14. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. *J Clin Epidemiol*. 1996;49:1215–1219.

Are the Effects of Rater Training Sustainable?

Results From a Multicenter Clinical Trial

To the Editors:

With the high rate of failed and negative trials,[1] the importance of the quality of the clinical assessments conducted in these trials has recently become the focus of much attention in the literature. Studies showing the relationship between good clinical interviews and signal detection illustrate the impact that poorly conducted interviews can have on study outcome.[2,3] Other studies have found that the most ratings in clinical trials, in general, are of fair or poor quality, demonstrating the need for intervention in this area.[3,4] Several rater training programs have recently been developed that have successfully improved raters' clinical skills before the study initiation.[5,6] Although these studies found that raters' clinical skills could be tested and improved before being certified to rate in a study, there has been no evidence, to date, on whether these skills are maintained during the course of a trial. The current study examined this issue in a multicenter depression trial.

Thirty-one raters from 15 sites went through a training and certification process on the Hamilton Rating Scale for Depression (HAMD) and Anxiety (HAMA)[7,8] before initiating the trial. All raters were required to have prior clinical experience with depressed patients and documented experience administering the HAMD, the primary outcome measure in the trial. Raters' education ranged from MDs and PhDs to BA- and MA-level clinicians.

The certification process consisted of 2 components, didactic training/testing and applied training/testing, and is described in detail elsewhere.[5,6] To be certified to rate patients in the trial, raters had to achieve passing scores on both components. The didactic component consisted of a Web-based tutorial on the HAMD, a handout on scoring conventions for the HAMA, and a 20-item multiple-choice posttest. The applied component consisted of live observation of the trainee conducting a HAMD and HAMA interview with a depressed patient via a 3-way teleconference. The trainer provided feedback to the rater on their interviewing technique and on their scoring rationale and rated the trainee's applied performance using the Rater Applied Performance Scale[9] (RAPS). Trainees who received a failing score on the RAPS scale (ie, mean score of fair or poor [see below]) were required to conduct another interview, incorporating the feedback given. Trainees were given a total of 3 opportunities to pass the applied component. If the trainee failed on the third interview, he or she was excluded from rating patients in the trial. The Structured Interview Guide for the HAMA and HAMD[10] was used in the training.

To evaluate whether trainees retained the skills learned during certification and to prevent rater drift, raters were required to be retested on their applied skills approximately halfway through the study (ie, roughly 12 months later). A similar assessment procedure was used as during their prestudy certification, that is, raters interviewed a depressed patient while being observed and evaluated by a trainer via 3-way teleconference. Raters who failed the midstudy evaluation were given feedback and additional opportunities to pass, as was done in the original training.

Results of the initial training are detailed elsewhere[6] and will be summarized here. A significant improvement was found after the initial training in both the trainees' didactic knowledge of scoring conventions and the trainees' applied clinical interviewing skills. On the initial applied training, 57% passed on their initial attempt (prior to any feedback), 30% passed on their second attempt, and 7% on their third attempt, and 7% failed all 3 attempts and were excluded from participating in the study. For persons who failed their first applied test in their initial training, their RAPS score improved significantly after feedback on their second attempt, from 9.05 to 11.58, $P = 0.001$. Similarly, for those who failed their second attempt, RAPS scores improved significantly after feedback on their third attempt, from 9.0 to 11.0, $P = 0.033$. The mean RAPS score for all raters who were certified to rate in the study on their final (passing) attempt was 12.22 (SD = 1.75) (maximum possible score = 16). A score of 14.5 to 16 is roughly equivalent to a rating of excellent performance for all RAPS dimensions, 10.5 to 14.4 is good, 6.5 to 10.4 is fair, and less than 6.5 is poor.

At midpoint, the mean RAPS score decreased significantly from their initial posttraining scores, from 12.22 to 10.68 (SD = 2.64), $t_{30} = 2.976$, $P = 0.006$. This change is clinically significant, as it represents a change from a solid to a borderline "good." The largest decreases were on the RAPS dimensions of neutrality (0.32 point) and follow-up (0.26 point). Eighteen raters (58%) passed the midpoint evaluation on the first try. Of the 13 raters (42%) who did not pass, the mean RAPS score significantly improved following feedback, from a mean of 7.82 on their initial attempt to a mean of 12.36 on their second attempt, $t_{10} = 5.590$, $P < 0.0001$. Two raters failed the second attempt, but passed on their third and final try, with the mean RAPS score improving from 9.5 on time 2 to 12.0 on time 3.

Results were also analyzed by educational degree. Raters with MD or PhD degrees (n = 12) did not have a significant drop in RAPS scores from prestudy certification to midpoint (12.25–11.50, $t_{11} = 0.828$, $P = 0.425$), whereas those with highest degree being MA or BA (n = 14) did have a significant drop (12.43–10.14, $t_{13} = 3.04$, $P = 0.009$). The mean drop in the latter group (2.29 points) was more than 3 times as large as the drop in the MD/PhD group (0.75 points).

DISCUSSION

The results indicate that although rater training can successfully improve raters' applied clinical skills, these skills can erode over the course of a trial. On the other hand, follow-up testing and training may be successful in re-establishing these skills. We did not do testing at study termination, and thus, we do not know if the reinforcement of the second intervention led to

maintenance of these skills for the duration of the trial. It also appears that level of education has an attenuating effect on this drop, with those having doctoral or medical degrees not showing a significant decline, whereas those with master's or bachelor's degrees did. It may be that this degradation of skills represents a slippage in raters for whom the skills are newer rather than a result of laxity over time.

These results suggest the need for follow-up testing and training in addition to training conducted before study initiation. Results also suggest there may be benefit to using more experienced raters to perform the outcome assessments in clinical trials. Given the shortage and expense of clinicians with MD or PhD degrees, alternative strategies may need to be considered. These include more intensive training and calibration up front, more intensive monitoring of rater performance, or the use of centralized raters. The significant impact that interview quality has on signal detection makes attention to the problem of rater quality critical for a successful clinical trial outcome. Finally, it should be noted that the methods used to evaluate the raters in this study (ie, observation of a full interviews in real time with ratings based on evaluations of specific clinical skills) are more rigorous than are generally used in most rater training programs. However, we feel this higher bar is justified, given the importance of these applied skills in clinical trial outcomes.

ACKNOWLEDGMENTS

This study was funded by a grant from GlaxoSmithKline. The authors thank Joseph Schmidt, who provided operational oversight for the study.

Disclosures: Dr Kobak is VP of Research, and Drs Engelhardt and Jeglic are Research Scientists at MedAvante, a company that provides rater training services. Dr Williams is VP of Clinical Development and Dr Lipsitz is consultant at MedAvante.

Kenneth A. Kobak, PhD*
Joshua Lipsitz, PhD*†
Janet B.W. Williams, DSW†
Nina Engelhardt, PhD*
Elizabeth Jeglic, PhD*‡
Kevin M. Bellew, MS§

*MedAvante, Inc
Princeton, NJ
†New York State Psychiatric Institute
and Columbia University
‡John Jay College of Criminal Justice,
New York, NY
and §Neurosciences Medicine
Development Center, GlaxoSmithKline
King of Prussia, PA
kkobak@Medavante.net

REFERENCES

1. Khan A, Leventhal RM, Khan SR, et al. Severity of depression and response to antidepressants and placebo: an analysis of the food and drug administration database. *J Clin Psychopharmacol*. 2002;22(1):40–45.
2. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *Am J Psychiatry*. 2005;162(3):628.
3. Feiger A, Engelhardt N, DeBrota D, et al. Rating the raters: an evaluation of audiotaped Hamilton Depression Rating Scale (HAMD) interviews. Presented at the National Institute of Mental Health, New Clinical Drug Evaluation Unit 43rd Annual Meeting; 2003; Boca Raton, FL.
4. Engelhardt N, Feiger AD, Cogger KO, et al. Rating the raters: assessing the quality of Hamilton rating scale for depression clinical interviews in two industry-sponsored clinical drug trials. *J Clin Psychopharmacol*. 2006; 26(1):71–74.
5. Kobak KA, Engelhardt N, Lipsitz JD. Enriched rater training using Internet based technologies: a comparison to traditional rater training in a multi-site depression trial. *J Psychiatr Res*. 2006;40(3):192–199.
6. Kobak KA, Lipsitz JD, Williams JB, et al. A new approach to rater training and certification in a multicenter clinical trial. *J Clin Psychopharmacol*. 2005;25(5):407–412.
7. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56–62.
8. Hamilton M. The assessment of anxiety states by rating. *Br J Med Psychol*. 1959;32:50–55.
9. Lipsitz J, Kobak KA, Feiger A, et al. The Rater Applied Performance Scale (RAPS): development and reliability. *Psychiatry Res*. 2004;127:147–155.
10. William JBW. *Structured Interview Guide for the Hamilton Depression and Anxiety Scales (SIGH-AD)*. New York, NY: New York State Psychiatric Institute; 1996.

Comments on ''Why Do Clinical Trials Fail? The Problem of Measurement Error in Clinical Trials

Time to Test New Paradigms?''

To the Editors:

In their guest editorial, Kobak et al[1] address "the problem of measurement error in clinical trials" as a partial answer to the high rate of failed clinical trials, suggesting it is "time to test new paradigms." Such calls have been sounded before[2]; methods to control the measurement error inherent in rating scales that undermine the validity of obtained data and the extreme difficulty of addressing such problems through rater training have been recognized.[3]

We agree that poor interrater reliability pursuant to dubious interview quality and rater bias are grave concerns. Such methodological confounds are most insidious at critical choice points in clinical trials, such as subject selection and final assessments that can determine study outcomes.[4,5] We contend, however, that interrater unreliability due to inconsistent interviewing procedures that compromise assessment quality and introduce rater bias is a problem *specific* to the use of human raters. We further contend that the absolute procedural reliability of computer interviews that unwaveringly follow up interviewee responses in a consistent manner and interpret replies in an unbiased manner may be the *only* means to address the critical clinical trial assessment issues raised by the editorial.

Kobak et al[1] acknowledged the futility of trying to train dozens of raters across different sites, with varied and diverse backgrounds, and hoping the imparted assessment skills will be equally effective for improving assessment quality across the whole of the clinical trial. Reducing the number of individual raters to "8 to 10 centralized raters" simply acknowledges that no 2 raters will ever be internally consistent. Consequently, reducing the number of raters involved in a trial merely minimizes the unreliability of human raters by reducing the number of individual raters employed in any given study. The implication is that only raters selected, qualified, and credentialed by individxual centralized rater companies are eligible to participate in the "new paradigm" whose time to be tested has come.

The editorial by Kobak et al[1] says little about the similarities and/or differences between centralized ratings and computer-automated assessments. It