

APPLIED CLINICAL TRIALS

Volume 18, Number 1 January 2009

YOUR PEER-REVIEWED GUIDE TO GLOBAL CLINICAL TRIALS MANAGEMENT

Trial Design



PHOTOGRAPHY: GETTY IMAGES

Adam Butler

Seeking Guidance on Rater Reliability

.....
The case for standardizing the use
of clinician rated outcome measures
to improve research studies.
.....

Rater reliability is the cornerstone of data accuracy for drug trials that depend on clinician-rated subjective instruments to assess efficacy. Every pharmaceutical company seeks unambiguous evidence that its product is the new best treatment for a given disorder. Increasing governmental scrutiny of new drug applications combined with rising costs of product development and rising industry competition are pushing companies to improve the margin of success for their clinical trials.

Demonstrating efficacy requires a confluence of excellence across multiple areas of clinical trial design and implementation. Protocol design issues, clinical features of the targeted disorder, the effectiveness of the drug, and study implementation factors all interact to affect signal detection. Pharmaceutical companies must consider all of these areas with meticulous attention to effectively demonstrate a significant drug effect.

In particular, measurements of efficacy must be reliable for a trial to achieve statistically significant results. Some clinical trials employ outcome measures generally considered to be objective in nature, such as blood tests or other laboratory values. Other types of clinical trials depend on the proficiency of clinician raters to assess the presence and severity of symptoms based upon subjects' reports during clinical interviews. Clinical trials that rely on subjective assessments either reported by a patient or collected by a clinician are far more susceptible to interpretation and fluctuations than trials whose efficacy measurements involve objective changes in laboratory measured variables.



Researchers involved in trials employing objective efficacy measures might be surprised to learn how heavily some clinical trials depend on the accuracy of a subject's self-report or a clinician rater's distillation of a subject's reported symptoms.

Currently, subjective ratings are the only practical and effective methods available to assess physical and emotional states—such as depression and pain—despite the almost certain biological origins of these central nervous system disorders. Clinician rated quality-of-life assessments also depend on patients' reports of their functional and emotional states. Confounding variables include the ability of subjects to be accurate when reporting their symptoms and symptomatic changes, clinician raters' interviewing skills, and their ability to be objective and investigative when collecting, assessing, and reporting accurate data from their subjects.

This article examines the rater reliability landscape in trials utilizing clinician-administered, clinician-rated outcome measures. Challenges to the consistent acquisition of data, guidance considerations, and methods to improve the outcome will also be addressed.

Measurement challenges

Obtaining reliable results with clinician-administered and -rated subjective measures poses unique challenges to the clinical trial industry. Variation in this type of subjective scale administration technique and scoring is widespread.¹ Generally, pharmaceutical companies select the trial sites that they believe will be best able to deliver accurate results. The importance of integrity at the level of the site cannot be overestimated. Previous experience with sites gives pharmaceutical companies and clinical research organiza-

These challenges require a comprehensive review of every rater's credentials and evidence of their interview skills.

tions useful information regarding the investigators at each site. While financial and time pressures may generate enrollment bias, sponsors and clinical research organizations work diligently to prevent this from occurring.

At the designated sites, clinical investigators may interview and rate subjects themselves or delegate one or both of these tasks to other clinicians at the site. When investigators delegate subject rating to other clinical staff members at their site, they rarely have time to supervise the scale-specific interviews and scoring on a regular basis. Raters in the United States, when examined as a group, have diverse educational backgrounds and varied experiences with patients in any given disease population.²

In addition, endpoint selection should be predicated on both regulatory pathways and scientific precedence—reliable, valid measurements are a requirement.

The Hamilton Depression Rating Scale-17 (HAM-D-17) is a clinician-administered and -rated instrument commonly used in antidepressant drug trials. Consider an antidepressant trial that designates the HAM-D-17 as the primary efficacy measure. In this trial, raters at all sites would be required to perform a HAM-D-17 research interview across visits and score this instrument so that the score accurately reflects the severity of the subjects' depressive symptoms during the week preceding the visit. To do this successfully, raters must have a working knowledge of the spectrum of mood disorders, as well as a thorough knowledge of the administration guidelines and scoring conventions of the HAM-D-17 rating scale.³ In addition, the raters must be consistently proficient at performing research interviews.

The challenge for the sponsor company is not only to be reasonably certain that a given rater will administer and rate the same scale in the same manner at every visit for all patients (intra-rater reliability), but to be reasonably certain that raters at all sites will administer and rate the same scale in the same way at every visit (inter-rater reliability). Studies have shown that improved rater reliability can decrease the minimum sample size necessary to demonstrate efficacy, thus decreasing the overall cost of a clinical trial.⁴

Meeting these challenges requires a comprehensive review of every rater's credentials and experience with a given scale, as well as evidence of each rater's interview skills and rating proficiency on the protocol designated scales.

Seeking guidance

Until recently there was little guidance available from regulatory agencies regarding the use of subjective endpoints in clinical trials. However, in response to the increasing number of failed studies, the increasing demand for evidence, and the increasing number of clinical trials that rely on subjective measures as pivotal endpoints, regulatory authorities are beginning to specifically address the use of subjective endpoints within clinical research.

Regulatory agencies exist in many countries, however, guidance requirements are not always consistent. The International Conference on Harmonization was formed to work toward a goal of international consistency in guidance recommendations for clinical trials. Primary participants in this effort have been the U.S. Food and Drug Administration (FDA), the European Agency for the Evaluation of Medicinal Products (EMA), and the Japanese Pharmaceuticals and Medical Devices Agency (PMDA). For purposes of this article, the focus will be on guidance recommendations made by the FDA and EMA.

Agency input

The FDA and EMEA have published general guidelines for clinical trials. In 2006, the FDA issued a draft guidance titled Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.⁵ Although no specific guidance regarding the use of clinician-rated scales has been issued, the FDA has indicated in recent presentations that it will likely address these instruments in a future guidance document.

Until a specific guidance is issued, much can be learned from existing FDA and EMEA guidance documents and the aforementioned draft guidance on patient-reported outcomes. One important FDA document is International Conference on Harmonization (ICH): E9 Statistical Principles for Clinical Trials.⁶ The section Composite Variables (2.2.3) contains this statement:

When a rating scale is used as a primary variable, it is especially important to address factors such as content validity, inter- and intra-rater reliability and responsiveness for detecting changes in the severity of disease.

In a move that emphasizes the importance of differential consideration of the different diagnostic categories of mental disorders, the FDA and EMEA have distributed draft guidance documents that focus on clinical trials of antidepressant medications. The FDA's recommendations for clinical trials in the field of depression are documented in Guidelines for the Clinical Evaluation of Antidepressant Drugs.⁷ Both this document and the EMEA's document "Note for Guidance on Clinical Investigation of Medicinal Products in the Treatment of Depression"⁸ contain reviews of study design, specifically targeted depressive syndromes, and populations. The authors of the FDA document also recommend the identification of target symptoms of depression so that these can be more closely monitored for change. Sponsors are urged to carefully consider site selection, with a focus on investigator experience with both depressive disorders and the evaluation of psychiatric drugs.

Regarding inter-rater reliability, the authors recommend that "[e]fforts should be made to establish agreement on the use of diagnostic and descriptive terms as well as the handling of assessment instruments." This recommendation highlights the need for training in a particular trial's specific diagnostic and assessment procedures.

In pursuit of efficiency in the journey of an investigational drug through the clinical trial and approval process, these recommendations further emphasize the importance of good communication between investigator, site monitor, and the regulatory agency. Good communication will ensure that the clinical trial design is adequate, which en-

sures that the data collected will allow all parties to "competently evaluate both the drug's efficacy and its safety" in a sensible manner.

The aforementioned EMEA document on antidepressant clinical trials, published in 2002 by the Committee for Proprietary Medicinal Products (CPMP)⁷ also attempts to focus the attention of the pharmaceutical industry on antidepressant drug trials. In addition to recommendations that it shares with the FDA regarding study design and execution, specific guidelines address measurement. EMEA's CPMP recommends the careful assessment of efficacy criteria and states that in antidepressant drug trials, in particular, the Montgomery-Asberg Depression Scale, the HAM-D-17, and the Clinical Global Impression Scale should be used. In addition, this EMEA guidance includes some detailed language regarding the recommended subjective endpoints and their use. In the section Study Design (5.2), the EMEA's CPMP advises that:

Investigators should be properly trained in evaluating the patient. Inter-rater reliability scores (kappa) should be documented for each investigator in advance and if necessary during the study, both with regard to the diagnosis and to rating scales used for efficacy and/or safety, where relevant.

Nearly identical language is used in similar EMEA guidelines published for anxiety disorders, bipolar disorder, schizophrenia, and other psychiatric disorders. This enriches the previously developed International Conference on Harmonization's ratified version of the E9 Statistical Principles for Clinical Trials guideline.⁶

Current best practices

In 2007 the FDA published draft guidance on the supervisory responsibilities of investigators, "Guidance for Industry: Protecting the Rights, Safety, and Welfare of Study

Subjects—Supervisory Responsibilities of Investigators."⁹ This document places distinct and detailed responsibilities on the investigators regarding the delegation of study-related tasks to qualified individuals, the adequacy of training

received by study staff, the supervision afforded to site staff, and the supervision of third parties involved in the conduct of the study, such as independent laboratories.

In section 3.1 of this guidance, Supervision of the Conduct of a Clinical Investigation, the authors state:

In assessing the adequacy of supervision by an investigator, FDA focuses on four major issues:

1. Whether delegated individuals were qualified to perform such tasks
2. Whether study staff received adequate training on how to conduct the delegated tasks and were provided with an adequate understanding of the study

The guidance stresses the criticality of training both patients and clinicians.

3. Whether there was adequate supervision and involvement in the ongoing conduct of the study
4. Whether there was adequate supervision or oversight of any third parties involved in the conduct of a study to the extent such supervision or oversight was reasonably possible.

This reinforces the need for staff to not only be adequately educated and experienced but also competent to perform the specific tasks—to administer a rating scale, take a blood pressure reading, perform an electrocardiogram, and so on. This guidance also places distinct responsibility on the investigator to ensure that staff members receive all relevant training provided by the sponsor.

The increasing use of patient-reported outcomes led the FDA to develop a guidance regarding the use of patient-reported outcomes in clinical trials. “Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims,”⁷⁵ released in 2006, supports the need to develop and administer patient-rated instruments with appropriate rigor. In particular, the guidance stresses the criticality of training both patients and clinicians in the effective use and application of these instruments with the goal of minimizing inter-subject variability. The FDA maintains that comprehensive, consistently applied training is necessary to standardize trial conduct. More specifically, the FDA has suggested three specific areas of standardization to support the use of patient-reported outcomes:

- Training and instruction for patients
- Interviewer training and interview format
- Instructions for investigators.

Governmental guidance for industry regarding pharmaceutical drug trials highlights the state-of-the-art philosophical and procedural considerations to be taken into account in the implementation of such studies. These docu-

Development of these initial guidance recommendations by the FDA and EMEA is only the beginning of what promises to be a complex educational process.

ments also present recommendations for the safe and successful implementation of clinical trials for drugs targeting different indications. All pharmaceutical sponsors and other companies closely involved in clinical trials should pay close attention to regulatory evaluations, recommendations, and decisions. A universal understanding of the relevant FDA and EMEA guidelines is useful. Researchers involved in implementing trials for which guidelines do not yet exist may extrapolate from the existing guidances to improve the likelihood of successful trials.

The emphasis on achieving improved rater performance cannot be overlooked. Governance agencies’ recommendations demonstrate their awareness of the pitfalls into which inexperienced or inadequately trained investigators and their staff may fall. The importance of rater experience and scale specific training is becoming an increasingly popular focus of study; this likely reflects the recent pharmaceutical company interest in detailed scrutiny of rater qualifications, experience, and scale-specific rating proficiency. Further, raters’ educational backgrounds and ranges of experience are being more carefully documented. Increasingly more investigator meeting time is being devoted to rater training, the subsequent evaluation of scale-specific rating proficiency, and research interview training and assessment.

Improving conduct and outcomes

Over the past decade, clinical researchers have been contributing to the literature in support of increased rater scrutiny and improved rater training in the use of clinician administered rating scales.¹⁰ A number of researchers have assessed the importance of scale-specific training in the development of inter-rater reliability. In addition, Kobak et al.¹¹ implicate inadequate rater training and inadequate rater competency in the failure of central nervous system drug trials. They list six general skills necessary for rater competency, including:

- Conceptual understanding of the disorder of interest
- Clinical experience with the population under study
- Generic interviewing skills
- Expertise in symptom rating scale administration
- Understanding of the research milieu
- Scale-specific expertise.

The development of these initial guidance recommendations by the FDA and EMEA is only the beginning of what promises to be a complex educational process. The increasing focus on improving data quality in clinical research highlights increased rater reliability as a means to improve signal detection in clinical trials that employ subjective endpoints. Careful rater selection, subsequent intensive training, and proficiency assessments of these raters are important first steps. When rater reliability concerns are addressed in combination with a careful study design and highly specific site selection, statistically significant drug effects will be more easily demonstrated.

Guidance from regulatory agencies will help direct pharmaceutical companies toward the implementation of practices that will improve rater reliability in clinician rated subjective scales. The combination of meticulous protocol development, careful site selection based on an examination of the raters’ credentials and experience, rigorous training and assessment programs, and close data monitoring should improve the reliability of rating these trials.

Rater reliability will only be consistently achieved with much consideration and with input from many different sources. Research in the area of rater reliability, combined with data from companies supporting CNS clinical trials should be used to inform future trials. Government agency recommendations represent a major leap forward in the standardization of industry efforts in the quest for achieving the best possible ratings reliability.

Acknowledgements

Thanks to both Joan Busner and Marian Ormont for their contributions to the article.

References

1. M.J. Muller and A. Dragicevic, "Standardized Rater Training for the Hamilton Depression Rating Scale (HAM-D-17) in Psychiatric Novices," *Journal of Affective Disorders*, 77, 65-69 (2003).
2. A. Bullinger and S. D. Targum, "Rater Experience in CNS Clinical Trials," Poster Presentation, New Clinical Drug Evaluation Unit, 44th Annual Meeting, Phoenix, AZ (2004).
3. M. Hamilton, "Development of a Rating Scale for Primary Depressive Illness," *British Journal of Social and Clinical Psychology*, 6 278-296 (1967).
4. K.O. Cogger, "Rating Rater Improvement: A Method for Estimating Increased Effect Size and Reduction of Clinical Trial Costs," *Journal of Clinical Psychopharmacology*, 27 (4) 418-420 (2007).
5. FDA Center for Drug Evaluation and Research, Center for Biologicals Evaluation and Research, Center for Devices and Radiological Health, *Guidance for Industry, Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims—Draft Guidance* (FDA, Rockville, MD, February 2006).
6. FDA Center for Drug Evaluation and Research and Center for Biologicals Evaluation and Research, *Guidance for Industry: E9 Statistical Principles for Clinical Trials*, ICH-ratified (FDA, Rockville, MD, September 1998).
7. FDA Psychopharmacological Drug Group, *Guidelines for the Clinical Evaluation of Antidepressant Drugs* (FDA, Rockville, MD, September 1977).
8. EMEA Committee for Proprietary Medicinal Products, Note for Guidance on Clinical Investigation of Medicinal Products in the Treatment of Depression (EMEA, London, 2002).
9. FDA Center for Drug Evaluation and Research, FDA Center for Biologicals Evaluation and Research, Center for Devices and Radiological Health, *Guidance for Industry: Protecting the Rights, Safety, and Welfare of Study Subjects—Supervisory Responsibilities of Investigators* (FDA, Rockville, MD, May 2007).
10. S.D. Targum, "Evaluating Rater Competency for CNS Clinical Trials," *Journal of Clinical Psychopharmacology*, 26 (3) 308-310 (2006).
11. K.A. Kobak, N. Engelhardt, J.B.W. Williams, J.D. Lipsitz, "Rater Training in Multicenter Clinical Trials: Issues and Recommendations," *Journal of Clinical Psychopharmacology*, 24 (2) 113-117 (2004).

Adam Butler is an associate vice president in the training and education group at United BioSource Corporation, 575 E. Swedesford Rd., Ste 200, Wayne, PA 19087, email: adam.butler@unitedbiosource.com.



United BioSource Corporation

Adam Butler
adam.butler@unitedbiosource.com