Clinician-reported Outcomes (ClinROs), Concepts and Development

William Lenderking, PhD, Senior Research Leader;
Dennis Revicki, PhD, Senior Vice President, Outcomes Research

In heathcare, there are many sources of information as to the benefits and potential harmful effects of treatments. Some information is objective and readily observable, such as laboratory findings, biomarkers, and most obviously, mortality. There are other very important endpoints, however, that are more subjective in nature, including symptom reports from patients, evaluation of symptom reports from clinicians, and observer reports. All of these can be used in assessing treatment benefit, which refers to how a patient feels, functions, or survives in daily life and can measure efficacy, effectiveness, and comparative safety. A patient-reported outcome (PRO) is a direct assessment of the patient's experience with, for example, symptoms, while a clinician-reported outcome (ClinRO) is an indirect assessment of the patient's experience. Although not all ClinROs are necessarily considered indirect assessments; both PROs and ClinROs are considered clinical outcomes assessments (COA).

The U.S. Food and Drug Administration (FDA) is increasingly involved in efforts to standardize COAs providing guidance to sponsors on these less objective measures of treatment. In 2009, the FDA released their final FDA Guidance for Industry—Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.[1] In 2010, the Draft Guidance for Industry—Qualification Process for Drug Development Tools[2] was released to address biomarkers and PROs and other rating instruments. In 2011, the FDA held the Clinical Trial Outcomes Assessment Workshop[3] to provide clarity around the key issues of developing and applying outcomes assessments. While much effort has been expended to guide the development and implementation of PROs, there is less clarity regarding the development of ClinROs from a regulatory perspective. The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) ClinROs Good Measurement Practices Task Force[4] was created recently as an ongoing

task force to examine the issues regarding ClinRO development. Below are some basic definitions:

- **PROs**—defined as "…any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else"[1]
- **ClinROs**—no formal definition to date from the FDA, however, defined in the ISPOR Task Force report as "assessment determined by observers with professional training relevant to the measurement"[4] and requirement training in order to perform the assessment successfully
- **Observer-reported outcomes**—An assessment performed by observers without professional training relevant to the measurement, but with knowledge of the patient; no training is required in order to perform the assessment.

While the patient perspective is critically important in understand the effects of treatments, there are many diseases or clinical states in medicine that require the observations of a trained clinician. For example, central nervous system conditions such as depression, Alzheimer's disease, and schizophrenia are all areas where the patient may not be able to provide complete or accurate reports of their medical state due to cognitive impairment, lack of insight, or the inability to assess the severity of negative symptoms. In these cases, clinical experience may be needed to understand the diagnostic meaning of certain symptoms (e.g., the relevance of olfactory hallucinations versus auditory hallucinations) or to clearly assess the severity of a patient's condition, as with patients with advanced cognitive impairment. It is important to note that clinicians include not only physicians, but anyone with specialized training pertinent to the assessment, including psychologists, physical therapists, nurses, EMTs, etc.

ClinRO development most often follows the PRO development requirements.

- **Concept of measurement (construct)**—What is being measured? This is critically important and must be clearly defined from the outset
- **Context of use**—Who is being evaluated? This should be very specific listing exclusion/inclusion criteria used for clinical trials to show efficacy
- **Content validity**—What is the evidence that the instrument is measuring? This requires input from the patient, who is considered the "expert" for PROs (and from clinicians for ClinROs)
- **Confirm measurement properties**—Psychometric testing
- **Consideration of further development**

There are different types of ClinRos, including rating scales, performance-based ratings, and clinical reading of medical signs.

RATING SCALES (E.G., HAMILTON DEPRESSION RATING SCALE)

Rating scales require a clear definition of the concept of measurement, context of use, content validity, and confirmation of measurement properties. The role of clinical experts is truly essential when using rating scales, since rating scales are often administered in a semi-structured way, making clinical judgment a critical part of the overall rating process. While these ClinROs are the most similar to PROs, the role of the patient is not as clear, and perhaps less central, than with PROs, since patients are not reporting on their own symptoms. It is not unusual to have ClinROs developed in parallel with PROs, such as in the Quick Inventory of Depressive Symptoms.

PERFORMANCE-BASED RATINGS (E.G., THE 6 MINUTE WALK TEST OR THE ALZHEIMER'S DISEASE ASSESSMENT SCALE [ADAS-COG])

Performance-based ratings require a clear context of use and measurement properties need to be confirmed; however, these ratings do not always have a clear concept of measurement. For example, some neuropsychological tests may be very good at assessing decrements in neuropsychological functioning, but it is not clear what those results mean in the real world. For example, performance on a word list recall task might be a good measure of decrements in memory, but it is not clear how those decrements would be related to the performance of activities of daily living such as going grocery shopping. In performance-based ratings, the role of the patient is unclear, and often the patient is more of a subject than an expert. In other words, the patient's knowledge of their own condition is irrelevant to the assessment of the measurement concept.

READINGS OF SIGNS (E.G., TENDER POINTS, SPLEEN SIZE, CT SCANS)

Again, clear definition of concept of measurement, context of use, and confirmation of measurement properties are required, but content validity may not be relevant. (Of course, validity is always relevant when evaluating a measurement, but content validity as it refers to the verbal content of items is not relevant in the evaluation of the validity of signs as diagnostic or evaluative predictors of clinical outcomes.) Reading of signs also requires the most training in order to perform them correctly, and patient input is not important to the development process.

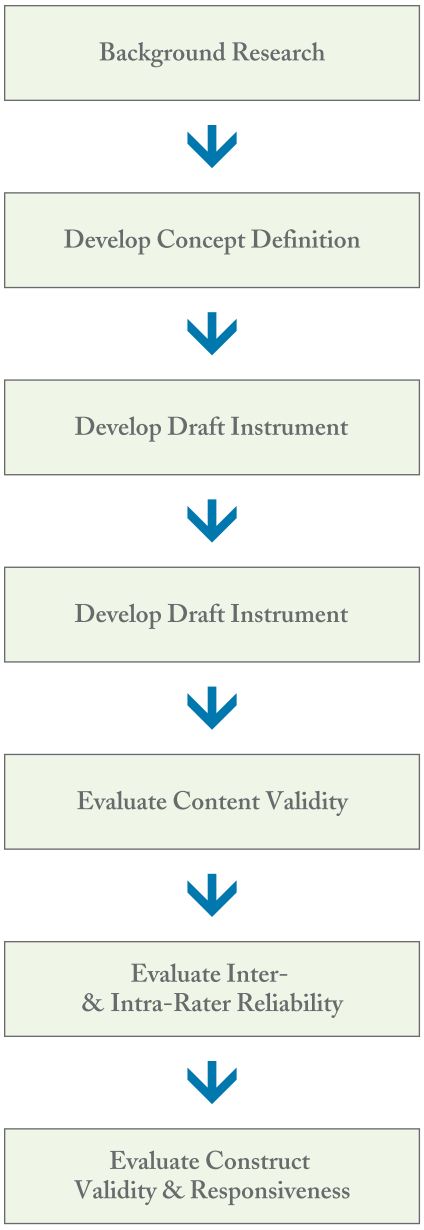


figure 1

DEVELOPMENT OF CLINICIAN-REPORTED OUTCOME MEASURES

As mentioned earlier, the process of developing a ClinRO is very similar to that of developing a PRO *(see Figure 1)*. A systematic approach to determining the conceptual framework, based on multiple sources of evidence, is required. This should consist of gathering appropriate background information through literature reviews, patient interviews, and expert clinician consultation. This clinician input is particularly important in the development of rating scales to define the nuances of individual item responses. The development process should be an iterative approach, where there are multiple cycles of development, review and revision. An expert clinician working group is necessary. Attention must be given to symptom definitions and descriptions, as there is often discrepancy between disease definitions in clinical practice and those used in clinical trials. Inclusion/exclusion criteria in clinical trials are often used as the basis for labeling, so things such as treatment setting, age and other demographic factors, and study design may be relevant. Meaningful and relevant response scales must be identified and detailed instructions for making the clinician ratings defined. Development of training materials, and appropriate training of those using the outcome assessment, is then necessary to achieve reliable results.

The measure must then be validated, including internal consistency and analyses of scale performance (only relevant for rating scales), test-retest reliability (are the same assessments being found over time), intra-rater reliability (is the same rater getting the same results over time), and inter-reliability (are different raters getting the same results over time). Inter- and intra-rater reliability is key to a successful ClinRO and minimizing rater variability should be a major focus. Considerations related to concurrent and divergent validity are also important—what is the evidence that the ClinRO fits into a theoretical network as expected? Equally important is the linkage of a measured change on a ClinRO to clinically meaningful outcomes.

IMPORTANT THINGS TO REMEMBER AND CONSIDER

- Is a ClinRO necessary at all? Can the patient report on their own experience? Can the clinician observe the concept?
- Many of the same considerations involved in developing a PRO are relevant to a ClinRO.
- The role of the patient varies depending on the type of ClinRO and how aware the patients are of the phenomenon being reported.
- The role of the expert/clinician is critical in all aspects—development, validation, and implementation.
- ClinROs require a systematic approach to instrument development, with attention to concept definition, rater instructions and training, and psychometric evaluation.
- Training in administration of the ClinRO is essential for reliable results across raters.
- ClinROs provide the clinician's perspective on patient-reported outcomes.

CASE STUDY— DEVELOPMENT OF A CLINRO INSTRUMENT: CLINICAL GLOBAL IMPRESSION FOR SCHIZOAFFECTIVE DISORDER[7]

Background

Schizoaffective disorder is a complex psychiatric condition characterized by concurrent psychotic and mood symptoms. Treatment can differentially affect the various symptom domains in schizoaffective disorder, including positive, negative, cognitive, manic, and depressive symptoms. Studies of schizoaffective disorder have traditionally used a combination of scales designed to assess schizophrenia or mood disorders (e.g., PANSS, Hamilton Depression Rating Scale, Young Mania Rating Scale), however, there is no scale to assess the global severity and change specific to schizoaffective disorder. The development of the Clinical Global Impression for Schizoaffective Disorder (CGI-SCA) was undertaken.[7]

Development

A literature review was completed on key symptoms and clinical assessments used in schizoaffective clinical trials. A working group, comprised of psychiatrists, clinical psychologists and psychometricians, was convened to identify key measurement concepts, develop concept definitions, and determine measurement approach and rating scales. The resulting CGI-SCA was based on the clinical global impressions (CGI) for bipolar disorder and schizophrenia and measures severity (CGI-S-SCA) using a 1–7 rating scale (normal to severely ill) and symptom change (CGI-C-SCA) using a 1–7 rating scale (very much improved to very much worse). Each measurement includes ratings of the four domains of schizoaffective disorder (positive, negative, depressive, and manic symptoms) as well as an overall rating measuring the clinician's assessment of overall severity of the patient. The rating scale The CGI-SCA instruction manual was then developed with definitions for symptom states and instructions for rating symptom severity, specifically defining the four domains to help clinicians differentiate manic from positive symptoms and depressive from negative symptoms.

Inter-rater reliability

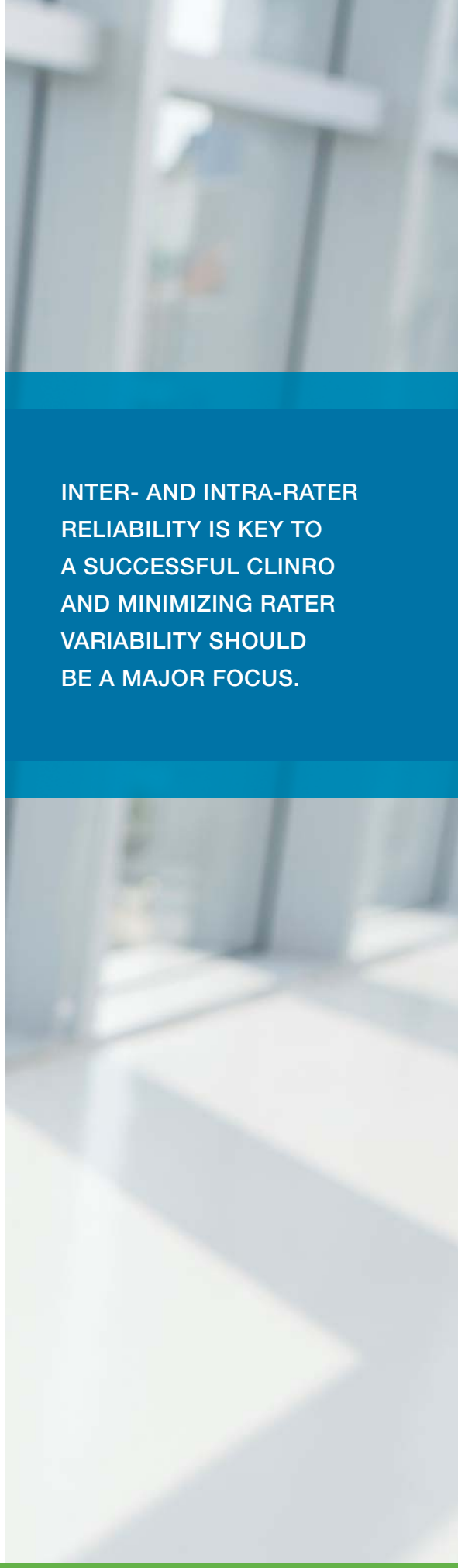
Videotaped interviews of 12 patients with schizoaffective disorder were independently rated by two trained raters. The intra-class correlation coefficient ranged from 0.62 to 0.88 for the four symptom CGIs, but the overall rater reliability scale was 0.5, which suggested that the training materials needed to be improved.

Test–re-test reliability

Two sets of videotaped interviews of a single actor representing two time points were rated by ten randomly selected clinical trial raters from the US, India, and Eastern Europe for a total of 30 raters. First and second ratings by each rater were separated by two weeks. In the first interview the actor portrayed a severely ill patient and in the second interview, he portrayed a patient who was much more improved. Test–re-test reliability was moderate to excellent, with scores of 0.60 to 0.89, while results from the severity scale and change scale were more moderate, from 0.48 to 0.63. These results showed that clinicians were having difficulty identifying the changes that were present in the actor in the second video, indicating that the training materials needed to be refined to provide more clarity.

Validity

Convergent and divergent validity at baseline for schizoaffective subjects in two pooled international trials (n=614) were reviewed and both showed good evidence of convergent and divergent validity, which was encouraging. Moderate to large correlations were observed between



INTER- AND INTRA-RATER RELIABILITY IS KEY TO A SUCCESSFUL CLINRO AND MINIMIZING RATER VARIABILITY SHOULD BE A MAJOR FOCUS.

EFFECTIVE SIZES FROM POOLED CLINICAL TRIALS (N=614)	
Scale	*Week 6*
CGI-SCH	
Overall	0.32
Positive	0.30
Negative	0.21
Depression	0.27
Manic	0.37
PANSS	
Total	0.33
Positive	0.37
Negative	0.21
YMRS	
Overall	0.43
HAM-D-21	
Overall	0.26

figure 2

the symptom CGIs and PANSS, Young Mania Rating Scale, and Hamilton Depression Rating Scale scores.[7]

The correlations between the CGI-SCA overall global scores and the other psychiatric rating scales were small to large.

Sensitivity to Change

Allen et al.[7] also evaluated the responsiveness of the CGI-SCA to changes in clinical status using data from two pooled clinical trials *(see Figure 2)*. The observed effect sizes were comparable for the CGI-SCA to the effect sizes for the PANSS, Young Mania Rating Scale and Hamilton Depression Rating Scale.

Conclusions

The CGI-SCA was developed to evaluate global impressions across the four relevant domains of schizoaffective disorder and to provide a global overall rating of patients' clinical status. The results provide support for the reliability and validity of the CGI-SCA for application in clinical trials of patients with schizoaffective disorder. More important, the CGI-SCA was able to assess severity and change in individual symptom domains, and in the overall status of patients with schizoaffective disorder. Further research is needed to confirm the psychometric qualities of the CGI-SCA in broader patient populations with a range of disease severity.

SUMMARY

Assessing treatment benefit from an outcomes perspective is growing in attention and use, and clinicians are an important resource when the patient cannot accurately report on their own experiences. While there is clear guidance on PRO development, the regulatory guidance has not been clearly defined for ClinROs to date. The FDA has made recommendations on clinical outcomes assessments, which include ClinROs, stating they should be specific to a context of use; specific to a version of an instrument including mode of administration and training materials; and specific to the concept of a measurement. Additionally, for COAs that measure treatment benefit indirectly (e.g., some ClinROs), qualification also includes a review of the evidence that the concept assessed is an adequate replacement for how patients feel or function in daily life. Specific guidance on ClinROs should be expected in the future, although the timing for such guidance is unknown.

For more information, please contact William.Lenderking@evidera.com or Dennis.Revicki@evidera.com.

References

[1] Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. December 2009. Available at: http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf. Accessed 23 September 2013.

[2] Draft Guidance for Industry—Qualification Process for Drug Development Tools. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf.

[3] Clinical Trial Outcome Assessments: Identifying Intrinsic Characteristics. Clinical Trial Outcome Assessment Workshop White Oak, October 2011. http://www.fda.gov/downloads/Drugs/NewsEvents/UCM277124.pdf. Accessed on: 23 September 2013

[4] ISPOR PRO Task Forces—Clinician-Reported Outcomes Good Measurement Practices Task Force. Available at: http://www.ispor.org/meetings/neworleans0513/releasedpresentations/FORUM-1-Clinros_TF_Powers.pdf.

[5] Papodopoulos E. Clinical Trial Outcome Assessments: Measuring Treatment Benefit Across the Lifespan. Gastroenterology Regulatory Endpoints and the Advancement of Therapeutics: Eosinophilia Esophagitis. September 19, 2012. Available at: http://www.meritagepharma.com/pdf/10.Papadopoulos-COA.pdf.

[6] U.S. Food and Drug Administration. Clinical Outcome Assessment Qualification Program. Available at: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm. Accessed on 232 September 2013.

[7] Allen MH, Daniel DG, Revicki DA, Canuso CM, Turkoz I, Fu DJ, Alphs L, Ishak KJ, Bartko JJ, Lindenmayer JP. Development and Psychometric Evaluation of a Clinical Global Impression for Schizoaffective Disorder Scale. *Innov Clin Neurosci.* 2012 January; 9(1): 15–24.