Volume 10 • Supplement 2 • 2007
VALUE IN HEALTH

Interpreting and Reporting Results Based on Patient-Reported Outcomes

Dennis A. Revicki, PhD,[1] Pennifer A. Erickson, PhD,[2] Jeff A. Sloan, PhD,[3] Amylou Dueck, PhD,[3] Harry Guess, MD, PhD,[4] Nancy C. Santanello, MD, MS,[5] the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group

[1]Center for Health Outcomes Research, United BioSource Corporation, Bethesda, MD, USA; [2]Department of Biobehavioral Health & Department of Health Evaluation Sciences, Pennsylvania State University, and OLGA (On-Line Guide to Quality-of-Life Assessment), State College, PA, USA; [3]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; [4]Department of Epidemiology, University of North Carolina, Chapel Hill, NC (deceased), USA; [5]Department of Epidemiology, Merck Research Laboratories, West Point, PA, USA

ABSTRACT

This article deals with the incorporation of patient-reported outcomes (PROs) into clinical trials and focuses on issues associated with the interpretation and reporting of PRO data. The primary focus and context of this information relates to the evidentiary support and reporting for a labeling or advertising claim of a PRO benefit for a new or approved pharmaceutical product. This manuscript focuses on issues associated with assessing clinical significance and common pitfalls to avoid in presenting results related to PROs. Specifically, the questions addressed by this manuscript involve: What are the best methods to assess clinical significance for PROs? How should investigators present PRO data most effectively in a Food and Drug Administration (FDA) application? In labeling or in a scientific publication? Guidelines for interpreting clinical significance of PROs and for comprehensively reporting on the methods, measures and results of clinical trials that incorporate PROs are important for clinicians, regulatory agencies, and most of all to patients. Clear specifications for considering a finding on a PRO measure, as clinically meaningful, need to be determined by instrument developers and psychometricians; they need to be reported for all clinical trials involving PRO end points. Clinical trial reports need to be comprehensive, clear, and sufficient to enable any reader to understand the methods, PRO measures, statistical analysis, and results.

Keywords: clinical significance, clinical trials, health-related quality of life, minimal important differences, patient-reported outcomes, statistical analysis.

Introduction

This article focuses on issues associated with the interpretation and reporting of patient-reported outcome (PRO) data. The primary focus and context of this information relates to the evidentiary support and reporting for a labeling or advertising claim of a PRO benefit for a new or approved pharmaceutical product. Nevertheless, the issues and recommendations discussed are not unique to pharmaceutical and regulatory applications. Therefore, the information may be generalizable to other clinical trials and settings that include PRO end points. For this article, we assume that the PRO is an important effectiveness end point in the study and that the intent of the clinical development research program is to achieve a labeling or promotional claim. We also assume that the PROs were selected based on strong rationale, that the instruments selected are credible and relevant, and have evidence supporting systematic development and psychometric qualities in the particular study population [1]. See other articles in this series regarding best practices for PRO instrument development and psychometric evaluation.

Ensuring fair and complete reporting of PRO end points based on a clinical development program for a new medication is important. The focus should be primarily on prespecified PRO end points and those end points that reach statistical and clinical significance criteria. Nevertheless, even with psychometrically sound measures, a priori specification of primary PRO end points, and well-designed clinical trials, unexpected patterns of findings may be observed. In these situations, it is important to report all of the prespecified PRO end points, whether they are supportive or nonsupportive of the treatment. Here, we focus on interpreting statistically and clinically significant PRO findings and the effective reporting and presentation of PRO data.

Address correspondence to: Dennis Revicki, Center for Health Outcomes Research, United Biosource Corporation, 7101 Wisconsin Ave., Suite 600, Bethesda, MD 20814 USA. E-mail: dennis.revicki@unitedbiosource.com

10.1111/j.1524-4733.2007.00274.x

 © 2007, International Society for Pharmacoeconomics and Outcomes Research (ISPOR) 1098-3015/07/S116 S116–S124

What Are the Best (Alternative) Methods to Assess the Interpretation of Clinical Significance for PROs?

Definitions and Methods of Minimal Important Differences

The minimal important difference (MID) has been defined as the smallest change in a PRO measure that is perceived by patients as beneficial, or that would result in a clinician considering a change in treatment [2]. The term MID has been widely adopted but the use of the term may be problematic in that it has been interpreted in many different ways.

Several anchor-based and distribution-based methods have been used to determine the MID for PRO measures [2,3]. Nevertheless, the current situation for determining the MID is evolving, and no clear consensus exists as to the recommended approach for determining the MID [2]. The usual approach is to estimate the MID using several anchor-based methods with relevant clinical or patient-based indicators, examine various distribution-based estimates (i.e., effect size, standardized response mean, standard error of measurement [SEM]), and then triangulate on a single value or small range of values for the MID. Confidence in a specific MID value evolves over time and is confirmed by additional research evidence. We all need to realize and accept that aspects of human measurement include some error and that no PRO measure is perfect and should not be expected to be perfect to be used in clinical research. Additionally, it is likely that the MID value varies by the population in which the PRO is used and that there is no single MID value for a PRO instrument across all applications and patient samples. The anchor-based methods for determining MID are preferred, with the distribution-based approaches providing supportive evidence.

Anchor-Based Methods

The anchor-based approaches use an external indicator, either clinical or patient-based, to assign subjects into several groups reflecting no changes, small positive changes, large positive changes, small negative changes, or large negative changes in clinical or health status [2–4]. The anchors can be clinical (hematocrit, American College of Rheumatology (ACR) response, clinician-rated change), or patient-based, such as global ratings of change or actual changes in PRO measures that have previously demonstrated MID and have been demonstrated to be responsive in the patient population. Regardless of the selected basis, investigators need to collect data to understand the practical value of scores based on the relationship to clinical outcomes (e.g., morbidity and death), patient behavior (e.g., adherence to medication use and resource utilization), and consequences on work (loss of productivity or working days).

We recommend that MID and clinical significance be based on multiple independent anchors. Selecting anchors should be done using criteria of relevance for the disease indication, clinical acceptance and validity, and evidence that the anchors have some relationship with the PRO measure. One important test conducted before calculating an MID is to determine the strength of the association of the anchor measure with the MID. An anchor that has a very low or even moderate correlation may provide misleading information in defining what is important to patients.

In general, patient-based anchors are considered important for estimating the MID for PROs. The degree of change observed may depend on the direction of change, and it may not be of consistent magnitude for those who deteriorate and those who improve. In addition, investigators need to consider what constitutes a clinically significant change in the target disease. For example, in some diseases (e.g., cancer, chronic obstructive pulmonary disease [COPD]) for which the goal of treatment is to maintain rather than improve function, clinically relevant change may be related to demonstrating a lack of disease progression in a treated group compared with an untreated group which is expected to deteriorate. In these cases, the clinical significance of differences between those who maintain clinical status and those who deteriorate may be of interest.

Distribution-Based Methods

The distribution-based methods include various forms of the effect size, standard response mean, and SEM methods [2–4]. Basically, these methods provide descriptive statistics on the magnitude of change observed in a study in standard deviation units or, in the case of SEM, reliability-adjusted standard deviation units. For example, the observed effect size may depend on the level of change found in a particular study, the impact of intervention used, the characteristics of the population being studied (e.g., disease severity, gender, sex, age), and the representativeness of the population studied. Effect sizes may vary widely according to condition. Small effect sizes may be meaningful to patients with severe disease, whereas only moderate to large effect sizes may be important to patients with milder disease. Although these indicators can be informative as to how large the change is, they do not provide any indication of whether the observed change is important to either patients or clinicians.

Determination of Change over Time

Several methodological issues and open questions remain to be answered in regard to the MID. Researchers need to understand the study design used to establish an MID for a PRO [5]. For example, they need to determine whether the MID was based on change within a population over time or calculated in a clini-

cal trial looking at change over time in a treatment group correcting for changes in a placebo group. In the first case, the MID may actually reflect a minimally important change within a group. One might want to apply the minimal change determined in an observational cohort to define what constitutes a minimal difference between treatment groups. This MID value can be used to evaluate within-treatment group or between-treatment group changes in the clinical trial population (e.g., calculating the percent of patients in each treatment group meeting the MID). Nevertheless, there is some uncertainty whether MIDs derived from observational studies can be applied in clinical trials which focus on mean differences between treatment groups. These issues have not been firmly resolved and little rigorous research has been applied.

Variation Based on Impact and Population

The MID of a PRO can vary according to the impact of the intervention and characteristics of the population being studied (e.g., disease severity, gender, sex, age) [6]. An MID value calculated for a PRO measured in a less severely diseased patient population may be different from the MID value calculated in a more severely diseased group. Even if an MID has been established within a broad range of patients and, hence, reflects a generally applicable MID, the use of the MID value to establish meaningful change in a given population may not be valid.

Even the global question used to anchor a PRO and how the data are combined may affect the calculation of an MID for that measure. A study of two different global questions showed different MIDs for the same PRO. Additionally, this same study demonstrated that the MID differed between groups of patients who improved and those who deteriorated over time even with treatment [7]. This finding is contrary to findings in some observational studies in which the MID for improvement and deterioration were the same [8]. Other researchers have observed asymmetry in improving and worsening groups [9,10].

Instrument-Specific Evidence from Published Studies

For research topics in which PROs may be more common in clinical trials, review and synthesis of previously reported data from clinical trials may help inform the choice of an MID. As a PRO instrument is incorporated into more clinical trials, for example the St. George's Respiratory Questionnaire (SGRQ) [11] in COPD studies, more evidence emerges on responsiveness and the value of the MID. Jones [12] reported, based on a review of completed clinical trials, that the MID for the SGRQ total score is 4 points. Conducting systematic reviews and meta-analyses of responsiveness and observed changes in PRO scores based on multiple clinical trials may be possible. Information from such work can help further understanding of the clinical significance and MID of PRO measures.

Triangulation

The use of multiple methods to determine the MID for a PRO instrument in a specific patient population typically yields a range of values for the MID. This is the essence of triangulation, which is, examining multiple values from different approaches and converging on a small range of values. We recommend that the different MID estimates be graphed to depict the range of estimates. To identify a single MID value (or narrow the range of MID values), we also recommend that the anchor-based estimates be assigned the most weight and experience from clinical trials be used to further support and perhaps further narrow the range of values. This MID range would evolve over time as more evidence becomes available from clinical studies. Interpretation of the MID from different anchors should also take into account the proximity of the anchor to the target PRO measure; that is, investigators should assign more importance to MIDs generated from closely linked concepts than to those from widely disparate concepts. The distribution-based methods can provide additional support, but they should not be the sole criteria for estimating an MID. In cases lacking any suitable anchor, however, the distribution-based approach may the best available method [13].

To arrive at a single MID value, we recommend that analysts conduct a systematic consensus process, based on Delphi methods, involving several clinicians and health outcome researchers. No consensus exists as to how much data are needed as supportive evidence for the MID of a PRO instrument. Clearly, more data and evidence are better, but a single, generalizable study with multiple patient-based and clinical anchors may be sufficient. If there are a range of MID estimates for a PRO instrument, based on relevant anchors, the investigator will need to select a priori a value for the clinical study and provide a rationale for selecting this MID value.

As with other aspects of construct validity, responsiveness and the MID value are confirmed based on accumulating evidence from multiple studies and, with additional data, we can be more confident in the MID value. It is unlikely that a single MID could apply to all applications and patient populations. For example, the MID derived for an asthma-specific quality of life measure in trials involving patients with only mild to moderate asthma may not be applicable or appropriate for clinical trials comparing an add-on treatment for patients with moderate to severe asthma [14].

Statistical and Clinical Significance of Prespecified PRO End Points over Time

Interpreting the PRO findings from clinical trials should be based primarily on those PRO end points

specified a priori that achieve statistical significance and that also meet clinical significance criteria. Clinical significance is most often based on a clearly specified MID for the PRO score that is relevant for the target population and study context. The primary PRO end points should be clearly stated in the statistical analysis plan and protocol. Other secondary PRO end points may be discussed, but these outcomes are of secondary importance and may be used to further support the primary PRO findings.

Making statements about PRO differences between treatment groups is most often done by comparing mean changes from baseline, modeling mean scores over time (as in a mixed model analysis of variance), or comparing the percentage of subjects in each treatment group exceeding an a priori MID. No consensus exists on the best way among various possible approaches, to compare treatment differences on a PRO end point. In general, the results of different methods are similar, but this pattern is not always observed.

Treatment differences on mean changes from baseline that exceed the MID for the PRO instrument do not imply that all subjects in the better treatment group achieved these improvements. A distribution of change scores exists such that some patients exceeded the MID value and others failed to achieve it. On average, however, the members of the treatment group with the higher mean change scores have done better than those in the group reporting lower mean change scores.

Often, to assist clinicians in interpreting PRO differences, analysts report and compare the percentages of subjects exceeding the MID value for the PRO. For example, a recent meta-analysis of the effects of omalizumab in moderate to severe asthma found a 1.6- to 2.0-fold increase in moderate (>1 point) and a 1.8- to 2.1-fold increase in large (>1.5 points) improvements in overall Asthma Quality of Life Questionnaire (AQLQ) scores [14]. These differences were apparent despite observing differences between active agent and placebo in the mean changes from baseline of less than the 0.5-point MID for the AQLQ. Clearly, there are real differences in AQLQ outcomes seen across these clinical trials, but differences between the trial populations and the asthma population used to determine the MID for the AQLQ may in part explain these findings [6].

Number Needed to Treat

Some outcomes researchers have suggested that the number needed to treat (NNT) may be an effective way to express clinical and PRO results that may be easier for clinicians to understand and interpret [15]. Basically, the NNT is the relative proportion of patients who achieved important PRO benefits from treatment. NNT values indicate the number of patients that would need to be treated to achieve the PRO benefit. Lower NNTs suggest more effective treatments. Previous research has demonstrated a relationship between the proportion of patients benefiting from treatment, the NNT, and the effect size [16]. Therefore, reporting the percentage of subjects by treatment group who benefit on a PRO end point may be just as effective as reporting the NNT. Note that the NNT still depends on an MID that has been developed for a PRO which is limited (as stated above) by methodology and patient population differences.

Reporting PRO Evidence and Studies

Patient-reported outcomes, along with clinician-reported outcomes, laboratory tests, and device measurements, are collected in clinical trials to evaluate the effectiveness of treatments. Instruments within each of the types of measurements have unique sets of characteristics that make them relevant for use in a given trial. Guidelines have been proposed to assist in the preparation of manuscripts and presentation of data. For PRO measures, in particular, guidelines have been proposed by Fayers and Machin [17], Revicki [18], Staquet et al. [19,20] and Sloan et al. [21].

In this section, we go beyond previously published guidelines by specifying the detailed type of information needed to describe adequately PRO instrumentation used in a study. Complete information is needed about the relationships among each instrument, the study population, and the interventions, as well as data collection procedures, methods of analysis, and methods of interpretation, for understanding the findings and for generalizing the results beyond any particular study.

By following the detailed reporting of information recommended here, users can compare PRO information across various studies, whether conducted as part of a submission to the Food and Drug Administration (FDA) or for other purposes. Consistent reporting can also lead to the creation of an evidence-based data set that can be used to support the construct validity of PRO instruments in a variety of settings. This information might simplify the FDA submission and approval process as well as contribute to a more informed instrument selection process in the future.

Description of the Instruments Used

A description of each PRO instrument used in any trial is critical for understanding the outcomes. In most publications, this information is minimal, consisting of the name of the instrument with key references to its development and to documentation of its measurement properties. This is typically accompanied by a statement that the instrument has been administered, scored, and analyzed according to the developer's documentation. Providing minimal information about

a PRO instrument in the Material and Methods section is acceptable only in trials that use an instrument without any modification.

A minimal description is satisfactory when the PRO instrument has been well documented by its developer and widely used, such as the St. George's Respiratory Questionnaire (SGRQ) [11] and the SF-36 [22]. Minimal information is insufficient, however, for instruments with little or no formal documentation. For these tools, sufficient information to be described in the Materials and Methods section includes the minimal details (given above) as well as domain names, recall period, and methods of scaling and scoring. All modifications to the original tool and/or conditions for use, e.g., changes in item wording, recall period, and scoring method, should be fully described to clarify interpretation and comparability of findings. A copy of the instrument should appear in the cited references, if possible, be readily available at a permanent website, or supplied by the authors on request.

Ideally, the Materials and Methods section should include all of the minimal and sufficient information as well as a copy of the instrument, in either the text or an appendix. This need not be in camera-ready format but, rather, may be in reviewable form that includes the item stem, response options, and recall period. With ready access to the tool, the reader can evaluate the instrument's content within the context of the trial and thus interpret the findings meaningfully.

If an instrument has been endorsed by a professional organization, such as the American College of Rheumatology, this point should be noted along with an instrument's description and taken as indication of the relevance of an instrument's content. For example, the ACR response criteria incorporate several PROs [23,24]. Endorsement also implies widespread use with results contributing evidence to support an instrument's construct validity. If an instrument lacks professional endorsement, then including a review version of the instrument in the article will help the reader determine its appropriateness. Also, without an endorsement and/or widespread use, evidence will need to be provided as to an instrument's validity for use in a particular study population.

Although reporting minimal and sufficient information seems basic, it is easy to find articles that cite incorrect instrument names and references and fail to describe changes in wording and/or scoring procedures. Most frequently, however, authors name and reference an instrument, but provide no information as to the conditions of its use. For example, in a review of approximately 100 articles reporting the use of either the General Well-Being or the Psychological General Well-Being Index [25] in clinical trials, 55% provided less than minimal information about how the instrument was actually used [26].

Description of the Study Population

Investigators should explicitly state all inclusion and exclusion criteria for the trial population; they should also describe the settings and locations of clinical trial sites. The same guidelines that apply to reporting clinical trial data involving clinical end points should also apply to studies with PRO end points. The external validity or generalizablity of PRO results should be apparent from the information provided regarding the study population. The inclusion and exclusion criteria define the study population and provide necessary information for understanding the characteristics of the study population.

Description of the Intervention

Details of the interventions intended for each group should be stated including the method, timing, and duration of treatment administration. Authors should describe the treatment given to the control group (more than merely stating that a group received a control regimen or "standard of care"), and they should provide information on the placebo treatment (if a placebo was used). Finally, they should state whether the treatment was masked (i.e., double-blind) or open label, as this may affect evaluation of effectiveness.

In addition, specifying the course of treatment is crucial so that the time points that the PRO instrument(s) was (were) administered can be examined in the context of treatment impact. From the description of the treatment(s) being evaluated and the timing of the PRO measurements (e.g., at baseline and weeks 2, 4, and 8), the relationship between treatment response and the likely ability of the instrument to have detected any change in status should be clearly understandable. For example, short reference periods (e.g., 4 or 24 hours) are better suited for assessing the impact of migraine treatment than are longer periods (e.g., 1 week or 1 month).

Data Collection Procedures

PRO instruments, when used in clinical trials, can be interviewer administered or self-administered either at the site of care by postal questionnaire, by telephone through an interactive voice recognition system, or via electronic data capture. Analysts should disclose the conditions of the administration site (e.g., a standardized location free of distractions) and any formatting specific to the study (e.g., large scale print to accommodate elderly persons with low vision or electronic forms of data capture). If the PRO measure was interviewer-administered, either in person or via the telephone, the extent to which interviewers were trained should be specified. Information on standardized training of study coordinators or patients (e.g., training videos) and use of practice diaries during the

placebo run-in period should also be provided. Such detail is important for understanding the nature of missing data and incomplete responses and potential sources of bias.

If the procedures used in the trial differ from those used in the development of the instrument then steps used to validate the alternative method of administration should be described or referenced. Although sponsors may use PRO data collected in a Phase III clinical trial to both validate a method of administration and evaluate treatment efficacy, they run the risk of not being able to use the PRO results for making a claim if the validation is unsuccessful. Thus, sponsors are ill advised to use a pivotal trial for PRO validation purposes. This same advice applies to using PRO data from a Phase III study to validate initially any aspect of an instrument that is also to be used as a trial end point. Developers can use clinical trial data as supportive evidence for the psychometric properties of a PRO but an independent validation study or another published study of the end point would ideally be available if sponsors intend for the PRO to be a primary end point.

Sample Size Determination

If a PRO is the primary end point, then the Materials and Methods section needs to state the methods and information used in determining the sample size; this requirement is similar to that for presenting equivalent information for a non-PRO end point. It is especially important to cite the source of the MID that is used so that its relevance to the current study is clear. With the increasing use of PROs in clinical trials, estimates of MID are more readily available from published sources. The power needs to be stated as well as any effort to oversample to account for anticipated dropouts, for those who prove to be ineligible, for key subgroups, or for missing data.

Many PRO instruments consist of multiple domains and subdomains, unlike laboratory tests, device measurements, and many clinician-rated assessments. Even if investigators intend to use a summary score (e.g., the Health Assessment Questionnaire Disability Index), the FDA may evaluate the subdomains. To accommodate this regulatory analysis, a larger sample size might be advisable. If the sponsor decides to use this strategy, the investigators should clearly state this point in the methods section.

Statistical Methods

In addition to sample size calculations, the Materials and Methods section needs to summarize the analysis plan to be used with the PRO data. The use of parametric or nonparametric methods needs to be stated. If a summary score is to be used to characterize patient response, the derivation of the score should be described and assumptions that might effect the interpretation of the findings outlined. Ideally, sensitivity analyses will be proposed to test the robustness of the assumptions and the findings summarized. Computer programs used in the scoring and analysis should be named and referenced. Procedures for handling missing data need to be stated and deviations from the instrument developer's specific instructions, when available, should be justified [27].

Presentation and Interpretation of Findings

Clear and thorough presentation is essential because PRO data provide the only statement of the patient's perspective on treatment impact. In addition, publication of the PRO data builds evidence to support the use of an instrument in a variety of populations and settings. That is, published findings help establish an instrument's construct validity.

Clinical trial data generally appear in tables or annotated figures. Tables are essential for presenting descriptive statistics that can be used by others to calculate estimates of effect sizes and perhaps other types of responsiveness statistics. This information should be presented for all PROs used in the clinical trial. Thus, PRO data need to be made publicly available, whether in journals or as part of a website of findings (perhaps as part of http://www.ClinicalTrials.gov).

Figures can be used to show relationships between PRO and clinical data. For example, Teeter et al. [28] used a scatter plot to show the absence of a relationship between FEV1 and total symptoms, thus illustrating the unique contribution of PRO data. Such plots can be more informative than a summary statistic such as a correlation coefficient. Bar charts or line drawings are also useful for showing either scores between groups at a particular time point or change from baseline scores. In these graphs, means, standard deviations, and sample sizes should be given to allow the reader to check for statistical significance or calculate responsiveness statistics. Charts without data are unacceptable unless the data are presented in accompanying tables.

For investigating the relationship between treatment groups in a PRO end point, the FDA in a recent guidance document [29] suggested a cumulative distribution plot (see page 19 in [29]) to supplement a table of sample sizes, means, standard deviations, and/or *P*-values for investigating differences between treatment groups. A cumulative distribution plot shows the distributional properties of the observed PRO end point data not readily extracted from a table of summary statistics. While a useful proposal, we present alternative graphics which may be easier for nonstatisticians to read (Fig. 1). A more detailed comparison of these alternative graphics is forthcoming.

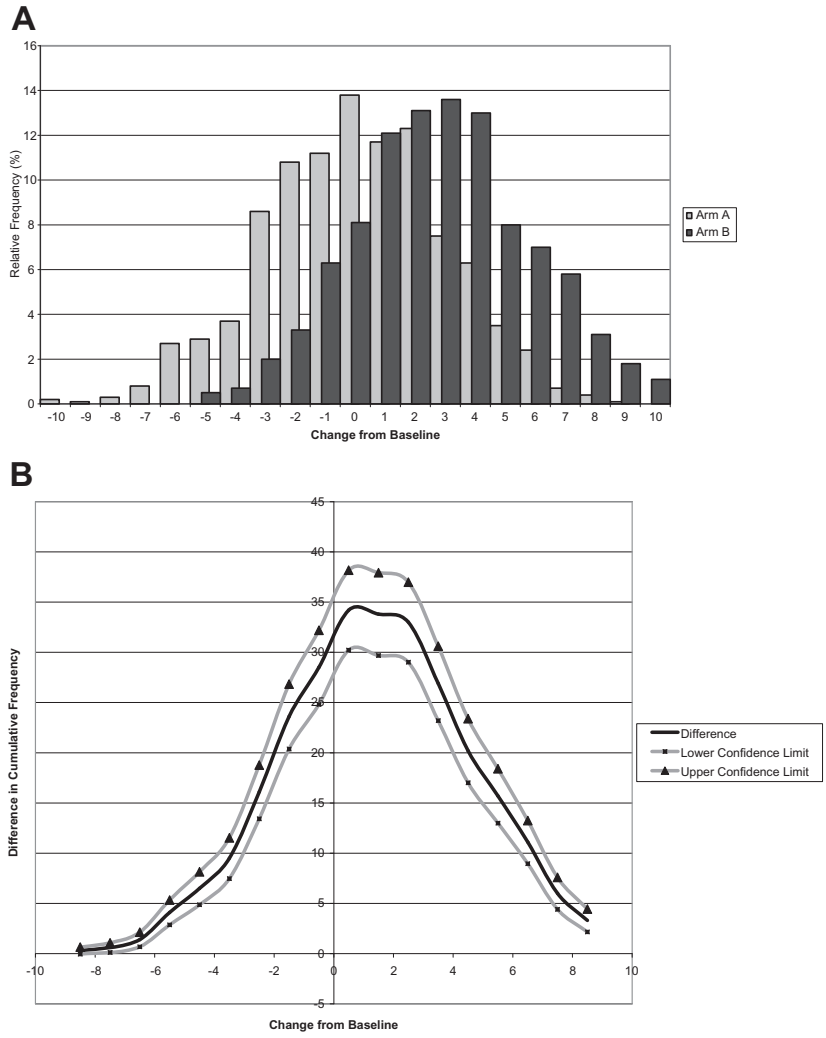

Figure 1 Overlaid (by treatment group) histogram (A) and difference between treatment groups in cumulative distribution plot (B) for change-from-baseline PRO data. PRO, patient-reported outcome.

Complete disclosure about the amount and type of missing data is imperative to the veracity of study results. Presenting a CONSORT diagram is a concise and complete method that has gained favor in recent years [18,30]. It delineates what happened to every person in the study from initiation to completion and provides sample sizes for the various analytical datasets (efficacy, safety, etc.). Alternatively, or complementarily, a table of the amount of missing data should be included to identify the exact number and proportion of missing data that occurred for each end point. Authors should comment about the amount of data that they consider be missing at random or missing owing to some systematic influence.

In reporting missing data, we recommend that authors include the following information: number of patients potentially able to provide data, number of patients who died, and number of patients who failed to provide data but were alive. This gives a clear indication of the proportion of patients who truly failed to provide data when they might have and indicates the degree to which missing data could have been avoided or may have influenced results. It also allows for an examination of differential dropout between treatment arms.

PRO data are usually one of several types of data collected in a clinical trial; other types include clinician-reported outcomes, laboratory tests, and biometric measures. Together, findings from these sources yield a composite picture of treatment efficacy and effectiveness. Relationships between the types of indicators need to be presented in a straightforward way that will be readily understandable.

For transparency, all results should be presented. What happens when there is a mix of positive and negative PRO results for an experimental treatment? For example, say a treatment group reported less fatigue (primary end point) but more diarrhea (secondary end point) than the control group. Interpreting such results will only be approachable on a case-by-case basis. The decision ultimately has to fit with the concept for the label claim. An analogy here is the listing of all potential adverse events that appear

during a clinical trial. It is not unreasonable therefore to suggest that any PRO domain that demonstrates a beneficial result might be included in a label claim or promotion as a potentially positive effect of a treatment, but that all other results, whether or not they are beneficial should also be presented to provide a full picture of the effects. The PRO labeling claim should be based on the a priori specified domains and whether between-group differences were statistically and clinically significant for these specified domains. Nevertheless, unexpected PRO findings, either positive or negative, may be seen in the secondary PRO end points, and these results may be informative to clinicians and patients.

Discussion

Guidelines for interpreting clinical significance of PROs and for comprehensively reporting on the methods, measures and results of clinical trials that incorporate PROs are important for clinicians, regulatory agencies, and, most of all, for improving the health care for patients. Clear specifications as to what is considered a clinically meaningful finding on a PRO measure need to be determined by instrument developers and psychometricians, and need to be reported for all clinical trials involving PRO end points. The reporting of clinical trial methods and findings need to be comprehensive, clear, and sufficient to enable any reader to understand the methods, PRO measures, analytic plans, and results. Clinical trials including PRO end points may be used for regulatory submissions to achieve labeling or promotional claims and/or to report on outcomes that are more meaningful to patients than the usual clinical end points, such as visual acuity or tumor response.

Summaries of PRO data collected in a clinical trial may be more detailed for an FDA submission than for a journal submission where page restrictions may limit the information provided. In either case, however, basic instrument-specific PRO data should be presented so that they can be used to build an evidence base for supporting the PRO labeling and for establishing construct validity and generalizability of findings. This evolving database is invaluable for understanding the relevance and validity (including responsiveness) of the PROs, and will help guide future clinical studies.

For all PRO instruments cited in product labels, a permanent, publicly accessible repository should be maintained either by the FDA (as part of the label information on the Drugs@FDA website: [http://www.accessdata.fda.gov/scripts/cder/drugsatfda/] or by the National Library of Medicine on the http://www.ClinicalTrials.gov website. This repository should contain a reviewable copy of each instrument, the user manual, documentation of reliability, validity, and responsiveness, as well as references to instrument-specific clinical trial findings. Submissions to this repository should be a mandatory requirement for making claims based on a PRO instrument regardless of whether the instrument is specifically named in the label. Availability of this detailed information will help researchers and clinicians understand and interpret findings and at the same time respond to critics' concerns about the perceived complexity of PRO measures.

The discussion on presentation need not significantly lengthen the Materials and Methods section of a journal article, especially if PRO instruments are used as developed and documentation about an instrument's development and measurement properties are publicly available. In this case, much of the required detail can be provided through a comprehensive list of references. When documentation is unavailable from a public source, then conditions of use of a PRO in trials intended for supporting a product claim through the FDA should be that 1) the developer/sponsor make the information available for public review; and 2) the regulator provide the access if the developer/sponsor is either unable or unwilling to provide this information. Clearly, full disclosure of information on the development, psychometric characteristics, and performance of PRO instruments in clinical trials is useful for reviewers and clinicians, and perhaps even most of all patients.

Acknowledgment

This article is dedicated to the memory of Harry Guess MD, PhD.

Source of financial support: Funding for the meeting was provided by the Mayo Foundation in the form of unrestricted educational grants; North Central Cancer Treatment Group (NCCTG) (CA25224-27) and Cancer Center grants (CA15083-32).

References

1 Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res 2000;9:887–900.
2 Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371–83.
3 Wyrwich KW, Bullinger M, Aaronson N, et al. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005;14:285–95.
4 Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221–6.
5 Sprangers MAG, Moinpour CM, Moyniyhan TJ, et al. Assessing meaningful changes in quality of life over time: a user's guide for clinicians. Mayo Clin Proc 2002;77:561–71.

6 Santanello NC, Zhang J, Seidenberg B, et al. What are minimal important changes for asthma measures in a clinical trial? Eur Respir J 1999;14:23–7.
7 Barber BL, Santanello NC, Epstein RS. Impact of the global on patient perceivable change in an asthma specific QOL questionnaire. Qual Life Res 1996; 5:117–22.
8 Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important chance in the disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81–7.
9 Cella D, Hahn EA, Dineen K. Meaningful changes in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207–21.
10 Yost KJ, Cella D, Chawla A, et al. Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. J Clin Epidemiol 2005; 58:1241–51.
11 Jones PW, Quirk FH, Baveystock CM. The St George's respiratory questionnaire. Respir Med 1991;85(Suppl. B):S25–31.
12 Jones PW. Interpreting thresholds for clinically significant changes in health status in asthma and COPD. Eur Respir J 2002;19:398–404.
13 Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life. The remarkable universality of a half a standard deviation. Med Care 2003;41:582–92.
14 Niebauer K, DeWilde S, Fox-Rushby J, Revicki DA. Impact of omalizumab on quality-of-life outcomes in patients with moderate to severe allergic asthma. Ann Allergy Asthma Immunol 2006;96:316–26.
15 Guyatt GH, Juniper EF, Walter SD, et al. Interpreting treatment effects in randomized trials. Br Med J 1998;316:690–2.
16 Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Med Care 2001;39: 1037–47.
17 Fayers P, Machin D. Quality of Life Assessment, Analysis and Interpretation. New York: John Wiley, 2000.
18 Revicki DA. Reporting analyses from clinical trials. In: Fayers P, Hays R, eds. Assessing Quality of Life in Clinical Trials, 2nd edn. New York: Oxford University Press, 2005.
19 Staquet M, Berzon R, Osoba D, Machin D. Guidelines for reporting results of quality of life assessments in clinical trials. Quality Life Res 1996;5:496–502.
20 Staquet MJ, Berzon R, Osoba D, Machin D. Guidelines for reporting results of quality of life assessments in clinical trials. In: Staquet MJ, Hays RD, Fayers PM, eds. Quality of Life Assessment in Clinical Trials: Methods and Practice. Oxford: Oxford University Press, 1998.
21 Sloan J, Symonds T, Vargas-Chanes D, Fridley B. Practical guidelines for assessing the clinical significance of health-related quality of life changes within clinical trials. Drug Info J 2003;37:23–31.
22 Ware JE Jr, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey: Manual and Interpretation Guide. Boston, MA: Health Institute, 1993.
23 Felson DT, Anderson JJ, Boers M, et al. The American college of Rheumatology: preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.
24 Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology: Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
25 Dupuy HJ. The Psychological General Well-Being Index. In: Wenger NK, Mattson ME, Furberg C, et al., eds. Assessment of Quality of Life in Clinical Trials of Cardiovascular Therapies. New York: Le Jacq Publishing, 1984.
26 Erickson P. Online Guide to Quality-of-Life Assessment (OLGA-QoL). Available from: http://www.olga-qol.com/index.html [Accessed online February 28, 2007].
27 Sloan JA, Dueck A, Erickson PA, et al. Analysis and interpretation of results based on patient-reported outcomes. Value Health 2007;10(Suppl. 2): S106–15.
28 Teeter JG. Use of pulmonary function tests in the diagnosis and management of asthma. Clin Pulm Med 1999;6:211–7.
29 U.S. Department of Health and Human Services Food and Drug Administration. Guidance for industry clinical studies section of labeling for human prescription drug and biological products—content and format. January 2006. Available from: http://www.fda.gov/Cber/gdlns/clinlab.pdf. [Accessed March 27, 2007].
30 Moher D, Schulz KF, Altman DG, the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001;357: 1191–4.