Exposing the Impact of Ongoing Assessment Feedback on Site Rater Performance: Does Our Work Matter?

Kim F. Baldwin, MA, MFT¹, Rolana Avrumson, MS², Elan A. Cohen, Ph.D.¹, Bethanne Friedman, Psy.D.¹, Melissa A. Carbo, MS¹, Natalie C. Glaug, BA¹, Andrew E. Komovosky, MS¹, Elian Raspomaki, PhD, Colleen R. Rock, BA¹, Michael F. Murphy, M.D., PhD¹

¹Worldwide Clinical Trials, King of Prussia, PA

Abstract

Objective: Rater training programs provide ongoing data surveillance to ensure appropriate scale administration, scoring, and protocol parameters are maintained. However, there is a paucity of research on the impact of feedback on reducing overall rater errors within industry-sponsored clinical trials. The current study adds to the literature by examining the impact of ongoing site rater feedback on rater accuracy as well as protocol adherence.

Design: Data from a global 36-week clinical trial evaluating negative symptoms and cognitive function in outpatient schizophrenia subjects were evaluated retrospectively. Participants were randomized to receive either 8 weeks of the active medication haloperidol or placebo and were submitted all screening and baseline diagnostic and symptom severity scales to external, expert clinicians who reviewed the scales to detect rating errors based on their not following administration and scoring conventions and protocol instructions.

Results: Data were derived from 27 sites across 27 centers in 137 patients and 217 visits. Statistically significant findings were observed for the effect of feedback on rater performance (ANOVA; p < 0.0001). Based on a mixed model for repeated measures (with numbers of errors logarithmically transformed) the number of errors per rater was 4.0 (95% CI: 3.3, 4.8) before feedback, a statistically significant reduction of 2.8 [1.7, 4.3] errors per visit afterward.

Conclusion: Though a causal relationship cannot be inferred without a concurrent control group, results suggest a significant relationship between ongoing assessment feedback and rater performance. Implications for training and quality assurance methodology, with suggestions for future studies, will be discussed in the poster.

Background

Literature examining the reasons for the increased number of failed clinical trials has emphasized the crucial role that quality ratings play in the overall integrity and approval of clinical trials (Kemp et al. 2010; Kobak et al. 2005). Quality assessments in multicenter psychiatric research in particular are operationally defined as those changes in scores consistent with expert adherence to practice conventions within, and between, investigations sites. Additionally, Ventura et al. (1999) and Targum et al. (2002) reported a significant increase in accuracy across sites, with more experienced rater performing at levels of the same or higher accuracy, than less experienced rater. These works demonstrate that rater training and calibration at the start of a trial is not sufficient to ensure ongoing quality and reliable scores on a ongoing basis.

Buxner et al. (2011) reviewed the effectiveness of data surveillance and remediation programs in four Major Depressive Disorder, Schizophrenia, and Bipolar Disorder multi-center clinical trials. The investigations found a large rate of subject visits flagged for clinical quality in the second half of the four trials and concluded that accuracy improved as a function of time as well as the study’s remediation program. Taking into account not only time in a site surveillance program but also feedback, it is the authors of the study who revised the quality scores in the results description. Analysis was done using a mixed effects model for repeated measures with feedback and visit number as fixed effects, rater as a random effect, and patient as a nested within rater random effect.

Statistically significant findings were observed for the effect of corrective feedback on rater accuracy (Table 1). The raw mean number of errors across all raters and visits reduced from 5.43±2.2 before feedback to 2.23±1.5 after feedback (Figure 1). The mixed model-Least Squares-Mean (LS) estimate for the mean number of errors per rater across all visits was 4.9 (95% CI: 2.7, 5.8) before feedback, and 1.8 (1.0, 1.5) after feedback, representing a statistically significant 32% decrease in errors per rater (Table 2). Visit number also had a statistically significant effect on the number of errors (Figure 2) shows the downward trend for the average errors per rater over the number of visits. In further analyses, no significant effects on rater accuracy were observed for the site location (US vs. CEE), or the interaction of visit number with feedback.

Results (cont.)

Table 2. Mixed Model Least-Squares Means Estimates for Number of Errors per visit per rater

Table 1. ANOVA Table (type 3 tests of fixed effects)

Design (cont.) All raters were required to have at least two years of schizophrenia population experience along with two years of experience in efficacy scale-specific experience, Logically speaking the number of standardized rating errors, i.e., (an)acta/erctor inmincent targeted behavior and psychopathology relevant to the trial. Remediation was provided raters to decrease standardized scale scores that did not agree (e.g., 80%) with the established gold standard video scores and whose applied interviews did not meet minimum quality standards on the four defined domains, rapport, instrument comprehension, level of questioning, and absence of therapeutic intervention.

Once the PIs were certified to rate in study, they continued to receive feedback from independent expert clinicians. Sites were electronically to submit data and report findings, as administered by the PIs, for each screening and baseline data submission, each rater in the data surveillance program could not know about the protocols' analysis and were therefore not based on providing feedback. A group of expert clinicians reviewed the types of errors identified in the data surveillance program and categorized them so as to be consistent with the following categories:

1. Scale Administration Errors: the rater did not follow administration conventions as defined by scale-specific instructions

2. Scale/Summation/Counting Errors: the rater incorrectly summed a diagnostic step by step error

3. Diagnostic errors: the rater coded an incorrect diagnosis

4. Protocol-specific procedures were not properly followed, per protocol

5. Missing/Insufficient Source Notes: the rater did not provide enough source notes to support the code or score given

Results

Data were derived from 27 PIs across 27 centers in 137 patients over 217 visits. Of the 27 raters, 26 received feedback during the trial period. Visits conducted by PIs accounted for 51% of the total number of screening and baseline visits reviewed in the study. The majority of raters who received feedback did so after their first visit (screening visit, 10/26 raters), and all 26 raters had feedback received by their seventh visit.

Visits were categorized as having occurred before or after feedback started. The median number of visits per rater and per site was 5 (IQR 2-18). The authors of this paper modeled the number of errors per visit or per rater as a function of the mean accuracy in the feedback included for each rater. Due to deviation from normal, the number of errors was log-transformed as: log(y+1) and back-transformed to the original scale in the results description. Analysis was done using a mixed effects model for repeated measures with feedback and visit number as fixed effects, rater as a random effect, and patient as a nested within rater random effect.

Statistically significant findings were observed for the effect of corrective feedback on rater accuracy (Table 1). The raw mean number of errors across all raters and visits reduced from 5.43±2.2 before feedback to 2.23±1.5 after feedback (Figure 1). The mixed model-Least Squares-Mean (LS) estimate for the mean number of errors per rater across all visits was 4.9 (95% CI: 2.7, 5.8) before feedback, and 1.8 (1.0, 1.5) after feedback, representing a statistically significant 32% decrease in errors per rater (Table 2). Visit number also had a statistically significant effect on the number of errors (Figure 2) shows the downward trend for the average errors per rater over the number of visits. In further analyses, no significant effects on rater accuracy were observed for the site location (US vs. CEE), or the interaction of visit number with feedback.

Design

This study assessed US and Central and Eastern European (CEE; 5 countries) raters' adherence to the study protocol, as well as adherence to administration and scoring conventions/guidelines on several diagnostic, symptom-severity, and global functioning outcome scales. All scales were administered in an industry-sponsored interventional outpatient Schizophrenia study. The Principal Investigators (PIs) were the raters for the current analyses. Prior to the start of the PI training, raters' initial indication and scale experiences were reviewed against predefined criteria.

Model Mixed Estimates

Table 1. ANOVA Table (type 3 tests of fixed effects)

Effect

Visit Number Feedback

NumDF DenDF FValue ProbF 1 1.71 0.00059

Feedback 51 28.35 < 0.0001

References provided upon request.