an existing PRO instrument as an adequate measure of a concept for nonregulatory or regulatory purposes (including by FDA) does not ensure that the instrument is adequate by current standards to support FDA-approved medical product labeling claims. Furthermore "expert opinion" supporting "face validity" of an existing PRO instrument or its development does not constitute sufficient evidence. Aspects of instrument selection and design of the measurement strategy are addressed in Snyder et al. [4] and Turner et al. [5].

Evaluating Content Validity and Adjusted Conceptual Framework

The essential first step in developing a PRO instrument is ascertaining that the measured concepts cover what patients consider the most important outcomes of the condition and its therapy. This is often called establishing the content validity of the instrument. Unfortunately, researchers often give insufficient attention to this step and concentrate more attention on establishing other measurement properties.

It is important that one take into account patients' perspectives when developing PROs, as the whole point of the endeavor is to measure patients' experiences. The measurement properties of the instrument have little meaning if one is not measuring something important to patients. The qualitative interview strategy, description of qualitative interviews and focus groups, transcripts, coding procedures, and justification for each version of the developing instrument support the adequacy of the development process.

Qualitative methods include both item generation and cognitive. This documentation provides a record for the assessment of each version of the PRO instrument used throughout development so that FDA reviewers can track evolution of a PRO instrument from its inception to the version(s) used in trials. Also valuable for content validity review is a listing of all items with all changes made to each item as each instrument version is developed, the rationale for decisions to drop, retain, or modify items, and a record of items added or dropped. All changes in response options and proposed recall periods should also be documented.

Table 5 illustrates how to track the history of item development in an instrument, using swallowing as an example. The long list of items elicited during qualitative interviews and focus groups or from existing instruments provides the basis for the table. Each item is tracked through the different stages of development (item generation and cognitive interviews). The final column indicates whether each item is dropped or retained for assessment of measurement properties and, eventually, clinical trial analyses and gives the reasons or evidentiary support for each decision.

Confirming Conceptual Framework and Assessing Measurement Properties of a PRO Instrument

FDA reviews measurement properties of the version of a PRO instrument sponsors used in clinical trials to confirm the conceptual framework such that scores generated from the instrument are reliable, valid, and able to detect change. Measurement properties of such instruments are a function of many factors, including instrument formats and mode of administration, population studied, and properties of the scoring system [6]. Sponsors should base documentation of measurement properties on the PRO instrument version they plan to use in trials. In evaluating an instrument, sponsors should use relevant study populations and proposed scoring procedures that correspond to those used for evaluation of treatment outcomes with the PRO instrument in the clinical trial. Frost et al. in this volume addresses issues of reliability and validity [7].

Reliability

Some assessment of the extent to which a PRO instrument yields consistent, reproducible estimates is important. *Test–retest reliability* confirms that the same assessment collected at multiple time points using identical methods produces the same results if the patient's condition has not changed. If a sponsor measures a concept using multiple highly correlated PRO items summarized into a single score, the sponsor can measure *internal consistency reliability* of these items (generally measured using Cronbach's coefficient alpha [8]) to determine agreement among responses to different questions. If a single score is used to summarize domain or subconcept measures (e.g., severity of symptoms of a disease; HRQL domain scores), internal consistency of the subconcept measures is not an appropriate measure of reliability [9].

Evidence of internal consistency reliability is not sufficient to assure that scores are stable over time, that scores change with or without treatment, or that scores are valid measures of a concept [10]. Sponsors should include in documentation of reliability testing a description of populations assessed, rationale for the time period used to evaluate the instrument under stable conditions, and if applicable, how sponsors used internal consistency to evaluate domain and total scores.

Validity

Documenting validity of a PRO instrument, i.e., the extent to which the test measures what it purports to measure, involves demonstrating that the sponsor has tested the instrument in the population of interest (patients with similar disease severity, condition, language or culture, and demographic characteristics) and in a context roughly similar to that in which the sponsor will eventually perform clinical trials. As stated above, at a minimum, sponsors should demonstrate content validity of a PRO instrument based on studies showing that the PRO instrument captures the issues that patients in the target population indicate are important about the concept reflected in the claims. For example, a PRO intended to assess fatigue may lack content validity if it measures only affective manifestations of fatigue without considering behavioral or physical components.

Food and Drug Administration also looks for other empiric evidence of PRO instrument adequacy as a measure of the concepts associated with product claims, including the following:

- confirmation of the conceptual framework for the PRO, items, domains, and total score if applicable and how obtained;
- correlation of PRO data with other important end points as hypothesized in the end point model;
- ability to discriminate among patients that differ on important characteristics, e.g., disease severity;
- comparable results among different language versions or alternative modes of administration for the PRO instrument.

A single PRO instrument used in one setting or population may not be valid in another population or setting, therefore there is no universal "validation" of a PRO instrument. Consequently, measurement properties should again be confirmed in each clinical trial that will be used to support claims.

Ability to Detect Change

When one expects a concept to change with therapy, values for the PRO instrument measuring that concept also should change with effective treatment. If they do not, one should question validity of the PRO instrument. A small, blinded study completed before Phase III may be sufficient to document that scores change based on a PRO when the patient's status on the concept of interest changes, even if the study is not powered to demonstrate a statistically reliable treatment benefit. Ability to detect change usually is reflected by effect size statistics in which higher values indicate larger effect sizes. Larger effect sizes usually imply that a smaller sample size will suffice to evaluate therapeutic benefit using a PRO instrument.

Data Analyses and Interpretation of PRO End Points in a Clinical Trial

Interpretability means the degree to which one can assign easily understood meaning to an instrument's quantitative score in a particular application. Interpretation is separated in Fig. 1 as it may vary with the clinical trial population and protocol and thus is not considered a property of the PRO instrument per se even though accumulated evidence with individual measures may suggest a particular meaning of a score change.

One of the major concerns of clinicians regarding use of PRO instruments is uncertainty in translating a score into a meaningful treatment benefit for patients. When clinical trials use changes in PRO instrument scores as study end points, sponsors should provide interpretation of the magnitude of changes in PRO scores that patients can perceive and that patients consider beneficial.

The degree of change that patients would consider a meaningful therapeutic benefit may vary depending on patients' medical conditions, disease severity, or stage of disease. The magnitude of change important to patients may be different when measuring improvement compared to measuring worsening. FDA reviewers consider these factors when reviewing the documentation to support recommended guides for

interpreting study results based on a PRO end point measure. We discuss approaches to interpretation of PRO scores below in the section on analysis and interpretation. Sloan et al. [11] and Revicki et al. [12] in this volume also address these issues.

Modifying an Instrument

Any change in a PRO instrument can potentially affect its measurement properties. One can influence answers to self-administered questionnaires by changes in wording of items or response options, changes in item order, addition or elimination of items, changes in recall period, changes in mode of administration, and by changes in visual presentation of items and responses. Modifying PRO instruments may affect distribution of item responses, leading to changes in domain and overall scores. Survey research literature contains examples of these effects in many situations [13,14].

What Modifications are Important to Study

An important point of discussion with the FDA is the degree of modification that may occur to an existing instrument before it is considered a new instrument, in turn triggering the need for further investigation of measurement properties in the population of interest. Not all modifications necessitate testing the new instrument version in a randomized trial or even a large study of measurement properties. Yet when an instrument is modified, sponsors should consider additional studies for confirmation of the modified instrument's measurement properties.

The extent of additional studies depends on the nature of modifications made to the instrument [15]. Sponsors may use small methodological studies to compare results from the original and revised instruments. Cognitive interviewing is a qualitative method for assessing respondents' interpretation of a question or item. For example, one can use "think aloud" techniques such as ask respondents to describe how they arrived at responses to illuminate thought processes involved in answering questions. A small study, for example, may be adequate to assess results of changing presentation of a response scale from a vertical to horizontal axis. Sponsors would need to perform a larger study to investigate effects of adding new items or deleting entire domains or changing recall periods.

Using a PRO instrument in an entirely new population of patients may affect an instrument's measurement properties, for example, using instruments developed in adults with children or adolescents or in patients with various degrees of severity of a disease. If the sponsor plans to use a PRO instrument in an entirely new population of patients, the sponsor may need to perform additional qualitative studies to confirm that items and response options are relevant and understandable to the new population and to identify any new issues of importance to these patients. FDA recommends an additional study of measurement properties or investigation of measurement properties in a randomized study to ascertain measurement properties in the new population. Such investigation may minimize the risk that an instrument will not perform adequately in a study used to support labeling or promotion.

Finally, an instrument that a sponsor translates for use in a new language and culture is considered a modified instrument. Randomized studies to test each culturally adapted version of an instrument may not be possible given that sponsors use PRO instruments in many different languages and cultures. It may be possible, however, to verify basic measurement properties, such as distribution of responses, internal consistency, test-retest reliability, and validity in at least a few different languages used in the largest populations to be studies in the clinical trials. Alternatively, in a pivotal study supporting a label claim, sponsors may evaluate basic measurement characteristics such as distribution of responses, means, standard deviations, etc. for different language groups, when sufficiently large.

Issues in Study Planning, Analysis, and Interpretation

Design and analysis considerations for clinical trials that use PRO end points as the basis for labeling claims are the same as for trials using any other type of clinical end point. Principles of appropriate study design and analysis are well described in other FDA guidance [16]. In particular, sponsors should identify and specify a priori the end points of interest that characterize the desired claims. Trial end points should be consistent with other study objectives and must not be arbitrary or chosen retrospectively as part of an exploratory effort, e.g., data dredging after collection of data on numerous unrelated end points with the hope that a few will show an effect of treatment. The trial protocol should prospectively define the place of a PRO instrument in terms of the hypothesis tested in the trial. For instance, end points measured with PRO instruments can be sole end points, part of composite end points, or serially tested end points. As with any end points studied in a clinical trial, those without reproducibility or lacking an adequate ability to detect change will be detrimental to the success of the trial and to interpretation of results.

In the protocol, sponsors should consider and plan for the risk of false positive results by designating criteria for concluding a positive treatment benefit and how many ways treatment benefit might be achieved, especially if end-point-specific claims are desired. This

requirement is part of specifying the hypothesis to be tested in any trial.

Simultaneous statistical testing of multiple study end points can increase the chances of a false positive conclusion, by inflating the Type 1 error rate as a result of allowing so many ways to conclude that a treatment benefit on one or more end points exists. Prospective articulation of the trial's statistical analysis plan addresses this risk and minimizes the likelihood of false positive conclusions. Acceptable statistical approaches include use of "closed testing" procedures, gatekeeper/fixed-sequence methods, or assignment of allowable type 1 error (alpha, the prespecified false positive rate) to different families of end points. If there is a clinically relevant order or a hierarchy to PRO end points (i.e., if some PRO end points are more important to patients than others), then a gate-keeper/fixed-sequence approach may be most appropriate. None of these techniques are new or specific to PROs; published guidance describes these methods and their applications [16].

The study planning process should also include a consideration of issues related to missing data [17]. Patients who fail to complete a clinical trial or who withdraw from assigned therapy before trial completion will not provide outcome data or measures of treatment response. This concern is important in every clinical trial. Missing data because patients withdraw early from a trial is particularly problematic when measuring outcomes using PROs, because missing outcome or response data may be a predictor of an individual's satisfaction with the treatment while in the trial. For example, patients may drop out of a trial because they consider the therapy ineffective, because of adverse effects, or because of dissatisfaction with some other aspect of the clinical trial experience. Withdrawal in studies involving severely ill patients may be related to inability to provide data or visit the clinical center. Quantity of missing data, time points when data were missed, and treatment group(s) affected may each adversely affect ability to interpret trial results. Of particular concern is how likely it is that the analysis of data from individuals who complete the trial will yield an accurate and unbiased estimate of the real treatment effect of the therapy. In their protocols, investigators must plan for an analysis of all data on all patients randomized, i.e., intent-to-treat analysis, even while recognizing that incomplete patient response profiles may exist for some subset of patients.

Trial results are more likely to be biased when the amount of missing data is substantial. Ideally investigators will have devised a plan for obtaining data on each patient at the time of withdrawal to determine reasons for withdrawal and if possible at another time subsequent to withdrawal from assigned treatment exposure but before planned trial completion. This information is often useful with regard to understanding treatment response and perhaps can be taken into account in the analysis. Thus the statistical analysis plan should provide approaches for analyses of missing data (e.g., missing items within domains, missing entire domains, or missing entire measurements) when evaluating treatment efficacy.

Researchers generally use a PRO instrument to ascertain a patient's status at baseline or study entry and at subsequent time points during the clinical trial. A patient's response to treatment is measured by changes in PRO scores from baseline to the end of that patient's assessments during the trial. All clinical trials evaluate the effect of medical interventions in groups of patients who receive a medical intervention compared with the effect in patients who receive a control intervention (either active or placebo, or both).

Two main conceptual approaches exist to clinical trial analyses of patient response to treatment as measured by PROs: 1) comparison of the average change from baseline across all patients in treatment and control groups according to some between group criteria, also called the *minimum important difference* (MID); and 2) comparison of the proportion of patients in the treatment and control groups who meet a prespecified criterion for response, also called a "responder criterion." Each approach requires planning to support interpretation of study results.

Analysis of Mean Change

Change for a treatment group averaged over all patients in a treatment group relative to the average change in a control group (active, placebo, or both) may be viewed as overall treatment effect size. The treatment effect size is an estimate of the true unknown treatment effect that the trial is intended to measure. Within a clinical study, researchers may propose a theoretical treatment effect size they consider to be important or clinically meaningful. This theoretical treatment effect size is the MID between or among treatment and control group means. The MID is thus presumed to represent a meaningful treatment benefit when the treatment and control groups each are considered as a group.

Point estimates of the difference in means between two groups, however, may mask important changes for individual patients or types of patients in each group. This problem arises because some types of patients within each group will tend to respond better or worse than others, often depending on baseline values of the end point. For example, groups of patients at different levels of severity at baseline may show higher or lower changes as measured by the PRO instrument than patients at other baseline starting points. All patients, regardless of either their baseline status or their change from baseline, contribute to a group mean estimate of response to treatment; for that reason the point estimate of the group mean change can be large or small

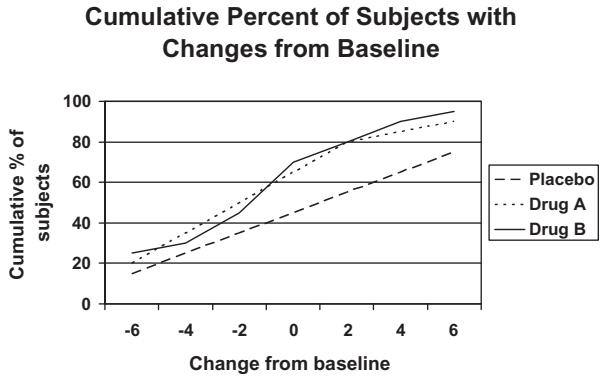

Figure 2 Example of a cumulative distribution curve.

depending on the distribution of individual PRO change scores within the group. Overall results may be driven by extremes of patient distribution at baseline or following treatment. The MID will thus not reveal whether some groups within the trial obtain a large benefit while other groups do not benefit at all.

Analysis of the cumulative distribution of patients' response to the experimental treatment within each group compared to responses of the control groups can help in evaluating the consistency of effects across the entire distribution. This distribution curve will reveal the extent to which overall results are driven by outliers who improve or worsen more than others. A cumulative distribution curve provides information on what type of responses contributed to the mean group response and provides more useful data than a simple point estimate of the difference between group mean changes.

Figure 2 provides an example of a cumulative distribution curve. FDA and sponsors have used this technique in product labeling to display treatment effect data [18–21]. The figure shows that change from baseline ranges from a negative 6 to a positive 6 with positive changes indicating improvement in the PRO. Fig. 2 shows that the two drugs (A and B) are clearly distinguished from placebo beginning at the no change point to 6 points above. Different cumulative distribution curves can be anticipated depending on the distribution of the effect as measured by the end point and its variance.

Results based on the MID may be more difficult for clinicians to interpret because the MID results are presented in terms of the score itself, rather than in terms of the proportion of patients who benefit from treatment. One may not be able to extrapolate differences among treatment groups to the magnitude of change in score that an individual patient may perceive as beneficial. This makes the determination of a clinically meaningful MID threshold challenging. For these reasons, FDA encourages sponsors to consider using cumulative distribution curves rather than MID criteria to demonstrate effect of treatment on PRO end points.

If sponsors do choose to use MID criteria in addition to the cumulative distribution to assess treatment benefit, the measured difference between the mean scores of each treatment group, usually including its entire confidence interval, should exceed the chosen MID. Fig. 3 illustrates that a difference in point estimates of approximately 10 points is the MID. Only the result at the far right of the figure in which the confidence interval exceeds the MID would constitute a "win." Nevertheless, the requirement that the confidence interval exceed the MID can be difficult to attain, because it entails modifying the null hypothesis from "is there a treatment effect greater than zero?" to "is there a treatment effect greater than the MID?" Achieving this objective may require a larger sample size than would otherwise be necessary or feasible in a clinical trial.

For these reasons, most commonly, only the point estimate needs to exceed the MID, and the lower bound of the confidence interval only needs to exclude no difference at all between groups. Therefore, most end point values are not judged against the requirement that the confidence interval exceed the MID. Regardless of the difficulty in interpreting the MID, estimates of the size of the treatment effect on each end point studied in a clinical trial is usually taken into account in benefit-risk judgments.

Analysis of Responders

An alternative way to analyze patient response to treatment as measured by PROs is to focus on within-individual patient changes in each treatment group and determine the proportion of patients who respond adequately to treatment. To compare the proportion of responders between treatment groups, one must establish responder criteria at the individual patient level that identifies within patient PRO changes known or

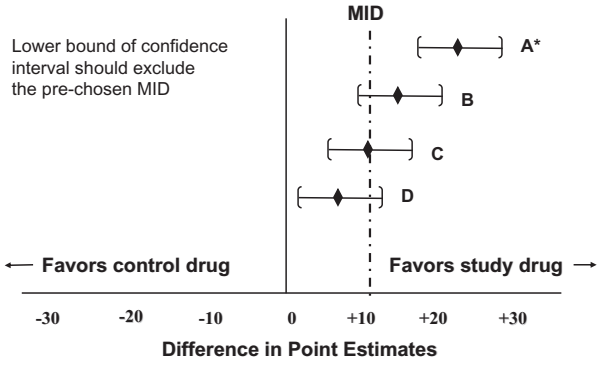

Figure 3 Minimum important difference—ruling out an important difference. *This result is only one that clearly exceeds the MID.

shown to be of perceptible importance and meaningful to patients. The responder criteria are used to measure the proportion or cumulative proportion of patients in each group who meet these predefined criteria of successful treatment response, e.g., the proportion of patients in the acetaminophen group who achieved a score of 2 or less on a 0–10 pain intensity scale compared to the proportion of patients in the opiate analgesic group meeting the same criteria.

In analyzing the proportions of responders, investigators must still interpret the degree of change that represents an important difference or clinical meaningful difference between the proportions of responders in each treatment group. Just as for analysis of means, examination of the cumulative distribution curve can be useful to interpret study results in an analysis of proportions.

A major difficulty with using a responder analysis is that sponsors must define criteria for an individual responder before starting the trial. This step may require some preliminary study to determine an appropriate value. Methodologies have been proposed to evaluate and define the minimum change that individuals perceive as being important and meaningful to them as part of the PRO instrument development process. However, appropriate standards have yet to be developed for responder definitions [11]. More research is needed in methods for identifying and applying responder criteria.

Conclusion: Toward Better Measurement of All Clinical Trial End Points

PROs and Clinical Trial End Points

Measurement principles stated in FDA's draft PRO guidance are an extension of principles that apply to all end point measures whether or not a PRO is involved. As stated above, a PRO instrument is but one way to define and measure end points for clinical trials. The draft guidance describes measurement properties of end points generally, with application to PROs in particular. All end points, however, should meet similar criteria. We hope that discussions regarding design of trials using PROs will shed light on issues regarding measuring other (e.g., laboratory, clinician-reported, caregiver-reported) end points as well.

The draft guidance provides recommendations, not requirements. No guidance can encompass all situations that may arise when designing clinical trials for a wide variety of disease areas. As science advances and new tools become available, FDA will continue to seek to consider new methods for demonstrating overall treatment benefit. Equally important, depending on the magnitude of effect observed, the Agency has considerable flexibility in allowing use of PRO instruments that may not fulfilll all of the principles articulated above.

PRO instruments may allow additional measurements of ways in which therapies may benefit or harm patients. For instance, a new therapy may have a similar survival advantage relative to existing therapies, but it may also provide other benefits in decreasing patients' symptoms or improving function among those who do survive. In addition, use of PRO instruments allows patients a voice in both development of clinical trial end points and measurement of benefits of therapeutic interventions. Finally, when investigators employ PRO instruments, they can provide information to patients that will enable them to assess the value of new therapies and to understand their treatment regimens.

In an age that is moving toward "individualized medicine" it seems logical that health-related needs as expressed by patients logically play a central role in the design and analysis of trial results. A successful medical product development program that employs a PRO instrument requires careful planning and clear objectives. Using the guidance as a basis for discussion between sponsors and FDA at all stages of product and PRO development offers the greatest prospect of success in demonstrating benefit from use of new therapies that will protect and promote the public health.

Novel End Points and Methods

PROs are routinely used in many therapeutic areas as clinical end points. Inevitably, of course, new PRO concepts and assessments will be proposed as the field of PRO research addresses assessment needs for new products. The FDA welcomes new approaches and ways of documenting treatment benefit, but the Agency will require time to understand the value, proper use, and interpretation of these innovations, depending on the speed of accumulating the evidence base and experience with innovative methods.

An important case in point is the developing use of new electronic methods for capturing and transmitting PRO data to the FDA, such as electronic diaries or momentary assessments, commonly referred to as e-PROs. Devices for collecting and transmitting data electronically offer many new ways to capture PRO assessments that have not been used in previous clinical trials with paper and pencil administration of PROs using retrospective recall periods. Sponsors, vendors for electronic monitoring and data capture, and the FDA all face a learning period to establishing confidence that the data provided are valid and reliable in a manner that is consistent with federal regulations and statutes, as well as with good clinical trials procedures. FDA welcomes the advancements promised by e-PRO assessment in trials but encourages sponsors to work closely with the Agency to ensure that the instrument and its implementation meet all the relevant requirements to support product claims.

Food and Drug Administration also is open to the use of new methods for evaluating the measurement properties of PRO instruments. Item-response methods or methods based on modern test theory offer new potential advantages over classical psychometric, such as dynamic testing. Applications using unfamiliar methods will likely require more review time so that reviewers can understand the methodology as well as learn to interpret the results. Sponsors are encouraged to provide reports that include traditional measurement property assessments as well as novel approaches to make the links among methods more apparent during FDA's review.

The draft PRO guidance acknowledges and reinforces the importance of PRO measures for understanding the treatment benefit of new therapies. It explains conditions for PRO use within the context of FDA regulations, and it highlights areas of research needed to resolve outstanding issues in PRO research for medical product development. With this guidance in place, sponsors, clinicians, researchers, and FDA will have a common understanding of FDA's concerns and a better opportunity to plan for successful clinical development programs so that safe and effective medical products can be made available to the US public as quickly as possible.

Acknowledgments

The authors would like to acknowledge the contributions of those FDA staff who participated in the February 2006 Mayo Clinic Workshop "FDA Guidance on Patient-Reported Outcomes: Discussion, Dissemination, and Operationalization:" Lisa A. Kammerman, Sandra L. Kweder, Scott Monroe, Elektra Papadopoulos, Bob A. Rappaport, William Pierce, Lilliam Rosario, and Stephen E. Wilson.

Source of financial support: None.

References

1 US Food and Drug Administration. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Available from: http://www.Fda.Gov/Cder/Guidance/5460dft.Pdf [Accessed May 4, 2007]. Federal Register: February 3, 2006, Vol. 71. Number 23) Docket no. 2006D–0044.
2 Delasko JM, Cocchetto DM, Burke LB. Target Product Profile: Beginning with the end in mind. Update 2005: January/February, 36–9.
3 Rothman ML, Beltran P, Cappelleri JC, et al. Patient-Reported Outcomes: conceptual issues. Value Health 2007;10(Suppl. 2):S66–75.
4 Snyder CF, Watson ME, Jackson JD, et al. Patient-reported outcome instrument selection: designing a measurement strategy. Value Health 2007;10(Suppl. 2):S76–85.
5 Turner RR, Quittner AL, Parasuraman BM, et al. Patient-Reported Outcomes: instrument development and selection issues. Value Health 2007;10(Suppl. 2):S86–93.
6 Fayers PM, Machin D, eds. Quality of Life: The Assessment, Analysis, and Interpretation of Patient-Reported Outcomes, 2nd edn. New York: Wiley, 2007.
7 Frost MH, Reeve BB, Liepa AM, et al. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 2007;10(Suppl. 2):S94–105.
8 Cronbach LJ. Coefficient alpha and the internal structure of test. Psychometrika 1951;16:297–334.
9 Bollen KA, Lennox RD. Conventional wisdom on measurement: Structural equation perspective. Psychol Bull 1991;110:305–14.
10 Patrick D, Erickson P. Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation. New York: Oxford University Press, 1993.
11 Sloan JA, Dueck AC, Erickson PA, et al. Analysis and interpretation of results based on patient-reported outcomes. Value Health 2007;10(Suppl. 2): S106–15.
12 Revicki DA, Erickson PA, Sloan JA, et al. Interpreting and reporting results based on patient-reported outcomes. Value Health 2007;10(Suppl. 2): S116–24.
13 Schwarz N, Sudman S, eds. Context Effects in Social and Psychological Research. New York: Springer-Verlag, 1992.
14 Schuman H, Presser S. Questions and Answers in Attitude Surveys. New York: Academic Press, 1981.
15 Groves RM, Fowler FJ, Jr, Couper MP, et al. Survey Methodology. New York: Wiley, 2004.
16 US Food and Drug Administration. ICH E9 'Statistical Principles for Clinical Trials.' Available from: http://www.fda.gov/cder/guidance/ICH_E9-fnl.pdf [Accessed May 4, 2007].
17 The European Agency for the Evaluation of Medicinal Products. Committee for Proprietary Medical Products: 'Points to consider on multiplicity issues in clinical trials.' Available from: http://www.emea.eu.int/pdfs/human/ewp/090899en.pdf [Accessed May 4, 2007].
18 Fairclough DL. Design and Analysis of Quality of Life Studies in Clinical Trials. Boca Raton, FL: Chapman and Hall/CRC Press, 2002.
19 Guyatt GH, Juniper EF, Walter SD, et al. Interpreting treatment effects in randomized trials. BMJ 1998; 316:690–3.
20 Premarin® Product Labeling. November 2005.
21 Aricept® Product Labeling. March 2005.