Assessing Clinician Agreement in Clinician Reported Outcomes: Many Options — Choose Carefully!

Randall H. Bender, PhD
Senior Psychometric Statistician, Outcomes Research, Evidera

Researchers familiar with patient-reported outcomes (PROs) may find that the transition to working with clinician-reported outcomes (ClinROs) is not as seamless as expected. Though both types of outcomes are required to meet the same criteria to be submitted as evidence to support a label claim ("well-defined and reliable"), the details involved in validating a ClinRO for use can be a bit more entangled than those involved with a PRO. One piece of the validation process that can often prove challenging is the assessment of inter-rater (inter-clinician) reliability or agreement.

Clinical trial researchers are looking for measures with as little measurement error as possible, hence a great deal of emphasis should be placed by measure developers on understanding all the potential sources of error in order to minimize or eliminate them. With ClinROs there is a new source of error in the measurement process: the "clinician." The FDA glossary describes Clinical Outcome Assessments (COAs), which include ClinROs, as "any assessment that **may be influenced by human choices, judgment, or motivation and may support either direct or indirect evidence of treatment benefit. … COAs depend on the implementation, interpretation, and reporting** [emphasis added] from a patient, a clinician, or an observer."[1] The definition of ClinRO in the same glossary expands a bit on this with reference to the special aspects of potential human error brought in by clinicians:

> **"Clinician-reported outcome (ClinRO)** — A ClinRO is based on a report that comes from a trained health-care professional after observation of a patient's health condition. A ClinRO measure involves a **clinical judgment or interpretation of the observable signs, behaviors, or other physical manifestations thought to be related to a disease or condition** [emphasis added]." [1]

All of this is to catalog the many potential sources of human error found in all COAs, PROs, and ClinROs alike (Table 1). Given the inclusion of the clinician in the measurement process, ClinROs include some unique sources of potential error which developers have the opportunity to evaluate in examining inter-rater/-clinician agreement. Determining how much a ClinRO measurement could vary, simply as a result of the clinician is who is using the ClinRO, is integral to the ClinRO's measurement properties. The degree to which internal and environmental variables for a given individual patient add unwanted measurement error in PROs is generally quite difficult to evaluate, though an attempt is often made to assess the effects of transient factors in examining test-retest reliability. However, with ClinROs additional access to the process is afforded, thus in developing a ClinRo, it is essential to evaluate how much clinicians tend to agree (or disagree) on a given assessment. There are many characteristics of good measurement that must be shown in addition to inter-rater reliability, including intra-rater reliability and validity. Assessment of agreement between clinicians establishes both reliability and generalizability. Beyond understanding the error sources, demonstrating high agreement among clinicians is an important part of a needed argument in any study for extending the given results beyond a particular study sample, i.e., the generalization justification. This work is also foundational to validity work included in any label claim for a clinician-reported outcome, because if clinicians cannot agree about what they report concerning a given patient, one can hardly consider such information valid.

As with PROs there is the same basic set of frameworks in which one can work on the reliability issue:

"Assessment of agreement between clinicians establishes both reliability and generalizability."

generalizability theory, classical test theory, and modern measurement theory including various modeling approaches such as latent trait modeling. Generalizability theory, while infrequently used, allows one to expand the number of dimensions along which reliability can be assessed. It provides a natural way of assessing both intra- and inter-clinician reliability, providing some insights into the sources of measurement "unreliability" and allowing for greater efficiency in study design. Classical test theory methods (Cohen's kappa, intra-class correlations, etc.) are most often used, though the other frameworks provide certain advantages or benefits. Modeling approaches can provide a deeper understanding of the sources of variability or bias in clinician reports, which can in turn be used in clinician training to increase the reliability of ClinROs.

When it comes to using classical test theory to assess the question of clinician agreement, or consistency (aka inter-rater, inter-judge, or inter-observer reliability, and "intra-" forms of these as well), researchers face a bewildering array of statistics or variants designed for this purpose.[2-8] Though most of these statistics have been in use for some time, design flaws or, at least, complications have been found in some cases of which researchers may not be aware. One of the more popular statistics, the Cohen's kappa has the potential for several problematic issues including biased estimates and paradoxical results.[9-12] Moreover, these statistics come in various forms from which researchers need to select the appropriate one.[6, 10] There is not a single kappa or intra-class correlation (ICC), but rather several versions that vary according to the data collection design and proposed purpose of the measure. McGraw and Wong (1996)[6] report five possible ICC statistics for individual scores and five more for combined scores. Those unfamiliar with these statistics and their potential problems may find themselves confused by results or making mistaken claims.

To correctly assess the inter-clinician reliability of the proposed ClinRO, the researcher will need to carefully consider and navigate a series of issues. To start with, the researcher will need to have a clear idea of how the measure will be used. Is this a measure that will be used for assessment in clinical practice, or is this measure intended for use in group comparisons in clinical trials? The researcher needs to select the agreement statistic that is most appropriate for assessing reliability or risk an endpoint failure. In most cases for clinical trial use, the group comparison use is all that is expected, thus using a statistic fitted to that purpose will be to the researchers advantage. The nature of the ClinRO itself needs to be considered, namely the level of measurement (nominal, ordinal, interval, continuous) it affords. Again, there are different statistics designed for different levels of measurement; not all are appropriate for every use. Another important issue to understand is what criteria your reviewers will require your measure to meet? Will you need to pass a statistical test (e.g., a test of the agreement level surpassing some value) or meet a certain descriptive criterion, i.e., show a level of agreement with some degree of precision? Some thought needs to be given to justify this choice when it is not explicitly defined. Finally, study design also needs to be carefully considered in choosing a statistical approach. In fact, if you have early input into the study design, it is very useful to be able to consider the needs of the statistical approach when designing the study. The relevant design concerns are the numbers of patients and clinicians, whether the selection is random in either or both cases, and the plans for study generalization. The question of minimum sample size and best design in view of various costs will inevitably arise in study planning. In many cases, there are methods for determining optimal design parameter values. The important thing to understand is

"The important thing to understand is that the choice of agreement statistic will be constrained by the design, and some statistics are less desirable for a given design."

Table 1: Sources of Variability and Bias: PRO Versus ClinRO

	Sources of variability and bias in target experience	Sources of variability and bias in reporting
PRO	• transient effects • learning effects • real change	• patient's reporting bias • interaction with the reporting instrument
ClinRO	All the above plus: • clinician effects • clinician x patient interaction effects	All the above

that the choice of agreement statistic will be constrained by the design, and some statistics are less desirable for a given design. If your statistical plan includes a specific statistic, it needs to have a matched design.

While most research has traditionally used the classical test theory approaches to establishing adequate agreement among clinicians, more revealing modeling approaches have been developed (latent trait models and latent class models, among others). Uebersax (1992)[13] provides a brief overview of select models that offer certain advantages over the classical approaches. A major weakness of all the classical statistics is that they do not identify the source of disagreement among clinicians where there is less than perfect agreement, and therefore afford little help in trying to *improve* a ClinRO while under development. Modeling approaches address this weakness. Agreement modeling is able to analyze undifferentiated aspects of the reporting process thereby identifying various sources of disagreement, whether it is overall reporting-level bias, different use of response categories, or just general measurement error which is responsible for the disagreement. This provides the researcher with important information which can often be used to target appropriate revisions to a ClinRO. As a result, modeling approaches provide a clear path to improving the performance of a ClinRO if used early in the development cycle.

ClinROs have come to play an important role among the broad group of COAs. In fact, they are probably much more widespread than PROs in many therapeutic areas where patient insight into their own condition is frequently in doubt (e.g., dementia). Given that importance, care needs to be exercised in the psychometric evaluation. As the reader can see, there are many decisions to be made in assessing inter-clinician agreement in ClinRO submissions. There is no single off-the-shelf approach one can use in every case. It would take a monograph-length tutorial to do justice to all of them, so no attempt has been made to do this here. Nonetheless, this article can serve to raise researchers' consciousness of the complications, alert researchers to potential pitfalls in the process, and increase their awareness of the role experimental design issues play in such appraisals. The statistical approach to assessing inter-clinician agreement needs careful consideration within an overall validation plan to avoid costly mistakes or schedule slippage—outcomes that can occur when not enough data were collected to support your claim, or if analyses need to be redone because regulators indicate the wrong statistical approach was used.

For more information, please contact Randall.Bender@evidera.com.

ACKNOWLEDGEMENTS

I would like to acknowledge the contributions made to this article by William Lenderking, PhD, Senior Research Leader; Dennis A. Revicki, PhD, Senior Vice President and Senior Research Leader; and Karin Coyne, MPH, PhD, Vice President Research, Outcomes Research, Evidera.

REFERENCES

[1] U.S. Food and Drug Administration Glossary of Terms. Available at: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm. Accessed September 16, 2015.

[2] Cohen J. A Coefficient of Agreement for Nominal Scales". *Educ Psychol Meas.* 1960; 20(1):37–46.

[3] Fleiss JL. Measuring Nominal Scale Agreement among Many Raters. *Psychol Bull.* 1971; 76(5): 378–382.

[4] Fleiss JL. *Statistical Methods for Rates and Proportions.* 2nd ed. New York: John Wiley. 1981; pp. 38–46.

[5] Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychol Bull.* 1979 Mar; 86(2): 420-428.

[6] McGraw KO, Wong SP. Forming Inferences about Some Intraclass Correlation Coefficients. *Psychol Methods.* 1996; 1:30-46.

[7] Cohen J. Weighed Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychol Bull.* 1968 Oct; 70(4):213-220.

[8] Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educ Psychol Meas.* 1973; 33:613–619.

[9] Feinstein AR, Cicchetti DV. High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *J. Clin Epidemiol.*1990; 43(6):543-549.

[10] Kraemer HC, Periyakoil VS, Noda A. Kappa Coefficients in Medical Research. *Stat Med.* 2002; 21:2109-2129.

[11]Spitznagel EL, Helzer JE. A Proposed Solution to the Base Rate Problem in the Kappa Statistic. *Arch Gen Psychiatry.* 1985 Jul; 42(7):725-728.

[12] Viera AJ, Garrett JM. Understanding Interobserver Agreement: The Kappa Statistic. *Fam Med.* 2005 May; 37(5):360-363.

[13] Uebersax JS. Modeling Approaches for the Analysis of Observer Agreement. *Invest Radiol.* 1992 Sep; 27(9):738-743.