VALUE IN HEALTH 20 (2017) 15–17

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

EDITORIAL

Reflections on ISPOR's Clinician-Reported Outcomes Good Measurement Practice Recommendations

Characterizing treatment benefit in terms that are meaningful and operationally sound is not only fundamental to clinical science but also essential to the credibility and clarity of the communication of this vital information. In this second report by the International Society for Pharmacoeconomics and Outcomes Research Task Force (TF) for Clinical Outcome Assessments (COAs), a clear conceptual foundation is provided for the development and evaluation of three types of clinician-reported outcome (ClinRO) assessments: reading, rating, and clinician global assessments [1].

Moreover, two additional terms of importance to the COA scientific community are discussed in the context of ClinROs: end points and biomarkers. The report emphasizes an important and often trivialized distinction between a COA and an end point. End points define how a COA is used to evaluate treatment benefit, analyzed to determine differences between groups, and interpreted to convey how any group differences reflect benefits in how patients feel, function, or survive. Furthermore, biomarkers are discussed as offering potentially useful information to clinicians when making a ClinRO assessment. When biomarkers are the sole determinant factor of a clinical decision, however, they are not ClinROs.

Good measurement practice (GMP) is elucidated via eight key areas: 1) defining the context of use; 2) identifying the concept of interest measured; 3) defining the intended treatment benefit in how patients feel, function, or survive, as reflected by the ClinRO assessment, and evaluating the relationship between the intended treatment benefit and the concept of interest; 4) documenting content validity; 5) evaluating other measurement properties once content validity is established (including intra- and interrater reliability); 6) defining study objectives and end point(s) objectives, defining study end points, and placing study end points within the hierarchy of end points; 7) establishing interpretability in trial results; and 8) evaluating operational considerations for the implementation of ClinRO assessments used as end points in clinical trials.

In their report and recommendations, the ClinRO TF authors have thoughtfully detailed many key considerations for good measurement practices, including clinical case examples to illustrate their points. Although the important conceptual and qualitative matters emphasized in defining the context of use (GMP 1) and in identifying the concept of interest (GMP 2) are indeed helpful in setting a solid foundation, the last six of these GMPs could benefit from moving beyond conceptual and clinical-qualitative case examples to also include some fundamental quantitative perspective with specific technical considerations matched directly to each recommended measurement practice. This, we believe, would strengthen the value and applied practical relevance of this important TF report.

For defining the intended treatment benefit reflected by the ClinRO (GMP 3), the report points out that most ClinRO assessments are indirect assessments of treatment benefit, and, therefore, an understanding is needed about the relationship between the ClinRO assessment's measured concept of interest and the meaningful health aspect of how patients feel, function, or survive. The relationship here should be based on the consistency of ClinRO and patient-reported outcome (PRO) scores, not their absolute agreement, because previous research in various diseases has highlighted discrepancies between patient and physician ratings of disease severity, with physicians tending to underestimate or underreport symptoms compared with patient reports [2].

Correlational and regression analyses of the ClinRO measure and a relevant PRO measure, which directly assesses treatment benefit, can provide the needed linkage. Although such correlation analyses would be informative, the magnitude of correlation at baseline is not expected to be sizable (and can be quite small) due to the typically restricted range of scores on both measures when a homogeneous sample of patients is assessed before treatment intervention. For this reason, postbaseline (or change from baseline) correlations would be preferred.

To capitalize on all available data from longitudinal measurements (baseline and all postbaseline assessments) on ClinRO and PRO assessments, their relationship can be evaluated using a repeated measures model with a suitable PRO measure as the outcome (or dependent variable) and the ClinRO as the predictor (regressor), with data pooled across all treatments. For example, it would be useful to know that a certain level or score on a ClinRO measure corresponds (on average) to a certain level or score on the PRO measure, with descriptors belonging to the response categories of both measures being used to enrich interpretation.

Furthermore, mediation modeling can be used to identify and explain the observed relationship between treatment group as the predictor and ClinRO of interest as the outcome at a given time point when a relevant PRO measure (which directly assesses how patients feel or function) is considered as the mediator variable. From this set of relations, pertinent questions can be addressed regarding what proportion of the total effect of treatment benefit on the ClinRO measure constitutes an indirect effect mediated through the PRO measures (the mediator) and what remaining proportion constitutes a direct effect (which represents all other possible effects other than those attributed to the mediator) [3].

For documenting content validity (GMP 4), the TF report notes that multifaceted items on a ClinRO can pose particular challenges, such as lack of clarity or justification on combining its different components. Confirmatory factor analysis can help in this regard by providing evidence regarding whether the ClinRO measurement model fits the data and, for example, whether it is appropriate to equally weight the items belonging to the ClinRO measure [4]. The point here is that content validity of the ClinRO assessment can be supported with suitable quantitative methods such as confirmatory factor analysis.

Regarding the evaluation of intra- and interrater reliability (GMP 5), the report recognizes that standardized and uniform training of clinician-raters is paramount at the clinical design stage to maximize agreement between different raters on the same set of patients at a given time (interrater reliability) and within a given rater for the same set of stable patients across time (intrarater reliability). At the psychometric analysis stage, the report appropriately recommends intraclass correlation and limits of agreement, which should be considered for continuous variables (or variables treated as continuous). It should be borne in mind, however, a family of intraclass correlations (not just one such correlation) exists, and which one is correct to use depends on the situation; using the incorrect intraclass correlation could be misleading [5]. Two major considerations for selecting the right intraclass correlation are 1) whether interest centers on only the raters in the study or a random sample of all possible raters (where the study raters are considered representative); and 2) whether interest centers on absolute agreement in rater scores or merely on their consistency (ranking). For ClinROs in this context, attention usually centers on absolute agreement (which is typically lower than consistency) among all possible raters. Like intraclass correlations, kappa statistics (which reflect chance-corrected agreement) for discrete or categorical variables have their own versions (simple and weighted) [6].

The application of ClinRO measures invites the more frequent use of generalizability theory in evaluation of their measurement properties—specifically on the accuracy and reproducibility of ClinRO assessments—thereby providing evidence for having (or not having) confidence in a ClinRO measure. Especially useful for quantitative ratings from ClinROs, generalizability theory concentrates on the dependability of measurements and enables multiple sources of error in a measurement to be estimated in a single analysis [7,8].

Suppose, for example, that a noninterventional methodologic study is to be undertaken to assess and quantify the reliability of a ClinRO measure before it is considered in a clinical trial. The study is to have multiple clinical raters scoring the same set of patients at multiple time points. In this repeated measurement of rating scores over time by different clinicians, a generalizability study can address several facets of the reliability of ClinRO measurement. This involves a comparison of the measurements of all patients as performed by different clinicians, across clinicians but not across time (i.e., interrater reliability); a comparison of one measurement by one clinician with another measurement by another clinician, across clinicians and time; a comparison of measurements performed over time by the same clinician, across time measurements but not across clinicians (i.e., intrarater reliability); and whether higher reliability is obtained by using the average of multiple measurements of a patient by one clinician or by using the average of one measurement by different clinicians, which can be used to make a decision in the planning of a subsequent (decision) study.

Regarding study objectives and end points within the hierarchy of end points (GMP 6), we concur with the guidance to be judicious regarding the role and position of the targeted ClinRO measure in establishing a hierarchy of all the end points chosen. The inferential testing of ClinRO for treatment effect can be based on the previous recommendations for PRO measures regarding adjustment for the multiplicity of end points intended for a label claim [9].

Among the most useful and efficient approaches include the gate-keeping method, which prespecifies a sequence or order of testing (a hierarchy) of comparisons that should first be satisfied before others are considered for testing. If the order of importance for a set of measures, which may include one or multiple ClinRO measures, is uncertain, then alternative approaches like the step-up method can be applied after the previous set of end points meet the gate-keeping criteria.

The authors also make a good point about a composite outcome, which may include at least one ClinRO measure. The advantages of such a multicomponent end point are that it may increase statistical precision of a treatment effect if this end point is correctly identified, may help to overcome the dilemma of which end point should be primary, and may deal with the multiplicity issue in an efficient manner. Disadvantages may arise, however, when the composite is not correctly identified or when its constituent components do not align with each other. A fundamental issue with a composite outcome is that its individual components must be associated with the primary objective. When this relationship is suspect, the interpretation of the composite outcome is also suspect. Components should be measurable concepts that can sensibly be added together as being aspects of the same underlying construct of treatment benefit, and interpretation is easier if the composite can fall under the rubric of a single name [10].

For establishing interpretability in trial results that have used ClinRO endpoints (GMP 7), the report gives this due emphasis. The same set of approaches for interpretation of PRO measures is also applicable for the interpretation of ClinRO measures as ratings [3]. For this purpose, two broad approaches—anchor based and distribution based—are available to enrich the understanding and meaning of ClinRO measures.

Anchor-based approaches include percentages based on thresholds, criterion-group interpretation, content-based interpretation, and clinical important difference. In addition to being a potentially useful secondary outcome, a clinical global assessment (like a patient global assessment) can serve as a useful anchor if it can be clearly interpreted and bears an appreciable correlation with the targeted ClinRO. Distribution-based approaches include effect size, probability of relative benefit, and responder analysis and cumulative proportions.

For a clinically important difference, an approach that uses all available longitudinal data involves a repeated measures model with the ClinRO of interest as the outcome (dependent) variable and a suitable PRO or another separate ClinRO measure as the anchor predictor. A clinically important difference in the ClinRO can be defined as the difference that corresponds to a one-category change on the anchor [3]. Mediation modeling can also help with interpretation by providing a framework to simultaneously assess the interrelationships of different variables (e.g., ClinRO, PRO, treatments).

The authors provide several insights regarding operational considerations of ClinROs (GMP 8). Nonetheless, one consideration not mentioned is the migration of ClinROs from paper to electronic administration. As with ePROs, the extent of validation for eClinRO depends on the level of modification (minor, moderate, or substantial) [11]. With eClinROs, however, two additional categories of classification should be considered—functional adaptation and instructional adaptation—because the clinician is likely to have a prior knowledge base associated with the disease indication, scale being used, and electronic devices [12].

A functional adaptation represents a change that allows the item to be administered in an electronic format—for example, use of radio buttons rather than circling a response or the addition of comment boxes to capture information. An adaptation of instructions refers to the addition of instructions from the training or administration guidelines, instructions not previously included in the paper scale, to increase the likelihood that scale administrators are following appropriate scale conventions (e.g., the addition of previously established guidelines, such as an instruction from the scale manual informing the clinician to read the question verbatim). Having more specific and better standardized instructions may help to ensure greater interrater reliability. The addition of such standardization to an electronic scale can therefore be considered a positive enhancement.

In conclusion, we applaud this International Society for Pharmacoeconomics and Outcomes Research TF work in producing this report that elucidates, delineates, and illustrates key principles of emerging GMPs for the development and assessment of ClinROs. In doing so, minimization of measurement error on the concept of interest and lucid interpretation of study results become more of a reality. These measurement practices will provide a solid foundation for future development and guidance on ClinRO assessments in clinical trials as the recommendations given evolve with more research and experience.

Joseph C. Cappelleri, PhD, Linda S. Deal, MS,
Charles D. Petrie, PhD*
Pfizer Inc
**Address correspondence to: Charles D. Petrie, Pfizer Inc, MS 8260-2511445 Eastern Point Road, Groton, CT, USA, 06340.*
E-mail address: charles.d.petrie@pfizer.com

http://dx.doi.org/10.1016/j.jval.2016.12.007

REFERENCES

[1] Powers JH III, Patrick DL, Walton MK, et al. Clinician reported outcome (ClinRO) assessments of treatment benefit: Report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force. Value Health. In press.

[2] Paul C, Bushmakin AG, Cappelleri JC, et al. Do patients and physicians agree in their assessment of the severity of psoriasis? Insights from tofacitinib phase 3 clinical trials. J Dermatolog Clin Res 2015;3:1048.

[3] Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. Stat Methods Med Res 2014;23:460–83.

[4] Cappelleri JC, Bushmakin AG, Harness J, et al. Psychometric validation of the physician global assessment scale for assessing severity of psoriasis disease activity. Qual Life Res 2013;22:2489–99.

[5] McGraw KO, Song SP. Forming inferences about some intraclass correlation coefficients. Psychol Meth 1996;1:30–46.

[6] Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions (3rd ed.). Hoboken, NJ: John Wiley & Sons, 2003.

[7] de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide. New York, NY: Cambridge University Press, 2011.

[8] Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use (5th ed.). New York, NY: Oxford University Press, 2015.

[9] Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. Fed Regist 2009;74:65132–3.

[10] Freemantle N, Calvert M, Wood J. Composite outcomes in randomized trials: Greater precision but with greater uncertainty? JAMA 2003;289:2554–9.

[11] Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force Report. Value Health 2009;12:419–29.

[12] Fuller RLM, McNamara CW, Lenderking WR, et al. Establishing equivalence of electronic clinician-reported outcome measures. Ther Innov Regul Sci 2016;50:30–6.