Volume 10 • Supplement 2 • 2007
VALUE IN HEALTH

Analysis and Interpretation of Results Based on Patient-Reported Outcomes

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

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; [2]Department of Biobehavioral Health & Department of Health Evaluation Sciences, Hershey Medical School, Pennsylvania State University, State College, PA, USA; [3]Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, NC (Deceased), USA; [4]Center for Health Outcomes Research, United BioSource Corporation, Bethesda, MD, USA; [5]Merck Research Laboratories, West Point, PA, USA

ABSTRACT

This article is part of a series of manuscripts dealing with the incorporation of patient-reported outcomes (PROs) into clinical trials. The issues dealt with in this manuscript concern the common pitfalls to avoid in statistical analysis and interpretation of PROs. Specifically, the questions addressed by this manuscript involve the analysis pitfalls with PRO data in clinical trials and how can they be avoided (e.g., missing data, multiplicity, null results etc.). The manuscript provides key literature for existing resources and proposes new guidelines.

Keywords: clinical significance, minimally important difference, missing data, patient-reported outcomes, statistical analysis.

Introduction

This article deals with issues of statistical analysis and interpretation of patient-reported outcome (PRO) data. The primary focus and context relate to supporting a labeling or advertising claim of a PRO benefit for a new or approved pharmaceutical product. The issues we discuss are not unique to pharmaceutical and regulatory applications, so the information may be generalizable to other clinical trials of other medical interventions involving PRO end points. For this article, we are assuming that the PRO is an important effectiveness end point in the study and that the intent of the research program is to achieve a labeling or promotional claim. We are also assuming that the PROs were selected based on a strong rationale, are credible, are appropriate, and have evidence supporting systematic development and psychometric qualities in the particular study population [2]. Other articles in this series focus on best practices for PRO instrument development and psychometric evaluation [3–9].

In addition, we think that achieving a PRO claim must require the a priori specification of a statistical analysis plan. We do not endorse basing a PRO claim on a post hoc analysis. The statistical analysis plan should be prescriptive and restrictive in terms of the analysis undertaken. The statistical analysis plan should detail the methods for handling missing data, multiplicity, and other relevant issues associated with the PRO data analysis. Any post hoc analysis that produces supplementary results should be considered interesting findings that should be confirmed in future studies.

This article focuses on four major areas related to statistical analysis of PRO data. In the first section of the article, we deal with commonly seen pitfalls including missing data, multiplicity of end points, blinding, choosing the correct end point for analysis, and the role of sensitivity analyses. In the second section, we discuss what to do when results turn out to be nonsignificant.

Address correspondence to: Jeff A. Sloan, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. E-mail: jsloan@mayo.edu

10.1111/j.1524-4733.2007.00273.x

What Are the Analysis Pitfalls with PRO Data in Clinical Trials and How Can They Be Avoided?

Missing Data

Much has been written about missing data and the analysis of PRO data [10–13]. Several methodological approaches have been proposed and tested in clinical applications in recent years. Missing data can have two major impacts on statistical analysis of clinical trials data. At a minimum, the missing data can produce wider confidence intervals and reduced statistical power to detect a treatment effect. At worst, missing data can distort treatment effects. The present recommended practice is to assess the degree to which missing data may affect the results of the clinical trial

1098-3015/07/S106 S106–S115

and thereby be considered "nonignorable" [13]. The primary goal of analyzing missing data is to consider whether a systematic influence or covariate could have introduced bias into the analytical results. For example, missing data closely linked to patients' health and PRO may bias the estimates of treatment effect. Wherever possible, the source of the missing data should be uncovered so that the study data set is as complete as possible and scientifically supportable. Data cleaning for PRO assessments should be handled in the same way an analyst would handle a missing laboratory test or clinical measure.

The best solution for missing data is, of course, to avoid missing data entirely by obtaining perfect compliance with protocol data collection requirements. Although this ultimate goal is not realistic, the amount of missing data can nonetheless be minimized by simple methodological practices. These include:

1. Treat PROs like any other end point in the trial. The PRO end points should be integrated into the data collection and clinical trial protocol. The PRO measures should not be seen as an additional burden but rather as a vital part of the scientific goals of the study.
2. Identify key personnel to oversee and coordinate the PRO aspects of the trial and serve as sounding boards and feedback experts for the trial.
3. Assure that the PROs will be administered in a standardized fashion in all study sites. Training of study site personnel in the administration of PROs and in the review of completeness of PRO measures is essential.
4. Consider the presentation format and patient burden of PRO data collection instruments from the patient perspective. Would they want to answer all these questions? Is there anything confusing, invasive, controversial, unnecessary, unclear, or omitted? Make the PRO instrument professional-looking and consider the formatting and appearance of the data collection instruments.

How much missing data renders various forms of imputation and statistical analysis invalid? Fairclough [10] and others have indicated that if more than 50% of the data for a particular end point is missing, then any analysis likely should not be undertaken because the analysis will be based on as much data present (with error) as are missing been acquired. If between 20% and 50% of data are missing for an end point, the analysis can be seriously compromised, but it can still be undertaken with a thorough checking of underlying analytical assumptions and extensive sensitivity analyses. If fewer than 20% of the data are missing for an end point, there may not be a major impact on study findings, but caution should still be employed and selected sensitivity analyses are recommended. If the amount of missing data is below 10%, the potential bias is slight and a simple imputation approach may be sufficient. The amount of missing data anticipated may influence the choice of PRO instruments and/or the PRO end point definitions.

Statistical analyses should be based on all of a subject's available data. Under no circumstances should only complete cases be used in an analysis; numerous studies have demonstrated that patients who provide complete data tend to be the "best performers" with better PRO scores than are representative of the general population [10,14]. Therefore, if a subject has data missing on the entire questionnaire on a particular occasion, that subject can still contribute on other occasions when her scores are available. If a subject discontinues or drops out of the study, the end of treatment visit data should be used as the subject's final assessment. We next discuss handling missing items within a PRO assessment and handling missing forms for individual patients.

Missing Items within a PRO Assessment

The handling of missing items on a particular questionnaire should be based on the recommendations given in the scoring manual of that questionnaire or, if not provided, by other forms of guidance (e.g., email or phone correspondence) from the developer of that questionnaire. If those recommendations are not available, missing items can be imputed for multiitem scales by considering alternative approaches. For example, if at least half of the items from the scale have been answered, one alternative approach is to assume that the missing items have values equal to the average of those items that are present for the respondent [15]. For instance, role functioning and cognitive functioning of the European Organization for Research and Treatment of Cancer (EORTC) each contain two items; thus, analysts can estimate these scales whenever one of their constituent items is completed. Physical functioning (PF), by contrast, contains five items, so at least three items need to be completed. For example, if only Question 3 (Q3) is missing on PF, then we would compute the Raw Score and PF Score as follows:

$$\text{Raw Score} = (Q1 + Q2 + Q4 + Q5)/4$$

$$\begin{aligned}\text{PF Score} &= [1 - (\text{Raw Score} - 1)/1] \times 100 \\ &= (2 - \text{Raw Score}) \times 100\end{aligned}$$

This approach is algebraically equivalent to using all items that were completed. In other words, the missing items are simply ignored when making the calculations. Otherwise, the subscale score must be considered missing.

Occasionally, using common logic, it is acceptable to complete an individual assessment form on behalf of the respondent, but this must be done transparently (i.e., reported) and with sound scientific rationale. For

example, in the McCorkle Symptom Distress Scale, there are two items for nausea; one for frequency and the other for severity. Patients without any nausea commonly indicate so on the frequency item and then leave the intensity item blank. In this situation investigators can reasonably complete the intensity item as indicating the lowest possible score as that conclusion is clearly logical. We recommend that these logic-based imputation methods be specified a priori in the protocol.

Missing Forms for Individual Patients

Investigators should summarize the total number (and percentage) of and reasons for missing forms by treatment. The analysis should be based on all available observations during the treatment period. The repeated measures, mixed effects model, which considers all measurements across time for each subject, can provide a simple yet sensible approach to address missing PRO data [11,16–18]. The validity of this model is based on the assumption that the missing data are missing at random, conditional on both model covariates, and observed values of the dependent variable. Under this assumption of ignorable nonresponse, statistical tests from the mixed effects model are expected to be valid and unbiased. Even when the data may not be entirely missing at random, results from a mixed effects model analysis of repeated measures may be quite robust when observed covariates, such as baseline score on the PRO assessment which may explain much of the missing data, are included [13,19].

To examine the association of missing data with observed changes in the PRO scores, analysts should include dropout patterns and depict average PRO scores stratified by time of dropout (by treatment and across treatments) [13]. These dropout patterns can be coupled with the descriptive profiles on the total number (and percentage) of and reasons for missing forms, to provide a better understanding of the mechanisms for missing data. These analyses can assist in determining whether the missing PRO data are most likely ignorable or nonignorable. We also recommend examining baseline demographic, clinical, and PRO differences between those subjects providing complete PRO data and those subjects with various amounts of missing data.

If the difference in the proportion of dropouts (and reasons for dropout) between treatment arms is substantial over time, results of the repeated measures model should be extended with those from a pattern mixture model. This is a more complicated model than the mixed effects model and it does not assume an ignorable missing data mechanism [11,13,17]. With a pattern mixture model, parameter estimates can be obtained by averaging over the missing data patterns. When there are nonignorable missing data, the recommendation is to use a pattern mixture model as the primary model [20] with a standard mixed effects model and other selected sensitivity analyses.

Resources for understanding and implementing mixed models and other statistical techniques are covered by Fairclough [10,21], Troxel [22], Fayers and Machin [23], Lenderking and Revicki [24], and Molenberghs and Verbake [25]. Also, guidelines regarding handling missing PRO data in statistical analyses are presented by Fayers and Hays [26], Fayers and Machin [23], Troxel [22], Fairclough [27], and Huntington and Dueck [14]. Revicki and Fairclough [28] cover methods for minimizing missing data in clinical trials.

Additional Ways to Handle Missing Data

In this section, we cover methods for handling missing data. Particularly, we cover the intention to treat (ITT) principle, the use of summary statistics (e.g., composite end points, indicator variables, and area under the curve [AUC] analysis), and imputation methods.

Intention to Treat Principle

Missing data may be handled analytically by the ITT principle or by the modified ITT principle in which at least one post-treatment measurement is required. Intention to treat reduces the amount of missing data or removes the issue entirely by labeling an individual as a success or failure once they have initiated treatment, regardless of whether they completed the trial.

In the following example, we translate the ultimate PRO outcome into a binary (dichotomous) outcome. In a trial of erythropoietin alpha [29] for improving hemoglobin [30], patients were scheduled to receive EPO for 16 weeks. Key end points were whether or not a patient achieved a specified clinically important increase in overall quality of life (QOL) from baseline as measured by a simple linear analog. All patients were evaluable for the PRO end point if they initiated treatment. If they failed to provide data for a given week, they were declared a treatment failure for that week. Unless an observed clinically meaningful increase in QOL was observed after baseline, the patient was declared a treatment failure. Complete data were obtained for this end point even though some patients had intermittent missing data.

Summary Statistics

Missing data can be handled by scientifically credible summary statistics. Summary statistics provide a method to combine available data into an overall end point. In anorexia trials, for instance, the maximum patient-reported improvement in appetite over a study period can be used as an end point to remove missing data issues. As another example, the American College of Rheumatology (ACR) response criteria constitute a composite end point developed as a

primary end point for clinical trials comparing rheumatoid arthritis treatments [31,32]. The response criteria combine clinical and PRO measures, including number of swollen and tender joints, patient-rated disease activity, clinician-rated disease activity, patient-rated pain, Health Assessment Questionnaire (HAQ), and c-reactive protein (CPR) values.

The scientific veracity of using summary statistics is well validated and must be identified a priori. It should also have evidence supporting reliability and validity in the target indication. The composite end point has to be validated by some reasonable scientific approach and not just concocted de novo. Going through a rigorous procedure including appropriate professional input and scientific scrutiny is essential to the development of any summary statistic. Summary statistics in general may be difficult to interpret because they typically lack a natural conceptualization of what is meaningful for what are essentially artificial summary indices.

An indicator variable is another type of summary statistic indicating whether a specific benchmark has been achieved (such as an a priori specified clinically meaningful benefit). Indicator variables produce complete data and binomial variables for which a standard statistical methodology such as Fisher's exact test and logistic regression readily apply. This is a preferred approach where applicable because it requires distinct a priori scientific justification for the handling of missing data and produces an appealing analytical approach. Unfortunately, the advantages of the binary approach come at the possible expense of loss of information.

Counterarguments to using this dichotomization approach of success and failure is that it loses the statistical power of the continuous end points. The key issue is whether the information gleaned from the PRO is sufficiently precise to be used as a continuous variable or whether the perceived level of accuracy is merely artificial numeritization. For example, the use of linear analog response scales data measured to the nearest hundredth of a millimeter is inappropriate because people are barely accurate to the nearest millimeter in responding to such items.

Area under the curve summary statistics are produced by combining longitudinal data into a simple, single numerical entity. At its simplest level the AUC represents the average value of the PRO over time for the entire treatment period. Some approaches use various distributional assumptions such as exponential decay between time points or the trapezoidal rule in estimating values between adjacent assessments. The justification for alternative assumptions needs to be based on scientific principles a priori rather than data-based post hoc determination.

Sloan [33] and Huntington and Dueck [14] provide examples of the technique in general. Sensitivity analysis is indicated if the distributional assumptions are questionable. Typically, the assumptions do not strongly affect the treatment comparison and so a small number of alternative assumptions can be tested before the consistency of the results becomes obvious. Missing data are handled in AUC construction in many ways, but typically by nearest-neighbor imputation or simply by constructing the AUC curve using the available data and prorating via the proportion of reporting periods.

Imputation Methods

In this section, we discuss single and multiple imputation methods. Single imputation is substituting a single value for each missing value in a data set. On the other hand, multiple imputations involve substituting several values for each missing value; in essence, creating multiple data sets. The statistic of interest is computed for each imputed data set and a final estimate is computed by combining estimates across the multiple data sets with readily available formulas in the multiple imputation literature.

There are many methods for selecting the values to impute whether using single or multiple imputation. Despite the myriad imputation techniques available, there is no uniformly accepted approach for imputation. All imputation methods are heavily reliant upon the underlying assumptions of the analytical technique. Results of any technique hence can be as much a result of the underlying assumptions as the true empiric result. Regardless of the technique, all methods involve statistically "guessing" what a person's PRO result would have been had it been observed. As such, imputation is "fudging the numbers" at its core. It is important to keep this truism in mind as alternative methods for imputation are considered to minimize the artificiality of the final results.

Sloan [33] provides examples of how different methods of handling missing data can provide startlingly different PRO average profiles over time. The most important issue is not the individual average profile, however, it is the relative difference between treatment regimens that is impacted by such missing data. More often than not (though definitely not in all cases), the missing data will impact both treatment average profiles similarly, keeping the comparative analytical results the same.

We recommend a differentiation between patients who do not provide PRO data because of death and those who do not provide data because of other reasons (e.g., missing pages in booklet, dropped study because of toxicity, etc.). Investigators should specify ahead of time how they will handle data for patients who die during the study; this permits them to examine this group of patients separately from those who were

potentially available to provide data at any given point but failed to do so. Imputation of zeroes (or the lowest possible score for a PRO) for those who have died, for example, is a technique in applying the intention to treat principle to patient PRO data but may result in significant bias in the presence of large amounts of missing form data [34]. Diehr and colleagues provide other approaches to incorporating mortality into PRO end points [1,35].

We do not endorse any particular method of imputation as being applicable to all situations; neither do we believe that any given method is prohibited in a specific circumstance. Each study should have an appropriate rationale that considers the expected pattern of missing data and imputation method chosen.

Single Imputation

The most common missing data approach in labeling claims is to use last-value carried forward in which the last observed value is substituted in for all subsequent missing data values. This method has been criticized for various reasons [24]. Alternative approaches involve imputing the average-value carried forward, the minimum or maximum value, or even imputing a zero value for patients who have died. This last approach has not seen much use but has the advantage of reflecting the average value for the complete sample that initiated the study. Some have argued that imputing a zero value for death may be inaccurate for a number of reasons [35]. The authors recommend that a single simple imputation method be specified for the primary analysis and justified within the context of the study.

Sensitivity analyses should include up to three different imputation approaches to verify that the results of the primary analysis are credible. If the results are not comparable, then this indicates the need for more complex missing data modeling and may suggest uncertainty regarding the PRO results.

Multiple Imputation

The use of complex statistical models to impute several values for missing data has become the topic of many statistical articles in recent years [21,22]. Whether linear and nonlinear models provide any better scientific "guesses" at missing numbers than common sense and single imputation has yet to be shown. The gains in terms of statistical power likely do not balance the amount of work and the number of strong distributional assumptions that are required. Nevertheless, because of the error inherent in "guessing," multiple imputations have the benefit over single imputation of a built-in measure of added variability in estimates computed from imputed data. (In multiple imputations, the estimation process considers the variability of estimates across the imputed data sets.) This benefit again likely does not balance the amount of work and required strong distributional assumptions because the added variability because of imputation can be assumed small if imputing for a small percentage of data and the simpler single imputation method is chosen appropriately.

In essence, if the multiple imputation methods provide a different answer than the simple single imputation methods, investigators may not know whether this is a result of improved precision or difference in technique. Recent literature has demonstrated that the result differs little across approaches [36].

Multiplicity

Analyzing multiple end points can be handled post hoc statistically or a priori scientifically. Although dealing with issues in the initial protocol is preferable, it is reasonable to expect that some issues are best addressed by statistical correction. We recommend, however, that investigators still prespecify the use of statistical correction in their statistical analysis plans. Many PROs, particularly those that assess health-related QOL or multiple facets of a disease are, by nature, multidimensional. Other clinical end points (such as tumor response) are also inherently multidimensional but have been worked into acceptable, agreed-upon single end points (e.g., a 50% reduction is a response). For example, trialists do not report average tumor size reduction or describe the characteristics associated with tumor response.

The best way to deal with multiplicity is to define in advance in the protocol or statistical analysis plan the PRO domains that the treatment is expected to affect and the domains that are not expected to be affected. Any given treatment is unlikely to affect any disease condition uniformly across all PRO subdomains. Additionally, in health-related QOL some dimensions may not change with treatment, such as concern about having a disease. These situations are analogous to the expectation that each treatment will produce variable toxicity profiles or disease-free survival, but not overall survival, may be affected. In all such situations, the end point of interest must be decided in advance and based on scientific evidence. We and others [37] recommend selecting a small number of PRO end points as being of primary interest and viewing remaining domain scores as secondary and supportive end points.

Multiple end points can be handled statistically by comparison-wise corrections to the Type I error rates. All these methods require an expanded sample size to account for multiple testing over and above what would be required for a single hypothesis test. The Bonferroni approach is most common; in it, the comparison-wise level of significance is set by dividing the number of tests involved into the overall experiment-wise Type I error rate. For example, if four PRO domains are to be tested, then an overall 5% Type I error rate is obtained if each test is carried out

at the 5%/4 = 1.25% significance level. This approach is considered quite conservative. Modifications of the Bonferroni approach involve specific algorithms of ordered testing, referred to as step-down, step-up, or hierarchical analyses [38,39].

Multiple End Points for Label Claims

Repeated or multiple testing of outcomes (or even of the same outcome at different times) will inevitably result, incorrectly, in a statistically significant treatment difference (say, *P*-value < 0.05) when no true difference exists between the treatments. Two forms of multiplicity exist. One form is multiple testing of multiple domains at the same time; a case in point is comparing multiple domains of the EORTC between a pair of treatments at week 12. Another form is multiple testing of a given domain at multiple times; comparing the emotional function domain of the EORTC between a pair of treatments at weeks 3, 6, and 12 illustrates this point.

Addressing the multiplicity of end points and time points is contingent upon study objectives and hypotheses, so no one definitive strategy for addressing multiplicity can be said to be appropriate for all studies. Nonetheless, general guidelines can be offered that are consistent with the specific hypotheses with regard to the domain(s) and the time point(s), both of which should be specified in advance. For a label claim, the number of prespecified domains of primary interest should be limited to no more than five and preferably to no more than three domains [23]. The number of key time points should also be limited and prespecified to testing treatment differences at no more than two time points for the primary analysis.

Suppose the objective is to show whether treatments differ in the mean change score from baseline to week 12 in five particular domains. Three types of alpha or *P*-value adjustments are recommended: 1) Bonferroni; 2) Bonferroni-Holm (Step-Down) Procedure; and 3) Hochberg's (Step-Up) Method. Of the three, Hochberg's Method is generally preferred. The following example illustrates testing for treatment differences in K = 5 domains where the five *P*-values are 0.20, 0.006, 0.011, 0.018, and 0.021.

Bonferroni:

If $P(i) > \alpha/K$, then accept the null; if $P(i) \leq \alpha/K$, then reject the null.

Ordered *P*-values:	$P(1) = 0.006$	$P(2) = 0.011$	$P(3) = 0.018$	$P(4) = 0.021$	$P(5) = 0.20$
α/K:	$\alpha/5 = 0.01$	$\alpha/5 = 0.01$	$\alpha/5 = 0.01$	$\alpha/5 = 0.01$	$\alpha/5 = 0.01$
Decision:	Reject	Accept	Accept	Accept	Accept

"Reject" means reject the null hypothesis that no treatment difference exists and therefore conclude that a treatment difference exists. "Accept" means do not reject the null hypothesis that no treatment difference exists and therefore conclude that no treatment difference exists.

Bonferroni-Holm (Step-Down) Procedure [38]:

- Step-Down: Start with smallest *P*-value.
- If $P(1) > \alpha/K$, then accept all null hypotheses (no treatment effects) and stop.
- If $P(1) \leq \alpha/K$, then the first null hypothesis [corresponding to $P(1)$] is rejected and then compare $P(2)$ with $\alpha/(K-1)$.
- If $P(2) > \alpha/(K-1)$, then accept all remaining null hypotheses and stop.
- If $P(2) \leq \alpha/(K-1)$, then the second null hypothesis [corresponding to $P(2)$] is rejected and then compare $P(3)$ with $\alpha/(K-2)$.
- Compare $P(3)$ with $\alpha/(K-2)$ and proceed in like fashion.

Ordered *P*-values:	$P(1) = 0.006$	$P(2) = 0.011$	$P(3) = 0.018$	$P(4) = 0.021$	$P(5) = 0.20$
$\alpha/K, \alpha/(K-1), \ldots =$	$\alpha/5 = 0.01$	$\alpha/4 = 0.0125$	$\alpha/3 = 0.0167$	$\alpha/2 = 0.025$	$\alpha/1 = 0.05$
Decision:	Reject →	Reject→	Accept→	Accept	Accept

Hochberg's (Step-Up) Method [39]

- Step-Up: Start with largest *P*-value.
- If $P(K) \leq \alpha$, then reject all null hypotheses and stop.
- If not, accept the first null hypothesis [corresponding to $P(K)$] and compare $P(K-1)$ with $\alpha/2$.
- If $P(K-1) \leq \alpha/2$, reject all remaining null hypotheses and stop.
- Otherwise, this second null hypothesis is accepted.
- Compare $P(K-2)$ with $\alpha/3$ in like fashion.

Ordered *P*-values:	$P(1) = 0.006$	$P(2) = 0.011$	$P(3) = 0.018$	$P(4) = 0.021$	$P(5) = 0.20$
$\alpha/K, \ldots, \alpha/2, \alpha =$	$\alpha/5 = 0.01$	$\alpha/4 = 0.0125$	$\alpha/3 = 0.0167$	$\alpha/2 = 0.025$	$\alpha/1 = 0.05$
Decision:	Reject	Reject	Reject	←Reject	←Accept

If interest centers on the difference in the mean change between treatments across time instead of at a specific time, then a summary measure of the domain scores can be created and a statistical difference between a pair of treatments tested using a multiple comparisons procedure like Hochberg's (Step-Up) Method.

Another way of handling multiple end points statistically is collective multiple testing, such as O'Brien's global test, to produce a single test of hypothesis. Similarly, one might use multivariate hypothesis testing such as Hotelling's T^2 or multivariate analysis of variance (MANOVA) [40–42]. We caution against using MANOVA, however, because it requires complete data which is rarely the situation in most pharmaceutical trials. Within the realm of MANOVA, however, a

hypothesis testing approach known as profile analysis may be applicable in testing PRO claims. A profile analysis is a hierarchical approach to multivariate comparisons between treatment groups. It typically proceeds in three steps: 1) a test for overall differences in average values; 2) a test for equality of levels; and 3) a test for differences over time. Collectively, the profile analysis represents a complete picture of the results. Again, the importance of a priori hypothesis specification is vital to the appropriate application of these statistical techniques.

Blinding

The absence of blinding (masking) of subjects to treatment group assignments is frequently raised as a potential source of bias for PROs, because subjects may believe that the newer treatment is somehow better and therefore may report improved health outcomes, even in situations in which they do not feel better. When possible, we recommend masking subjects to treatment and completely avoiding this possible source of bias in PROs. Sometimes, it is not possible to mask treatment assignments, but often in these situations two (or more) active treatments are evaluated in the clinical trial. In this case, it is important to make sure that the PRO assessments are performed before any clinical assessments or procedures are undertaken which might influence patient perceptions of their health state, and care must be taken to evaluate whether bias may be present. For the assessment of bias, it is important to examine whether the patient reports of symptom or health status improvement are tracking with objective clinical measures and clinician reports of change in clinical status. This bias in patient reporting may be most critical for short-term studies, as it may be difficult for subjects to continue to report improved symptoms and health-related QOL in the absence of any real effect over longer periods of time (i.e., the initial expectations for benefits may wear off with increased experience of no benefit).

Choosing the Correct PRO End Point

Much statistical literature has appeared about whether changes from baseline, average values at a particular time point, or percentage of successes should form the basis for analysis to compare treatment regimens. Although different statistical significance levels are possible for each of these three types of analyses, in most settings, analyses should result in consistent interpretations with respect to treatment efficacy. As an example, see the study by Rummans and colleagues of a psychosocial intervention which indicated statistical significance for all three end points [43]. Again, the a priori research hypothesis should be the primary source for the decision as to which end point is the most appropriate in a given situation.

As mentioned earlier, the intention to treat principle can be applied to identify a patient as a "success" or "failure" with respect to treatment outcome. This approach may be desirable if dichotomous outcomes are reasonable or where analysts expect to have many study dropouts (i.e., metastatic cancer). Otherwise, continuous end points such as average values per treatment (whether at a given time point or change from baseline) may be preferred to keep the sensitivity of a continuous end point. Change from baseline end points typically take precedence over average values at a given time point if researchers believe that treatment efficacy is related to baseline values. For an analysis of covariance model with baseline PRO as the covariate and treatment group as the key explanatory variable, the treatment effect and its standard error will be identical whether the dependent variable is the follow-up PRO or change in PRO from baseline.

Sensitivity Analyses

Sensitivity analyses are not always needed or required when conducting statistical analysis of PRO data. In some situations, properly conceived sensitivity analyses can help support and confirm the findings from the primary PRO data analysis. Most frequently, sensitivity analyses are recommended when the level of missing data is high (>20%), when a generally accepted method for imputing missing PRO data is lacking, or the best method for imputing missing PRO data is uncertain [44].

The planned sensitivity analyses should directly inform and address this uncertainty and the problems associated with missing PRO data. The sensitivity analyses can incorporate different approaches to imputing missing PRO data, such as substituting the worst possible (or observed) score for missing data, multiple imputation, and other methods [23]. Alternatively, different and somewhat more complicated statistical models, such as the family of pattern mixture models or selection models [11,24] can be used to compare the effects of treatment on PRO end points. It is best to complete a small, focused number of sensitivity analyses that are relevant and fit the particular situation and that help address any uncertainty related to the PRO analysis and findings. If these alternative imputation and statistical analysis strategies produce results that are similar to those of the primary PRO data analysis, the findings are further supported and confirmed. If the results of the sensitivity analyses are disparate with the primary PRO data analysis, then some question remains about the PRO results, and the investigators may need to provide further explanation for the PRO results.

Large clinical trials often include multiple clinical sites in potentially many different countries. Hence, using translated versions of the PRO of interest is common practice. Translations of the PRO measures

should be conducted according to standardized, accepted methods with cognitive testing (linguistic validation) of the translated measure in the countries where the measure will be used.

Analysts should test for interaction of treatment with study site, country, or region to provide statistical assurance that the translations of the PRO did not differ across sites by treatment group. "Revalidation" of the PRO in each country or region to examine its psychometric properties is not necessary. In a randomized clinical trial, any differences in the PRO from country to country will incorporate more noise in the measurement and, hence, decrease the ability to detect differences between treatments. There should be no reason that any decreased sensitivity of a PRO that has not been ideally translated would differ between randomized groups any more than a clinical measure would, and the test of interaction will provide this assurance.

How Should Null Results Be Interpreted?

Patient-reported outcome measures often include several different domains or summary scores. Positive treatment effects may be found in only a subset of these domain scores. The question then arises as to how to interpret so-called "null" results for scores that do not demonstrate statistically significant differences between treatment groups. The problem of how to interpret null results is not unique to PROs. It also arises with composite clinical end points where statistical power is often inadequate to show a statistically significant improvement in each individual component of a composite end point. Another example concerns different bacterial or viral subtypes in a clinical efficacy trial of a composite vaccine, where the end point will typically be infections caused by any of the types in the vaccine. The sample size is often insufficient to expect statistical significance for each subtype separately.

To approach the problem of interpreting null results, investigators should begin by appropriately prespecifying the hypotheses to be tested, the order in which they are to be tested, and any multiplicity adjustments. The choice of method and the ordering of the hypotheses to be tested are determined by the study objectives and by the power for different hypothesis tests. With the proper prespecified statistical hypothesis-testing plan, the interpretation of null results is greatly clarified. Hypothesis-testing plans are designed to preserve the overall Type I (alpha) error while providing adequate power for meaningful tests of a limited set of multiple hypotheses.

When data from a clinical trial are analyzed one should claim only what one can demonstrate with convincing supportive data. Null results are reported in the clinical study report and should also be addressed in journal publications. The question of whether they would need to be mentioned in drug labeling would depend on their importance for interpreting the positive results that have been shown. Null results that call into question the validity of positive results (e.g., those demonstrating an actual negative effect of the treatment) would need to be interpreted differently from secondary end points that simply failed to achieve statistical significance in the particular order of hypothesis-testing used in the trial. For example, clinical trials of new treatments for rheumatoid arthritis often find that new treatments demonstrate significant improvements in the SF-36 physical component but not, as expected, in the mental component. The labels for recently approved rheumatoid arthritis treatments (i.e., Remicaid, Enbrel, Humira) include statements on the treatment effects on both the physical summary and mental health summary measures.

In general, ensuring fair and complete reporting of PRO end points based on a clinical development program for a new medication is imperative. The focus should be primarily on prespecified PRO end points and those end points that reach statistical and clinical significance criteria. The sample sizes should allow for sufficient power to detect differences if these differences actually exist between the study treatments. In addition, investigators must provide sufficient evidence and rationale supporting the selection of PRO measures for the clinical trials, because measures with poor psychometric characteristics that do not adequately cover the relevant PRO domains are unlikely to detect treatment-related differences. Nevertheless, even with psychometrically sound measures, a priori specification of primary PRO end points, and well-designed clinical trials, unexpected patterns of findings may emerge. In these situations, investigators should report all the prespecified PRO end points, whether or not they support the treatment.

The problem of how to interpret null results is not unique to PRO measures. The approach to interpretation is greatly clarified by clear specification of the statistical hypothesis-testing and multiplicity-adjustment framework. Null results that call into question the validity of positive results need to be interpreted differently from those that simply represent hypotheses for which statistical significance was not achieved in the particular testing plan used in the trial.

A specific example of this, in a nonlabeling setting, involved a psychosocial intervention designed to impact overall patient health-related QOL [43]. Although other domains of QOL were included, the primary testing and analysis was carried out on overall QOL because it was the treatment target. The results indicated that overall QOL was indeed impacted by the intervention, although none of the other subdomains of QOL changed significantly.

In a similar fashion, if different aspects of fatigue were measured for a labeling claim, it would be the

sponsor's responsibility to define, a priori, which aspects of fatigue would be likely to be impacted. Ultimately, whether one end point is selected as primary, or multiple coprimaries are selected, it must be supportable by scientific argument a priori. Post hoc multiple testing should not be allowed under any circumstances to dredge the data for potential labeling claims.

Conclusions

Statistical analysis and interpretation related to results based on PROs can support a labeling claim with the same scientific integrity that is achievable for other end points, as long as the design elements necessary for credibility delineated in earlier manuscripts in this series are incorporated into the clinical trials. PRO data should be handled and viewed like any other effectiveness data in clinical trials.

The statistical analysis plan must be clear and consistent in justifying the various assumptions and processes used. It is critical to specify a priori what primary end point(s) will form the basis of the statistical analysis of the claim. Particular importance needs to be paid to the handling of missing data, the multiplicity of end points, and the longitudinal data structure. Methods for dealing with many of these analytical issues now exist and guidelines for their appropriate use are available. The FDA guidance document appropriately indicated that, methodological advances aside, there is need for further exploratory and confirmatory research in some areas. A body of evidence is accumulating, as articulated in this manuscript, that will continue to provide exemplary applications for statistical analysis and interpretation of PRO assessment in clinical trials.

Acknowledgments

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 Mayo Comprehensive Cancer Center grants (CA15083-32).

References

1 Diehr P, Patrick DL, McDonell MB, Fihn SD. Accounting for deaths in longitudinal studies using the SF-36: The performance of the physical component scale of the Short Form 36-Item Health Survey and the PCTD. Med Care 2003;41:1065–73.
2 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.
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 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.
7 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.
8 Osoba D, Bezjak A, Brundage M, Pater J. Evaluating health-related quality of life in cancer clinical trials: the National Cancer Institute of Canada Clinical Trials Group experience. Value Health 2007;10(Suppl. 2):S138–45.
9 Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health 2007; 10(Suppl. 2):S125–37.
10 Fairclough DL, Peterson HF, Cella D, Bonomi P. Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. Stat Med 1998;17:781–96.
11 Fairclough D. Design and Analysis of Quality of Life Studies in Clinical Trials. New York: Chapman & Hall/CRC, 2002.
12 Sloan JA, Dueck A. Issues for statisticians in conducting analyses and translating results for quality of life end points in clinical trials. J Biopharm Stat 2004;14:73–96.
13 Donaldson GW, Moinpour CM. Learning to live with missing quality-of-life data in advanced-stage disease trials. J Clin Oncol 2005;23:7380–4.
14 Huntington JL, Dueck A. Handling missing data. Curr Probl Cancer 2005;29:317–25.
15 Fairclough DL, Cella DF. Functional assessment of cancer therapy (FACT-G): Non-response to individual questions. Qual Life Res 1996;5:321–9.
16 Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. New York: John Wiley & Sons, 2004.
17 Hedecker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychol Meth 1997;2:64–78.
18 Singer JD, Willett JB. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York: Oxford University Press, 2003.
19 Brown H, Prescott R. Applied Mixed Models in Medicine. Chichester: John Wiley & Sons, 1999.
20 Pauler DK, McCoy S, Moinpour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. Stat Med 2003;22:795–809.

21 Fairclough DL. Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy. Stat Med 1997;16:1197–209.
22 Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical analysis of quality of life with missing data in cancer clinical trials. Stat Med 1998;17:653–66.
23 Fayers P, Machin D. Quality of Life Assessment, Analysis and Interpretation. New York: John Wiley, 2000.
24 Thijs H, Molenberghs G, Jansen I. Missing data: sensitivity analyses. In: Lenderking, WR, Revicki, DA, eds. Advancing Health Outcomes Research Methods and Clinical Applications. Mclean, VA: International Society for Quality of Life Research, 2005.
25 Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag, 2000.
26 Fayers P, Hays R. Assessing Quality of Life in Clinical Trials: Methods and Practice. New York: Oxford University Press, 2005.
27 Fairclough DL. Analysis and Interpretation of Results Based on Patient-Reported Outcomes. Value Health 2007;10(Suppl. 2):S106–15.
28 Revicki DA. Preventing missing data. In: Fayers P, Hays RD, eds. Assessing Quality of Life in Clinical Trials, 2nd edn. New York: Oxford University Press, 2005.
29 Fallowfield L, Gagnon D, Zagari M, et al. Multivariate regression analyses of data from a randomised, double-blind, placebo-controlled study confirm quality of life benefit of epoetin alfa in patients receiving non-platinum chemotherapy. Br J Cancer 2002;87:1341–53.
30 Witzig TE, Silberstein PT, Loprinzi CL, et al. Phase III, randomized, double-blind study of epoetin alfa compared with placebo in anemic patients receiving chemotherapy. J Clin Oncol 2005;23:2606–17.
31 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. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729–40.
32 Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology: Preliminary definition of improvement in rheumatoid arthritis. [see comment]. Arthritis Rheum 1995;38:727–35.
33 Sloan JA. Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. COPD. J Chron Obstr Pulm Dis 2005;2:57–62.
34 Revicki DA, Gold K, Buckman D, et al. Imputing physical health status scores missing owing to mortality: Results of a simulation comparing multiple techniques. Med Care 2001;39:61–71.
35 Diehr P, Patrick DL, Spertus J, et al. Transforming self-rated health and the SF-36 scales to include death and improve interpretability. Med Care 2001;39:670–80.
36 Enders CK. A primer on the use of modern missing-data methods in psychosomatic medicine research. Psychosom Med 2006;68:427–36.
37 Fayers P, Machin D, Fayers PM. Practical and reporting issues. In: Fayers P, Machin D, eds., Quality of Life Assessment. Analysis, and Interpretation. Chichester: John Wiley & Sons, 2000.
38 Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70.
39 Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800–2.
40 Sloan JA, Dueck A. Issues for statisticians in conducting analyses and translating results for quality of life end points in clinical trials. J Biopharm Stat 2004;14:73–96.
41 Mandrekar S, Kamath C. Presenting longitudinal data. Applying QOL assessments: solution for oncology clinical practice research, part 1. Curr Prob Cancer 2005;29:296–305.
42 Stevens J. Applied Multivariate Statistics for the Social Sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, 1996.
43 Rummans TA, Clark MM, Sloan JA, et al. Impacting quality of life for patients with advanced cancer with a structured multidisciplinary intervention: a randomized controlled trial. J Clin Oncol 2006;24:635–42.
44 Fairclough D. Analysing studies with missing data. In: Fayers, P, Hays, RD, eds. Assessing Quality of Life in Clinical Trials, 2nd edn. New York: Oxford University Press, 2005.