Methods to Evaluate and Correct for Bias in Patient-Reported Outcomes in Clinical Trials: A Discussion

Jessica Roydhouse[1,2], Joseph C. Cappelleri[3], Elizabeth Colantuoni[4], Demissie Alemayehu[3]

[1]Menzies Institute for Medical Research, University of Tasmania, 17 Liverpool Street, Hobart, TAS Australia 7000

[2]Department of Health Services, Policy, and Practice, Brown University School of Public Health, 121 S. Main Street, Providence RI USA 02912

[3]Pfizer Inc., 235 E 42nd St, New York, NY 10017

[4]Johns Hopkins University, 615 N. Wolfe Street, Baltimore MD USA 21205

Abstract

Improving the underlying disease or condition is a central goal of drug development. However, understanding patient experience while on therapy is increasingly of interest. The goal is accurate and interpretable patient-centric information that can inform providers and patients when making treatment decisions. Understanding the patient experience requires collecting data from patients. Patient-reported outcomes (PROs) such as symptoms and function are frequently collected on trials on a quantitative scale. These outcomes can provide valuable insight into the patient perspective.

However, like all trial data, PRO results may be biased. PRO data can present additional analytic challenges, and a better understanding of methods to analyze and interpret this data, while taking into account the potential for bias is needed. In this discussion paper, we consider two situations: 1) bias in responder analyses and 2) estimands for analyzing patient function in trials with severely ill patients.

Key Words: Trial, patient-reported outcome, bias, responder, estimand, principal stratification, composite

1. Background

The process of drug development and approval is becoming increasingly patient-centered. Major regulatory agencies such as the European Medicines Agency (EMA) and Food and Drug Administration (FDA) include patients in the regulatory process.[1] Furthermore, patient-reported outcomes (PROs) can be used to demonstrate treatment benefit.[2] PROs can be defined as patient reports of their own health or symptoms, rather than a report or interpretation about a patient's health or symptoms from other individuals such as clinicians or caregivers. PROs included as part of efficacy submissions for regulatory consideration to the FDA were frequently discussed in FDA clinical reviews.[3] Depending on the clinical context, PROs may be primary or non-primary endpoints. Although PROs can contribute important data about a patient's experience while on therapy, and the benefit of the treatment to a patient, PRO data also poses analytic challenges. The FDA's 2009 PRO guidance highlighted several of these challenges, including, in particular, missing data.[2] Missing data in the guidance was described as including trial attrition and non-

completion of PRO instruments while on trial. The National Research Council report discussed the challenge of missing data in trials, and also noted that some unavailable data such as PROs after a patient has died would not be defined as missing.[4] The delineation between missing data and post-randomization intercurrent events (ICEs) such as treatment discontinuation or death that impact a clinical question of interest was further explicated in the International Council for Harmonisation (ICH) E9 (R1) estimand framework.[5] Terminal ICEs such as death are particularly salient in clinical contexts such as critical care or oncology. Evaluating the potential for bias in PRO data and mitigating it where possible is important.

In this article, we will consider challenging methodological issues for PROs in different trial contexts and discuss methods for analyzing PRO data in these contexts, as well as evaluating and correcting for bias. We will also consider contexts with high anticipated mortality (critical illness) and others where it is not (sexual medicine).

This article is organized as follows. We will begin by discussing the challenge of misclassification when operationalizing PROs as dichotomous variables and consider a method for correcting misclassification for a dichotomous PRO (Section 2). This will be in the context of sexual medicine, where mortality is not anticipated. Next, we will discuss different estimands for continuous or ordinal PRO data in critical illness, where high mortality can be anticipated but patient reports of function are a co-primary or key secondary outcome (Section 3). Finally, we will conclude in Section 4 with a brief discussion of the issues raised and considerations for the way forward.

2. Addressing Bias in Responder Analyses of Patient-Reported Outcomes

2.1 Introduction

One way to lend meaning and interpretation to a quantitative PRO measure is to dichotomize between values where within-patient changes are considered clinically important and those that are not.[6,7] Responder analysis is in common use in clinical trials and has been described in regulatory documents,[2,8] especially where "soft" clinical endpoints such as PRO measures are used. The procedure is useful because a between-group difference in responder proportions or percentages may be understood more intuitively than a between-group difference in mean scores from rating scales.

Anchor-based methods, which examine the association between the targeted concept of the PRO measure and the concept measured by the anchor measure (i.e., an interpretable external measure, which serves as a "gold standard" criterion, related to the PRO measure), can provide the primary empiric evidence to estimate a cutoff or threshold score for the responder definition of the targeted PRO measure.[2,6,7,9] Nonetheless, even a PRO scale with a cutoff score that discriminates well between responders and non-responders is fraught with some misclassification or measurement error: Some individuals classified as responders (based on the cutoff or threshold score on the PRO measure) may in fact be non-responders; some individuals classified as non-responder may in fact be responders. Yet there has been no attempt in research to adjust for responder misclassification on a PRO measure.

2.2 General Methodology

In standard epidemiologic settings, formulas exist for correcting misclassification on disease or exposure, or both, for a two-way cross-classification table of disease status (yes, no) and exposure status (yes, no).[10,11] But these formulas have not been applied in the context of responder analysis in general and for PRO measures in particular. In this current research, no misclassification of treatment is assumed, a reasonable assumption in experimental and quasi-experimental studies where the investigator directs treatment allocation (be it randomly or non-randomly). The formulas with misclassification on disease only (and no misclassification on treatment) can be applied directly and modified by replacing disease (yes, no) with responder status (yes, no).

In the corresponding oral presentation version of this article, formulas are provided to correct for responder misclassification under the assumption of no treatment misclassification in a two-by-two contingency table. A generalizable framework is provided to illustrate how responder misclassification affects measures of treatment effect (responder ratio, responder difference, odds ratio). In the oral presentation, estimates of treatment effect are compared between unadjusted and adjusted estimates of treatment effect using two cases studies from sexual medicine to illustrate the methodology.

2.3 Methodological Considerations

It should be emphasized that the main analysis of patient-reported measures with quantitative (ordinal or continuous) data should be analyzed as such, rather than a dichotomized version of them, in order to preserve the full information and natural structure inherent in the original data.[8,12] A responder analysis is intended to supplement, not replace, such a main analysis for the purpose of advancing interpretation of a quantitative PRO measure above and beyond its primary analysis and interpretation from original data using a type of regression model.[13,14]

A limitation of the formulas intended to correct for nondifferential misclassification of binary responder status may yield negative and hence inappropriate results for the corrected cell frequencies in certain circumstances. Although not a perfect solution, one viable way to address this problem is to select the closest alternative value to sensitivity or specificity that changes a cell count from negative to positive.

Anchor-based methodology is used to determine sensitivity and specificity.[2,6,9,15-19] The two examples given in the oral presentation use, in particular, a receiver operating characteristic (ROC) curve analysis to obtain sensitivity and specificity.

The simple bias-correction analysis introduced here for responder analysis of PRO measures is an improvement over its conventional counterpart, which implicitly assumes no misclassification error at all on responder status (100% sensitivity and 100% specificity). But this simple bias-correction implies that the diagnostic parameters (i.e., sensitivity and specificity) are fixed and known without error, a situation that is rarely realized. This limitation is not restricted to PRO measures but applies generally to many exposure and outcome variables in epidemiology.[11]

2.4 Conclusion

In the context of PRO measures, formulas are available that correct for responder misclassification under the assumption of no treatment misclassification and, therefore, corresponding estimates of the treatment effect can be adjusted accordingly. As such, treatment effect bias from misclassification of responder status on PRO measures is

addressed and corrected, leading to their having a more trustworthy interpretation for effective decision-making.

3. Estimands for PRO Data in a Clinical Setting of High Mortality

3.1 Introduction

In trials evaluating treatments for critically ill patients, mortality is common and is often the primary outcome. However, given the high value that patients place on outcomes other than mortality, including cognition, physical function, and quality of life, these and other PROs are increasingly being studied in critically ill populations as either co-primary or key secondary outcomes.[20,21] Comparing treatment effects on PROs is complicated when a subset of patients die before the PROs can be assessed. Patient death defines an intercurrent event[5] and the PROs for patients who die are "truncated" due to death and do not exist.[22,23] As outlined in the ICH E9 (R1) estimand framework and described below, there are several approaches for defining the treatment effect for PROs "truncated due to death"; however, there is no single best approach. Regardless of which approach is applied; clearly defining the estimand and justifying and supporting the required assumptions is essential.

3.2 General Methodology

The approach most commonly applied to PROs "truncated due to death" is the "survivors only" analysis where the mean PRO among survivors from each treatment arm are compared. Treatment comparisons using PRO data available only from survivors violates the intention-to-treat (ITT) principle and can yield biased results because mortality is a post-randomization event determining which patients provide PRO data for comparison.[22,24] If treatment assignment has no impact on survival, then the benefits of randomization are preserved. Specifically, the cohort of patients who are observed to survive in the two randomized treatment groups will probabilistically have the same distribution of measured and unmeasured baseline covariates.

Causal methods, known as principal stratification,[25] seek to address the bias created when mortality differs by treatment arm. The principal stratification approach is based on the potential outcomes framework of Rubin.[26] This approach conceptualizes, for each patient, the survival time and PRO at a specified follow-up assessment (if the patient survives to the assessment time) for each possible treatment assignment. The difference between the PRO at a specified follow-up assessment under the two possible treatment assignments is only defined for "potential survivors," i.e. those patients who would survive to the assessment time on both treatments. The survivor average casual effect (SACE)[23] is defined as the mean difference in the PRO among the "potential survivors." Like the "survivors only" analysis, the SACE defines a treatment effect for only a subset of patients thus violating the ITT principle. Further, the "potential survivors" are a subset of study patients; however, since patients are only assigned to one treatment arm during a trial, assumptions are required to identify the proportion of "potential survivors" and estimate SACE.[27-30]

An alternative to the two conditional approaches described above is to generate a composite outcome that combines mortality and the PRO among survivors into a single outcome allowing for treatment comparisons based on all randomized patients, i.e. satisfying the ITT principle. In general, the composite outcome approach requires a natural ordering or ranking of the value of death and the PRO with the composite outcome comparison based on a non-parametric test such as the Mann-Whitney test. Several PRO measurement scales include mortality, e.g. the EQ-5D utility score where death has a numeric score of 0 and

negative values to represent health states worse than death; whereas for others, death may be ranked as the worst health state or the timing of death may be incorporated into the treatment arm comparisons.[31-33]

3.3 Methodologic Considerations

Comparing PROs across treatment arms when patient mortality is anticipated is challenging and we have briefly reviewed three approaches to address this challenge. When mortality is the primary outcome of a trial, given the anticipated difference across treatment arms, pre-planned analyses of PROs should consider approaches beyond the "survivors only" analysis. In addition, regardless of which approach is selected, the statistical methods section should include a precise definition of the patients used in treatment comparisons of PROs, as well as definitions of and support for any required assumptions. Further, the impact of possible violations to the required assumptions should be addressed in the discussion/limitations.

3.4 Conclusion

There is not one single best approach for comparing PROs across treatment arms in trials conducted among critically ill patients where patient mortality is anticipated. Improvements in reporting statistical methods, required assumptions and the subsequent impact of violations of the assumptions can lead to better interpretation of trial results.

4. Conclusion and Discussion

The current paradigm of drug development is heavily reliant on efficacy and safety data generated with minimal or no regard to the experience of the patients. However, there is a growing realization of the need to incorporate the patients' perspective to better inform patients, healthcare providers and other stakeholders about the relative risks and benefits of alternative treatment options. This recognition is reflected in the various guidelines issued by regulatory agencies and HTA authorities. In addition, there is a huge drive by patient advocacy groups to ensure that the role of the patients is established as an integral component of a drug development program.

To understand the patient experience adequately, it is essential to collect PROs using well-validated instruments and to have a well-planned strategy that considers the issues that may arise with regard to the analysis of PRO data. Although paper-based data collection has been the customary approach, with the increasing use of technology to enhance the efficiency of clinical trials, ePROs are now routinely implemented in data acquisition. Irrespective of the mode of data collection, the analysis and interpretation of PRO data require consideration. Data quality, validity and reliability are especially germane in research involving PROs. While use of technology can mitigate some of issues, technology can also introduce additional challenges, including measurement validity and interpretability, especially in areas where information about the PRO measure has been established using paper-based instruments. Beyond this, however, there are a number of issues that should be reviewed before proposing a PRO endpoint and the statistical analysis for that endpoint.

As discussed in this paper, bias can be introduced in several ways in the reporting and collection of PRO data. Improper definition of a responder with respect to a PRO endpoint can lead to inadequate characterization of the risk and benefit profile of a given treatment. The proposed approach, while not without limitations, can help mitigate responder

misclassification under suitable assumptions. Since responder analyses are frequently used for PRO endpoints, consideration of possible misclassification is important.

In studies involving mortality, in which PRO endpoints are likely to be truncated, caution is advised in the choice of the analytical method. Consideration and adequate reporting of the patient population to which the endpoint corresponds, as well as underlying assumptions, is also relevant. The estimand framework from the ICH E9 (R1), which considers population to be an estimand attribute, can be helpful in this regard. Pre-trial planning for how PRO data should be collected and analyzed given intercurrent events such as mortality, as well as how to account for PRO data that are missing, is essential. The approaches of "survivors only," principal stratification and composite have advantages and disadvantages, and including a sensitivity analysis is critical.

A topic that has not specifically been addressed in this paper is the handling of potential bias associated with PRO outcomes in open-label studies, including single-arm trials. In situations where double-blind studies are not feasible for ethical or operational reasons, estimation of treatment effects may be challenging without accounting for the true "placebo effect," emanating from the patients' knowledge of the treatment they are receiving. Caution should, therefore, be exercised in the interpretation of PRO results obtained from such trials.

In summary, as the focus on personalized medicine heightens, PRO evidence is likely to continue to play a crucial role in clinical and policy decision-making, as well as in labelling claims for medical products. While the issue of missing data has garnered considerable attention in the PRO literature, it is noted that other sources of bias are equally important, and require concerted efforts by all researchers involved in the design, conduct, analysis and reporting of PRO data to appreciate the problems. Furthermore, with the advent of the estimand framework, differentiating between missing data and intercurrent events is important. It is hoped that the ideas put forth in this paper can serve as a useful resource for PRO experts, sponsors and other stakeholders in their efforts to fully integrate the patient's perspective in healthcare decision making.

References

1. Hoos A, Anderson J, Boutin M, et al. Partnering With Patients in the Development and Lifecycle of Medicines: A Call for Action. *Ther Innov Regul Sci.* 2015;49(6):929-939.
2. Food and Drug Administration. *Guidance For Industry: Patient-Reported Outcomes Measures: Use In Medical Product Development to Support Labeling Claims*. 2009. Silver Spring, Maryland: US Department of Health and Human Services.
3. Kim J, Kanapuru B, Roydhouse JK, et al. 2017-2018 Hematology Drug Approvals at the Food and Drug Administration (FDA): Communication of Patient-Reported Outcomes (PRO) Information in FDA Clinical Reviews and Prescribing Information (PI). *Blood.* 2019;134:3450.
4. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med.* 2012;367(14):1355-1360.
5. International Council for Harmonisation. *E9(R1) Estimands and Sensitivity Analysis in Clinical Trials.* 2019.
6. Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. *Stat Methods Med Res.* 2014;23(5):460-483.

7. McLeod LD, Coon CD, Martin SA, Fehnel SE, Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Rev Pharmacoecon Outcomes Res.* 2011;11(2):163-169.
8. European Medicines Agency. *Guideline on Multiplicity Issues in Clinical Trials.* 2017. London, United Kingdom: European Medicines Agency.
9. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol.* 2008;61(2):102-109.
10. Kleinbaum DG, Sullivan KM, Barker NB. *ActivEpi Companion Textbook: A Supplement for Use with the ActivEpi CD-ROM*. Second ed. New York, NY: Springer; 2013.
11. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analyses to Epidemiologic Data.* New York, NY: Springer; 2009.
12. Food and Drug Administration. *Patient-Focused Drug Development Guidance Series for Enhancing the Incorporation of the Patient's Voice in Medical Product Development and Regulatory Decision Making. Draft guidance documents.* https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical. Published 2019. Accessed April 15, 2020.
13. Snapinn SM, Jiang Q. Responder analyses and the assessment of a clinically relevant treatment effect. *Trials.* 2007;8:31.
14. Uryniak T, Chan ISF, Fedorov VV, et al. Responder Analyses – A PhRMA Position Paper. *Stat Biopharm Res.* 2011;3:476-487.
15. Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Qual Life Res.* 2018;27(1):33-40.
16. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol.* 2003;56(5):395-407.
17. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide.* New York, NY: Cambridge University Press; 2011.
18. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res.* 2011;11(2):171-184.
19. King MT, Dueck AC, Revicki DA. Can Methods Developed for Interpreting Group-level Patient-reported Outcome Data be Applied to Individual Patient Management? *Med Care.* 2019;57 Suppl 5 Suppl 1:S38-S45.
20. Fried TR, Bradley EH, Towle VR, Allore H. Understanding the treatment preferences of seriously ill patients. *N Engl J Med.* 2002;346(14):1061-1066.
21. Turnbull AE, Rabiee A, Davis WE, et al. Outcome Measurement in ICU Survivorship Research From 1970 to 2013: A Scoping Review of 425 Publications. *Crit Care Med.* 2016;44(7):1267-1277.
22. Rubin DB. Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Stat Sci.* 2006;21(3):299-309.
23. Rubin DB. Discussion of "Causal Inference with Counterfactuals" by A.P. Dawid. *J Am Stat Assoc.* 2000;95:435-437.
24. Zhang JL, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J Educ Behav Stat.* 2003;28(4):353-368.
25. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics.* 2002;58(1):21-29.

26. Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *J Educ Psychol.* 1974;66(5):688-701.
27. Hayden D, Pauler DK, Schoenfeld D. An estimator for treatment comparisons among survivors in randomized trials. *Biometrics.* 2005;61(1):305-310.
28. Egleston BL, Scharfstein DO, Freeman EE, West SK. Causal inference for non-mortality outcomes in the presence of death. *Biostatistics.* 2007;8(3):526-545.
29. Chiba Y, VanderWeele TJ. A simple method for principal strata effects when the outcome has been truncated due to death. *Am J Epidemiol.* 2011;173(7):745-751.
30. Tchetgen Tchetgen EJ. Identification and estimation of survivor average causal effects. *Stat Med.* 2014;33(21):3601-3628.
31. Diehr P, Patrick DL, Spertus J, Kiefe CI, McDonell M, Fihn SD. Transforming self-rated health and the SF-36 scales to include death and improve interpretability. *Med Care.* 2001;39(7):670-680.
32. Lachin JM. Worst-rank score analysis with informatively missing observations in clinical trials. *Control Clin Trials.* 1999;20(5):408-422.
33. Wang C, Scharfstein DO, Colantuoni E, Girard TD, Yan Y. Inference in randomized trials with death and missingness. *Biometrics.* 2017;73(2):431-440.