

## Site Versus Centralized Raters in a Clinical Depression Trial Impact on Patient Selection and Placebo Response

Kenneth A. Kobak, PhD,\* Andrew Leuchter, MD,† David DeBrot, MD,‡ Nina Engelhardt, PhD,\* Janet B.W. Williams, DSW,\*§ Ian A. Cook, MD,† Andrew C. Leon, PhD,|| and Jonathan Alpert, MD, PhD¶

**Abstract:** The use of centralized raters who are remotely linked to sites and interview patients via videoconferencing or teleconferencing has been suggested as a way to improve interrater reliability and interview quality. This study compared the effect of site-based and centralized ratings on patient selection and placebo response in subjects with major depressive disorder. Subjects in a 2-center placebo and active comparator controlled depression trial were interviewed twice at each of 3 time points: baseline, 1-week postbaseline, and end point—once by the site rater and once remotely via videoconference by a centralized rater. Raters were blind to each others' scores. A site-based score of greater than 17 on the 17-item Hamilton Depression Rating Scale (HDRS-17) was required for study entry. When examining all subjects entering the study, site-based raters' HDRS-17 scores were significantly higher than centralized raters' at baseline and postbaseline but not at end point. At baseline, 35% of subjects given an HDRS-17 total score of greater than 17 by a site rater were given an HDRS total score of lower than 17 by a centralized rater and would have been ineligible to enter the study if the centralized rater's score was used to determine study entry. The mean placebo change for site raters (7.52) was significantly greater than the mean placebo change for centralized raters (3.18,  $P < 0.001$ ). Twenty-eight percent were placebo responders (>50% reduction in HDRS) based on site ratings versus 14% for central ratings ( $P < 0.001$ ). When examining data only from those subjects whom site and centralized raters agreed were eligible for the study, there was no significant difference in the HDRS-17 scores. Findings suggest that the use of centralized raters could significantly change the study sample in a major depressive disorder trial and lead to significantly less change in mood ratings among those randomized to placebo.

**Key Words:** clinical trials, randomized, placebo effect, methods, outcomes assessments, patient, depressive disorder, Hamilton Depression Rating Scale

(*J Clin Psychopharmacol* 2010;30: 193–197)

Problems associated with clinical measurement have become the focus of increased attention as a possible contributor to the high rate of failed antidepressant clinical trials.<sup>1</sup> Poor interrater reliability has long been known to increase variance,

which reduces the between-group effect size, and as a result, decreases statistical power or requires a corresponding increase in sample size, study duration, and cost.<sup>2</sup> Rater bias, in the form of baseline score inflation<sup>3</sup> and expectancy effects<sup>4</sup> (the tendency to see improvement over time), can increase placebo response and decrease signal detection, as higher pretreatment depression scores are associated with greater change with antidepressants, whereas lower baseline scores are associated with greater change with placebo<sup>1</sup> (ie, inflation results in overinclusion of subjects whose true scores are lower in severity and thus more likely to show a placebo response). More recently, interview quality, that is, the skill of the clinician administering the scale, has been found to significantly affect signal detection.<sup>5</sup>

One methodological approach to addressing these problems is to use blinded centralized raters to administer the screening and outcome measures in psychiatric clinical trials. *Centralized raters* refers to a relatively small group of highly skilled and tightly calibrated raters who are independent from the study sites. They are linked to the various study sites through videoconferencing or teleconferencing and remotely administer the screening and/or primary outcome measure(s) to study subjects during their regularly scheduled study visits.

Centralizing raters could improve reliability of clinical measurements by reducing the number of raters involved, enhancing the training and experience of the cadre of raters, and through blinding of the raters to inclusion criteria and study visit number. Because centralized raters are independent from the study site, they also are protected from the pressure to enroll subjects, thus helping to ensure a truly independent evaluation at each study visit. The use of different raters at each visit (chosen from a pool of interchangeable centralized raters) may also minimize the potentially psychotherapeutic impact of repeated assessment, which was found in one recent study to account for approximately 40% of the placebo response.<sup>6</sup> Several studies have found lower placebo response and better signal detection when using different raters at baseline and end point.<sup>4,7</sup>

The current study evaluated the use of centralized raters in an investigator-initiated 2-site depression study. The primary objective of that study was to examine predictors and indicators of response to placebo treatment in patients with major depressive disorder (MDD). The comparison of site-based versus centralized raters was added as a secondary objective to compare the performance of 2 assessment methods. The objective of this report was to compare those 2 methods on 2 critical aspects of randomized controlled trial (RCT) implementation: (1) selection of subjects deemed eligible for inclusion into the study and (2) magnitude of placebo response, as measured both by mean change from baseline to end point and percent responders among those randomized to the placebo cell. We hypothesize that relative to site raters, (1) centralized raters would have a lower subject inclusion rate and (2) central raters would show lower change from baseline to end point for subjects randomized to placebo.

From the \*MedAvante, Inc, Madison, WI; †Semel Institute for Neuroscience and Human Behavior and Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA; ‡Eli Lilly & Co, Indianapolis, IN; §Department of Psychiatry, Columbia University; ||Weill Medical College of Cornell University, New York, NY; and ¶Harvard University and Massachusetts General Hospital, Boston, MA. Received May 5, 2009; accepted after revision December 30, 2009.

Reprints: Kenneth A. Kobak, PhD, MedAvante, Inc, 7601 Ganser Way, Madison, WI 53719 (e-mail: kkobak@medavante.com).

This study was cofunded by grants from Eli Lilly & Co and Pfizer, Inc. Copyright © 2010 by Lippincott Williams & Wilkins

ISSN: 0271-0749

DOI: 10.1097/JCP.0b013e3181d20912

## METHOD

Eighty-one subjects with MDD were recruited via newspaper ads at 2 sites Semel Institute for Neuroscience and Human Behavior at UCLA, Laboratory of Brain, Behavior, and Pharmacology ( $n = 39$ ) and Massachusetts General Hospital, Department of Psychiatry ( $n = 42$ ). Subjects were screened and consented at visit 1. At visit 2 (baseline), subjects were administered several scales to assess mood and symptoms of depression. If eligible, subjects entered a 1-week single-blind placebo lead-in phase. Because this was a study of placebo response, subjects were not excluded because of placebo response during the single-blind placebo lead-in. This placebo lead-in was used primarily to be consistent with other studies to manage expectancy bias in subjects. After the placebo lead-in, those subjects with normal laboratory test results at visit 3 were randomly assigned to 5 weeks of treatment with either sertraline (up to 150 mg daily) or placebo, with a 3:1 randomization allocation ratio favoring placebo. The enlargement of the placebo group minimized the total number of depressed subjects enrolled in the trial, yet allowed testing the parent study hypotheses regarding prediction of placebo response.

All subjects were interviewed twice with the 17-item Hamilton Depression Scale (HDRS-17) at each of 3 time points: baseline (visit 2, day 0), visit 3 (day 7), and end point (visit 8, day 35)—once by the site rater and once remotely via video-conference by a centralized rater. Subjects were also rated by the site raters on the intervening visits between baseline and end point. The central rater was blind to study visit, design, and inclusion criteria. Administrative staff acted as an interface between the raters and the remote sites and patients to maintain blinding. Site raters were blind to many aspects of the study design, but they were aware of the subject's overall history, that an HDRS-17 score of 17 or greater was necessary for study entry, of the study visit number, and that this was primarily a study of placebo response (although they were not aware of the randomization ratio). Site and central raters were blind to each others' scores. A counterbalanced order for site and central assessments was used only at visit 3 and at end point; site raters went first at baseline. All raters used the Structured Interview Guide for HDRS.<sup>8</sup> Subjects had to have a site-based HDRS-17 score of 17 or greater at baseline to be entered in the study (centralized raters scores were not used for study inclusion or exclusion). There were 8 centralized raters (3 men and 5 women, 7 with a doctoral degree and 1 with a masters degree) and 13 site raters (7 men and 6 women, 10 with an MD, 1 with a PhD, 1 with a MA, and 1 with a BA). Centralized assessments were per-

formed by unique raters at each visit, although the sites generally used the same rater at each visit for each subject.

Remote centralized raters were connected to the on-site subjects via a Polycom videoconferencing unit connected over Integrated Services Digital Network telephone lines at an industry standard rate of 384 kbps. There was no connection to or access from any other device, including the Internet. To establish connection, a site entered the phone number on the remote system, and the connection was established. The connection was secure and encrypted to the same standards required by the US Department of Defense.

## Data Analytic Procedures

The paired ratings resulting from the 2 assessment methods were compared using Wilcoxon rank sum tests. This includes analyses of cross-sectional assessments and of pre-to-post change. Comparison of percent responders was calculated using the McNemar test, which assesses the significance of the difference between 2 correlated proportions, as is the case when the proportions are based on the same sample of subjects or on matched-pairs samples. Internal consistency reliability was evaluated using Cronbach coefficient  $\alpha$ . Cronbach  $\alpha$  measures how well a set of items in a scale hang together, that is, measure a single unidimensional latent construct. Cronbach  $\alpha$  is a common test of whether items are sufficiently interrelated to justify their combination in an index and will generally increase when the correlations between the items increase. Each inferential statistical test used a 2-tailed  $\alpha$  level of 0.05. Owing to the randomization allocation ratio of 3:1, only 16 subjects were randomized to drug (12 subjects if using centralized raters for eligibility). Because of small sample size of medication-treated subjects, those data are not examined in analyses of pre-post change; instead, those analyses focused on the subjects randomized to placebo.

## RESULTS

### Study Flow

Eighty-one subjects who had been prescreened at visit 1 were continued to visit 2 (baseline). At visit 2, these 81 subjects were examined, and 66 subjects were found eligible and started on single-blind placebo. At visit 3, 62 of these 66 subjects were reexamined (4 dropped between visit 2 and visit 3) and randomized to drug or placebo. Of these 62 subjects, 51 completed the study (39 on placebo and 12 on active drug) and had end points for both central and site raters.

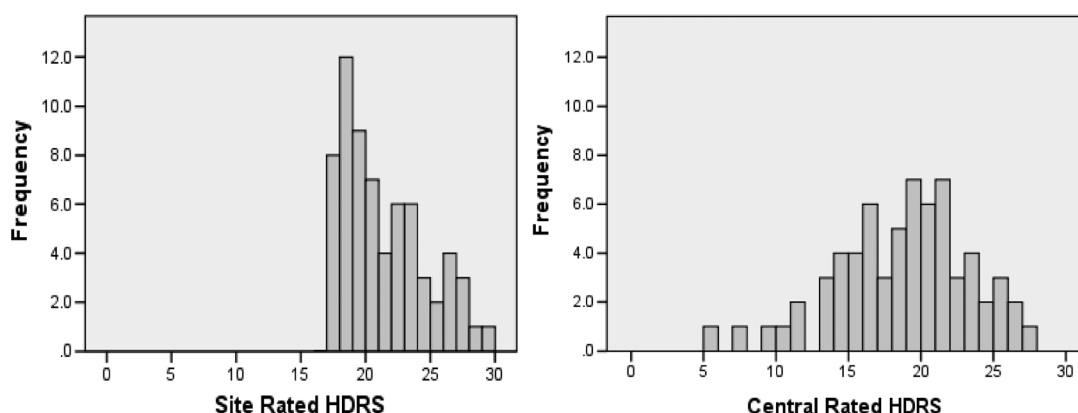


FIGURE 1. Frequency distributions of site and central HDRS ratings at baseline for site eligible subjects.

## Agreement Between Assessment Modalities

Site HDRS ratings were significantly higher than centralized ratings at baseline (visit 2) and visit 3 but not at end point. At baseline (visit 2), there was a 2.67 difference between site and centralized raters (mean [SD], 20.92 [3.25] vs 18.26 [4.64],  $Z = 4.621$ ,  $P < 0.001$ ); at visit 3, the difference was 2.91 points (18.90 [4.69] vs 17.24 [6.31],  $Z = 2.917$ ,  $P = 0.004$ ); however, at end point, the mean difference was less than 1 point (0.98; 13.08 [6.58] vs 14.06 [7.14],  $Z = 1.665$ ,  $P = 0.096$ ). At visit 2, 65% of site raters' scores were higher than centralized rater scores; at end point, only 35% were higher. Intraclass correlation coefficients between site and centralized ratings at baseline (visit 2) and visit 3 were 0.34 and 0.48, respectively and improved to 0.71 at end point.

At baseline (visit 2), 35% of subjects given a HDRS total score of 17 or greater by a site rater were given a HDRS total score of lower than 17 by a centralized rater and, thus, would have been ineligible to enter the single-blind placebo phase if the centralized rater's score was used to determine study entry (Fig. 1).

## Internal Consistency Reliability

The centralized ratings had significantly higher internal consistency reliability at baseline (including all subjects screened) and visit 3, and at end point, both central raters' and site raters' internal consistencies were high—centralized ratings: 0.67 (95% confidence interval [CI], 0.551–0.768; baseline), 0.79 (95% CI, 0.714–0.863; visit 3), and 0.83 (95% CI, 0.766–0.894; end point) and site ratings: 0.375 (95% CI, 0.156–0.570; baseline), 0.52 (95% CI, 0.350–0.692; visit 3), and 0.78 (95% CI, 0.697–0.859; end point; Fig. 2).

## Placebo Response Between Modalities

The mean (SD) change (baseline to end point) on placebo for site raters (7.26 [5.99]) was significantly greater than the mean (SD) change on placebo for centralized raters (3.18, [6.60], Wilcoxon  $Z = 3.872$ ,  $P < 0.001$ ). When analyses are limited to only those subjects who would have been eligible based on centralized ratings ( $n = 25$ ), the mean (SD) change on placebo for site raters (7.52 [7.52]) was numerically greater than the mean (SD) change on placebo for centralized raters but no longer was statistically significant (5.24 [6.99],  $Z = 2.184$ ,  $P = 0.09$ ). (Study power is greatly reduced when reducing the sample size in half, ie, 51–25).

Using the a priori definition of responder status of 50% or greater change on the HDRS-17, 28% (11/39) were classified as placebo responders based on site ratings, whereas 14% (7/39) were responders based on centralized ratings ( $P < 0.001$ ; 95%

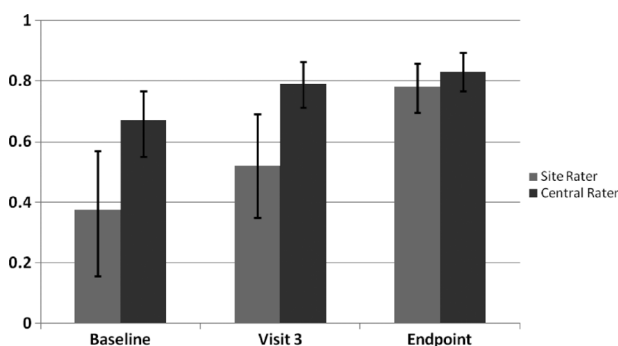


FIGURE 2. Internal consistency reliability (coefficient  $\alpha$ ) for site and central raters by visit.

CI, 9.157–1.747). When analyses are restricted to those subjects who would have been eligible to be randomized based on centralized ratings, 36% (9/25) were placebo responders by site raters and 24% (6/25) by centralized raters ( $P = 0.052$ ; 95% CI, 6.815–1.044).

It has been suggested that HDRS-17 rating interviews may sometimes be quite brief.<sup>9</sup> In this study, durations of HDRS-17 interviews by site raters were not measured, but durations of HDRS-17 interviews by central raters were measured and were quite consistent across visits (35.6, 33.2, and 32.3 minutes for baseline, visit 3, and end point, respectively).

## DISCUSSION

The objective of this study was to compare 2 assessment modalities in patient selection and magnitude of placebo response. The results indicate that site and centralized ratings had limited concordance. Use of centralized raters who were blinded to study design and visit would have resulted in a smaller number of study subjects than those enrolled based on site ratings. Blinded centralized raters generally scored depression severity lower at baseline, but the ratings coalesced with site raters' scores at end point. This mirrors what has been found when patient self-ratings were compared with site-based ratings in clinical trials<sup>3</sup> and thus seems not to be solely attributable to differences between self-report and clinician report. Interestingly, at end point, there was a greater, though nonsignificant, percentage of ratings where the centralized raters scored higher than the site raters, the opposite of what was found at baseline and visit 3. One interpretation of this finding is that expectation bias caused site raters, aware of study visit, to rate higher severity at the beginning, and lower severity at the end of the trial, whereas raters blinded to study visit and entrance criteria would not be affected in this manner.

Also of interest was the comparison of internal consistency reliability by visit and rating method. Although some attenuation of internal consistency reliability is expected at screen and baseline because of truncation of severity due to exclusion criteria, central raters had significantly higher internal consistency than site raters at baseline (0.67 vs 0.38) and visit 3 (0.79 vs 0.52) but not at end point (0.83 vs 0.78). This suggests that the individual items for the site-based ratings did not hold together as well, that is, did not seem to all vary in the same manner, at baseline as they did at end point. Centralized raters' internal consistency was more stable across the visits. There are several possible explanations for this finding. One is that study subjects, who are highly motivated to enter a study, may augment their reports of symptoms in an in-person interview to enhance the likelihood of their being enrolled. Other explanations include the possibilities that study coordinators, either to help so as not to disappoint subjects or to help meet enrollment goals, increase their ratings of various items, a phenomenon that has been observed in other studies. All of these possibilities are speculative, and it is not possible to determine the source of the ratings discrepancy in these subjects.

The mean change on placebo was significantly smaller for centralized raters than that for site raters. This finding was statistically significant only if the site raters' scores were used to determine inclusion criteria, suggesting that the discrepancy in ratings affects primarily a subset of subjects screened for a trial in MDD. It is possible that these subjects have certain common clinical characteristics or that the nature of the interaction between these subjects and either in-person or off-site raters affects results. These possibilities should be examined in future studies. Because site raters were aware that this was a study of placebo response and thus a greater percentage of subjects were likely to be on placebo, one might anticipate that site

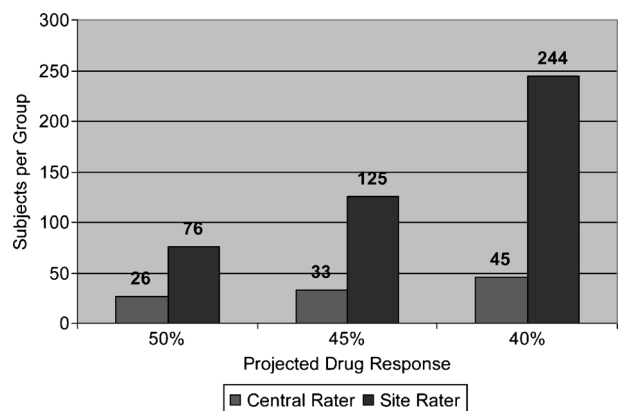
raters would show a lower placebo response than central raters, who were unaware of the study's purpose. However, the opposite result was found. Another possible explanation for the difference might be that centralized raters tended to rate generally lower than site raters. However, the fact that site and central raters coalesced at end point argues against this possibility.

There are several other possible reasons why the central raters may have rated subjects as significantly less depressed at baseline than the central raters. First, the site raters had more involvement in execution of the protocol than did the central raters. They usually obtained the consent from the subject and participated in the subjects' management, including dispensing of study compound. It is possible that if the raters solely performed ratings, their assessments could have been different. Second, the site raters had more information regarding the protocol than did the central raters. Specifically, the site raters read the consent form and therefore knew more about the study procedures, the minimum HDRS-17 score for entry into the study, and that there was a higher than usual likelihood of the subjects receiving placebo. Had the site raters been blinded to all details of the design, including entry criteria, this might have altered baseline ratings. Note that subjects, who may also be influenced by expectation bias if they think they have a high chance of receiving active treatment, were blinded to the randomization ratio. Third, it is possible that site raters' ratings on the HDRS-17 reflected information from the subjects that the central raters did not possess. Because site raters commonly performed telephone screening and scheduling, met the subjects upon arrival at the site, and escorted them to the interview room, they may have received verbal or nonverbal information from the subjects that the central raters did not. Fourth, the subjects may have been more likely to communicate to the site raters the severity of their depression as a function of spending more time with them or because the subjects were in acute distress and wanted help. Fifth, the baseline ratings were the only ones not counterbalanced for order (the site raters always occurred first). Thus, order may have affected severity, although sites were also higher the following visit when order was counterbalanced. Sixth, some have hypothesized that the nature of in-person communication is more intimate than communication through a camera and a television monitor, and thus, patients may self-disclose more in face-to-face communication. On the other hand, as discussed earlier, several studies support the equivalence of scores found by the 2 methodologies, and some have found that remote administration may actually increase disclosure of information of a sensitive nature, such as sexual behavior and suicidal ideation. It would be useful to examine empirically the impact of remote administration alone (vs blinding and bias) by altering the interaction of the subjects and the site raters to make the experience of site and central raters with subjects more similar. For example, site raters could be totally blinded to the protocol and have no interaction with subjects other than the rating session. If this led to greater concordance between raters, then this would help clarify the source of the higher site ratings. In practical terms, however, most sites would be unable to provide complete blinding, as it would require a different rater at each visit and, thus, a staff of, for example, 12 raters for a 12-visit study. Finally, the site raters were primarily MDs, whereas the central raters were primarily PhDs. The extent to which this may have impacted differences between site and central ratings is unknown. There has been little research on the impact of professional degree and clinical trial ratings. One study found that both MDs and PhDs had significantly higher clinical skills versus raters with other degrees<sup>10</sup> and that they were more likely to retain these skills throughout the trial.<sup>11</sup> Another study found

that rater competency was more a function of amount of training received versus years of experience per se.<sup>12</sup>

If the central raters' judgments are deemed to be more accurate than those of the site raters, it would be important to consider the potential impact of such a reduction in placebo response on sample size requirements for power of 80% in the design of future RCTs. If, for example, a drug-administered population's drug response rate of 50% was hypothesized, a reduction in the placebo population response rate from 28% to 14% (as was found in this study) would reduce the sample size required from 76 to 26 subjects (per cell), more than a 50% reduction (Fig. 3). Alternatively, assuming a hypothesized population drug response rate of 45%, a reduction in the placebo response rate from 28% to 14% would amount to a corresponding reduction in sample size required for power of 80% from 125 to 33 subjects per arm. At a cost of \$15,000 per subject, this could save several million dollars and significantly reduce trial duration. Conversely, a number of the subjects who were enrolled in this study would have been excluded (ie, a higher screen fail rate) through reliance on central raters. Exclusion of a significant proportion of subjects would lead to increased need for subject screening and unknown potential costs of increased advertising and trial duration. However, this screen fail rate may eventually diminish as sites learn new methods to recruit subjects who meet severity criteria. Another possibility is that lower centrally rated baseline scores are equivalent to higher site-rated baseline scores, and thus, a lower inclusion threshold may be used with central raters; however, this has not been empirically determined. Any projected cost savings must take all these factors into account in a comprehensive economic model. If the reduction in placebo response rates from centralized raters that was seen in the current study is replicated in future RCTs, the cost savings could be significant. Conversely, if the central rater judgments are not more accurate than those of the site raters, then a number of eligible subjects would have been excluded from the study for invalid reasons.

In the current study, there were almost as many central raters ( $n = 8$ ) as site raters ( $n = 13$ ). This is quite different from what would occur in a typical multicenter study with 10 or more investigative sites, where the number of site raters would be significantly greater than central raters (eg, a 30-site multicenter trial that used 60–75 raters [ie, 2 or 3 raters per site] could be conducted with 8–10 centralized raters). Reducing the sheer number of raters should in itself improve reliability and makes



**FIGURE 3.** Differences in sample size requirements as a function of population placebo and drug response rates. Based central rater placebo response rate of 14% and site rater placebo response rate of 28%. These calculations assume 0.80 power, a 0.05  $\alpha$  level, and a 2-tailed test between 2 independent groups.

the current findings even more compelling. One goal of the current study was to compare site and central raters on selection of subjects deemed eligible for inclusion. When there is disagreement on scores, the authors do not claim to know the right score. It may be that both raters are wrong and that the correct rating is in between the 2, higher than both or lower than both. Whether differences between site and central raters are due to factors related to rater behavior or factors associated with patient behavior (eg, subjects acting differently when interviewed by videoconference vs when interviewed in person) cannot be determined from the present study design. However, prior studies generally found high concordance between video and face-to-face ratings.<sup>13,14</sup>

There are limitations to the results presented here. First, the sample size was small. Second, by design, all subjects were rated twice at each point in time. Despite the overall counterbalanced design, this duplication could have bearing on the ratings. Future studies should be conducted that randomize subjects to both treatment and assessment modality. Future studies are also needed to examine whether change by centralized raters is also smaller (or larger) on active treatment and, more importantly, what the relative degree of change on drug and placebo is found by each rating method (site vs centralized).

#### AUTHOR DISCLOSURE INFORMATION

*Dr Kobak is a consultant to MedAvante, Inc; Drs Williams and Engelhardt are employees of MedAvante, Inc; and Dr Leon is on the scientific advisory board of MedAvante, Inc. Dr Leon has served as an investigator for research funded by the National Institute of Mental Health (NIMH) and the National Institute of Drug Abuse. He has served on data safety monitoring boards for AstraZeneca, Daiippon Sumitomo Pharma America, and Pfizer; and served as a consultant to the Food and Drug Administration, NIMH, Cyberonics, MedAvante, and Takeda. He has equity in MedAvante.*

*Dr Alpert received research support from Abbott Laboratories; Alkermes; Lichtwer Pharma GmbH; Lorex Pharmaceuticals; Aspect Medical Systems; AstraZeneca; Bristol-Myers Squibb Company; Cephalon; Cyberonics; Eli Lilly & Company; Forest Pharmaceuticals Inc; GlaxoSmithKline; J & J Pharmaceuticals; Novartis; Organon Inc; PamLab, LLC; Pfizer Inc; Pharmavite; Roche; Sanofi/Synthelabo; Solvay Pharmaceuticals, Inc; and Wyeth-Ayerst Laboratories. He participated on advisory boards for or consulted to Eli Lilly & Company; PamLab, LLC; and Pharmavite LLC. He received speakers' honoraria from Eli Lilly & Company; Janssen; Organon; and Reed Medical Education. He declares no relevant equity holdings, patents, or royalties.*

*Andrew Leuchter, MD, has provided scientific consultation or served on advisory boards for Aspect Medical Systems, Bristol-Myers Squibb, Eli Lilly and Company, Merck & Co, Otsuka Pharmaceuticals, and Pfizer. He has served on a speaker's bureau for Bristol-Myers Squibb, Eli Lilly and Company, Otsuka Pharmaceuticals, and Wyeth-Ayerst Pharmaceuticals. He has received research/grant support from the National Institute of Mental Health, the National Center for Complementary and Alternative Medicine, Aspect Medical Systems, Eli Lilly and Company, Wyeth-Ayerst Pharmaceuticals, Merck & Co, Pfizer,*

*Sepracor, Vivometrics, and MedAvante. He also is a former equity shareholder in Aspect Medical Systems.*

*Dr DeBrotta is an employee of Eli Lilly & Company.*

*Dr Cook (in the past 12 months) had research grants from NIH, Sepracor, Neuronetics, and Aspect Medical Systems; speaking honoraria from Neuronetics; consultation/advisory boards for Bristol-Myers Squibb; patents on biomedical devices/methods are assigned to the University of California, and he received no royalty income from them.*

#### 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; 22:40–45.
2. Leon AC, Marzak PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Arch Gen Psychiatry.* 1995;52(10): 867–871.
3. DeBrotta D, Demitrack M, Landin R, et al. A comparison between interactive voice response system-administered HAM-D and clinician-administered HAM-D in patients with major depressive episode. NCDEU 39th Annual Meeting, Boca Raton, FL; 1999.
4. Kobak KA, Lipsitz JD, Feiger AD, et al. Same versus different raters and rater quality in a clinical trial of major depressive disorder: impact on placebo response and signal detection. American College of Neuropsychopharmacology, 48th Annual Meeting, Hollywood, FL; 2009.
5. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *Am J Psychiatry.* 2005;162(3):628.
6. Posternak MA, Zimmerman M. The therapeutic effect of follow-up assessments on the placebo response in antidepressant efficacy trials. American Psychiatric Association, 158th Annual Meeting, Atlanta, GA; 2005.
7. DeBrotta D, Gelwicks S, Potter W. Same rater versus different raters in depression clinical trials. 42nd Annual Meeting, New Clinical Drug Evaluation Unit, Boca Raton, FL; 2002.
8. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry.* 1988;45:742–747.
9. 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.
10. 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.
11. Kobak K, Lipsitz J, Williams JBW, et al. Are the effects of rater training sustainable? Results from a multi-center trial. *J Clin Psychopharmacol.* 2007;27(5):534–535.
12. Targum SD. Evaluating rater competency for CNS clinical trials. *J Clin Psychopharmacol.* 2006;26(3):308–310.
13. Kobak KA. A comparison of face-to-face and videoconference administration of the Hamilton Depression Rating Scale. *J Telemed Telecare.* 2004;10(4):231–235.
14. Hyler SE, Gangure DP, Batchelder ST. Can telepsychiatry replace in-person psychiatric assessments? A review and meta-analysis of comparison studies. *CNS Spectr.* 2005;10(5):403–413.