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

Poster presented at the Annual CNS Summit Meeting, October 2015, Boca Raton, FL

Kim F. Baldwin, MA, MFT[1], Rolana Avrumson, MS[1], Elan A. Cohen, Ph.D.[1], Bethanne Friedmann, Psy.D.[1], ,Melissa A. Carbo, MS[1], Natalie C. Glaug, BA[1], Andrew E. Komorowsky, MS[1], Eleni Rapsomaniki, PhD, Colleen R. Rock, BA[1] , Michael F. Murphy, M.D., PhD[1]

[1]Worldwide Clinical Trials, King of Prussia, PA

Abstract

Objective: Rater training companies provide ongoing data surveillance to ensure appropriate scale administration, scoring, and protocol parameters are maintained. However, there is a paucity of data exploring whether site raters improve with ongoing data surveillance. While Targum (2006) and Busner et al. (2012) demonstrated the effectiveness of ongoing training in reducing overall rater errors within industry-sponsored clinical trials, the current study adds to the literature by examining whether external rater feedback impacts individual rater accuracy as well as protocol adherence.

Design: Data from a global, 26 week clinical trial evaluating negative symptoms and cognitive function in outpatient schizophrenia subjects were evaluated retrospectively. Previously qualified and credentialed site raters submitted all screening and baseline diagnostic and symptom severity scales to external, expert clinicians who reviewed the scales to detect raters' errors based on their not following scales' administration and scoring conventions and protocol instructions.

Results: Data were derived from 27 raters across 27 centers in 137 patients and 217 visits. Statistically significant findings were observed for the effect of feedback on rater accuracy (ANOVA; p <0.0001). Based on a mixed model for repeated measures (with number of errors logarithmically transformed) the number of errors per rater was 4.0 [95% CI, 2.7, 5.8] before feedback, and 1.2 [1.0, 1.5] after feedback, representing a statistically significant reduction of 2.8 [1.7, 4.3] errors per visit per rater.

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 outlined 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 success of clinical trials (Kemp et al. 2010; Kobak et al., 2005). Quality assessments in multicenter psychiatric research in particular, are operationally defined as those characterized by consistent adherence to scale conventions within, and between, investigative sites. Additionally, Ventura et al. (1998) and Targum et al. (2006) demonstrated that rater performance improves significantly across experience levels with repeated applied training exercises. These findings make compelling arguments for comprehensive rater training at the outset of a trial. However, Rothman et al. (2011) also illuminated the tendency for reliability coefficients to decrease significantly over time following this initial training - a phenomenon labelled as "Rater Drift". Ventura (1998) found reduced agreement in scoring and interview quality at biannual quality assurance checks for both experienced and neophyte raters. This documented phenomenon demonstrates that rater training and calibration at the start of a trial is not sufficient to ensure ongoing quality and reliable ratings once a trial is underway.

Busner 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 investigators found a lower 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 studies' remediation program.

Taking into account not only time in a study surveillance program but also feedback to site raters (pre and post feedback), the authors of the current study sought to further verify that clinical and protocol oversight yields tangible reduction of errors.

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 analysis. Prior to the start of the trial, PIs' educational degrees as well as indication and scale experiences were reviewed against pre-determined criteria.

Design (cont.)

All raters were required to have at least two years of schizophrenia population experience along with two years of diagnostic and efficacy scale-specific experience. Once approved, raters received didactic scale training by expert clinicians either at the Investigators' Meeting or via training web portal. For two common clinical trial schizophrenia scales (PANSS and SANS) didactic training was followed by a scoring exercise of a video-taped mock interview as well as an applied skills exercise which required the rater to conduct a live-interview with a standardized patient (i.e., an actor/actress mimicking targeted behavior and psychopathology relevant to the trial). Remediation was provided to raters whose scores did not agree (< 80%) of the established gold standard video scores and whose applied interviews did not meet minimum quality standards on the following domains: rapport, instrument comprehension, line 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 required to electronically submit source documentation, as administered by the PIs, for each screening and baseline visit throughout the study. As part of the data surveillance program, expert clinicians provided detailed written feedback via email to the PIs for each screening and baseline submission, noting each separate error with reminders of scale and protocol-specific conventions. PIs were also informed (documented) when no issues or errors were identified during the course of the review. This methodology ensured that the raters included in this analysis received feedback directly and had the opportunity to apply the feedback in subsequent visits.

The development of the current study was conceived after the international clinical trial terminated; as such, the independent expert clinicians within WCT[1] who participated in the data surveillance program did not know about the present investigation's analysis and were therefore not biased when providing feedback. A group of expert clinicians reviewed the types of errors identified in the data surveillance program and the errors were categorized as:

1. **Scale Administration Errors-** the rater did not follow administration conventions as defined by scale-specific instructions
2. **Scale/Symptom Coding Errors-** the rater incorrectly coded a symptom
3. **Diagnostic errors-** the rater coded an incorrect diagnosis
4. **Protocol adherence errors-** specific procedures were not properly followed, per protocol
5. **Missing or Insufficient Source Notes-** the rater did not provide enough source notes to support the code or score given

Results

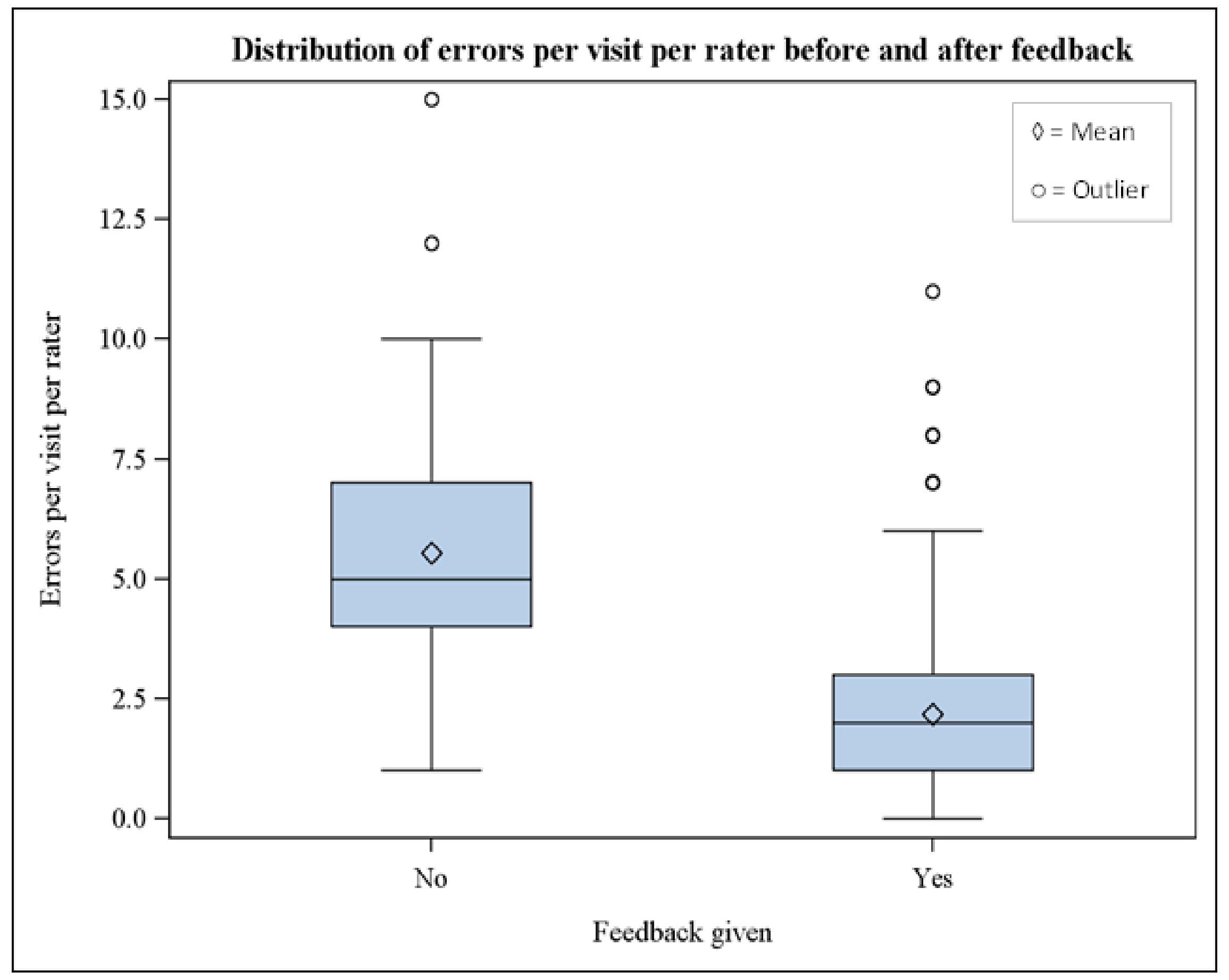
Data were derived from 27 PIs across 27 centers in 137 patients over 217 visits. Of the 27 raters, 26 received feedback during the study period. Visits conducted by PIs accounted for 51% of the total number of screening and baseline visits reviewed in the study. The majority of raters who received feedback did so after their first visit (screening visit; 16/26 raters), and all 26 raters had received feedback by their seventh visit.

Visits were categorized as having occurred before or after feedback started. The median number of visits per rater was 6 (1 pre-feedback, 5 post-feedback). The authors of this poster modelled the number of errors per visit per rater as a function of corrective feedback received [Yes/No] and number of visits. Due to deviation from normality, the number of errors was first log-transformed as: log(x+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 (SD=2.2) before feedback to 2 (SD=1.5) after feedback (Figure 1). The mixed model Least-Squares (LS) means estimates for the number of errors per rater across all visits was 4.0 [95% CI, 2.7, 5.8] before feedback, and 1.2 [1.0, 1.5] after feedback, representing a statistically significant reduction of 2.8 [1.7, 4.3] errors per visit 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.

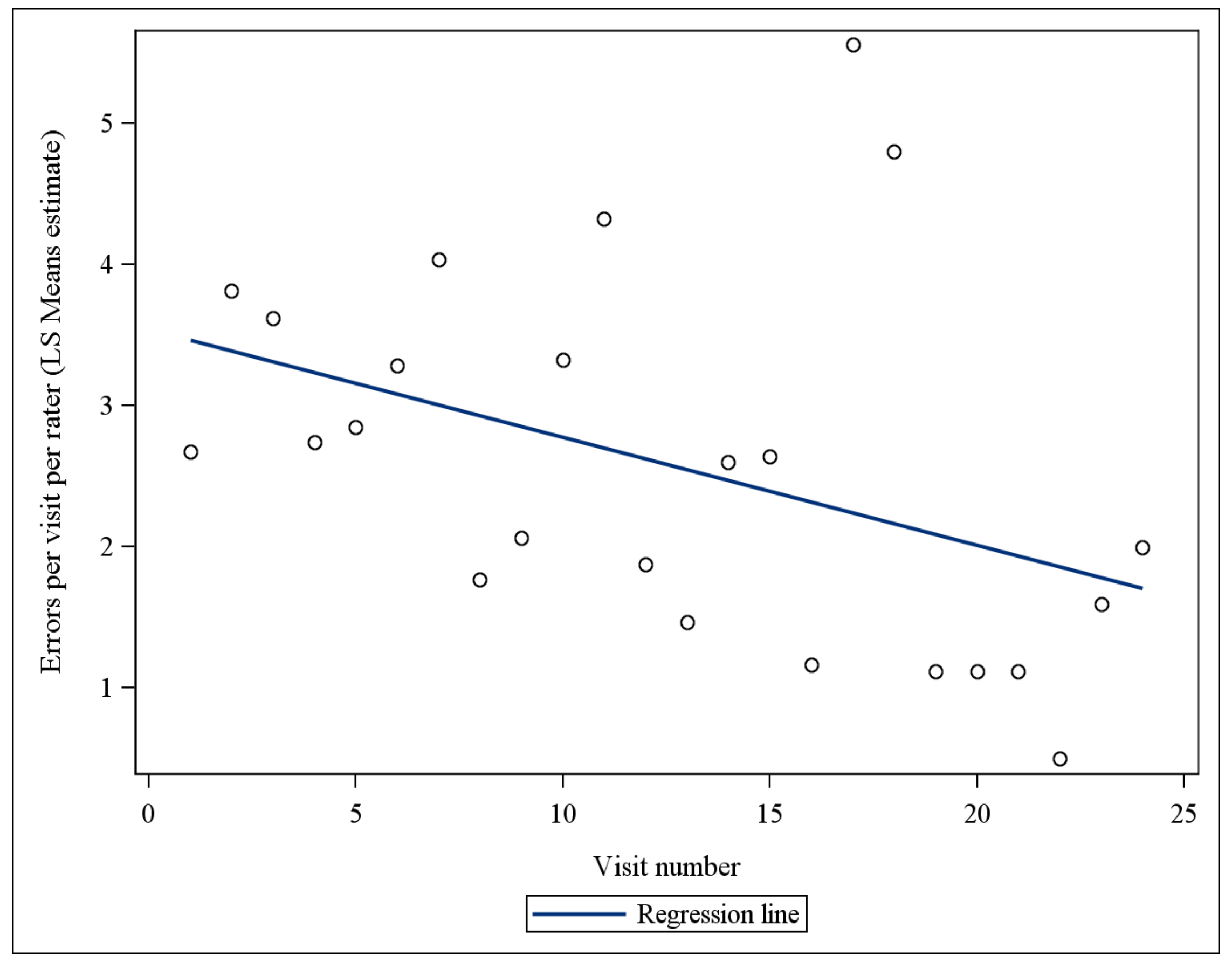
Results (cont.)

Figure 1. Distribution of errors per visit per rater (raw data) before and after feedback.



Number of errors: Before feedback mean=5, After feedback= 2, p <0.0001

Figure 2. Errors per visit per rater over number of visits based on mixed model (LS means) estimates.



Mixed Model Estimates

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

Effect	NumDF	DenDF	FValue	ProbF
Visit number	23	58	1.71	0.0505
Feedback	1	58	28.35	<.0001

Results (cont.)

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

	Category	ESTIMATE	95% CI	P-value
LS Means				
	Errors before feedback	4.0	(2.7, 5.8)	
	Errors post-feedback	1.2	(1.0, 1.5)	
Diff. in LS Means	Errors before feedback – errors after feedback	2.8	(1.7, 4.3)	<.0001

Conclusion

The results of this study indicate that fully-certified and experienced raters still make errors on crucial study assessments, indicating that experience and certification are not enough to ensure accuracy in psychometric evaluation in clinical research. However, the role of independent data surveillance and monitoring demonstrates in the current study that raters significantly improved with ongoing constructive feedback. Improvement was related to "time in study" while receiving constructive feedback. Busner et al.'s (2011) also demonstrated that ongoing data surveillance and remediation efforts lead to verifiable results in enhancing rater accuracy. The current data adds to the literature by demonstrating that improvement does occur even among the most credentialed and experienced raters – PIs.

It is noteworthy that protocol adherence errors were among the types of errors monitored in this trial and included in this analysis. For example, protocol adherence errors included scoring the primary and secondary efficacy scales without interviewing an informant/caregiver when required to do so or failing to follow the order of assessments specified in the protocol. Getz et al. (2008) shed light onto the growing complexity of study protocols between 1999 and 2005, noting a 6.5% rate of growth in the number of unique procedures per protocol across all therapeutic areas and all phases of development. Thus, this study's inclusion of protocol adherence errors along with clinical scale errors may reflect the growing demands placed on study raters to marry clinical knowledge with awareness of protocol and population-specific requirements.

Also notable, the current study found that US and CEE raters alike scored more accurately following external feedback, regardless of differences in prior scale, research, and clinical experience. This finding supports Cohen et al.'s (2015) assertion that both US and CEE raters perform similarly through the life of the clinical trial, despite significant differences in years of scale experience. Given that between 1995 and 2005 the number of trial sites outside the U.S. more than doubled while the number of U.S. and Western Europe-based trial sites decreased (Glickman et al., 2009), the current and Cohen et al.'s investigation results are applicable as it pertains to the increased use of CEE sites in clinical trials.

As this sample of raters were part of an industry-sponsored trial reviewed retrospectively, it was not possible to include a control group of raters, free from ongoing intervention. One could argue, for instance, that the decrease in error rate found in the current study was due to not only the feedback itself but also the raters' knowledge and expectation of oversight (as in the observer, or Hawthorne Effect). However, analyses presented indicated that corrective feedback was an additive influence in significantly reducing rater error. Another potential limitation to the present investigation is that the analysis used a composite endpoint with no ability to separately examine the impact of corrective remediation on each of the contributing parameters. Regardless of the type of raters employed or the primary agent of change, the net result is clear and quantifiable in regard to rater improvement with continued feedback. The results of expert feedback on rater performance call for further study to specify the elements which determine best practices in standardized rater training and quality assurance methodology.

[1]Worldwide Clinical Trials (WCT)

References provided upon request.