

# Placebo Response Assessed by Site and Blinded Centralized Raters in a GAD Trial

## ABSTRACT

**Background:** The percent of anxiety trials in which an FDA-approved active comparator has failed to show a signal has exceeded 50%. Lack of standardization across sites and raters, poor inter-rater reliability, and possible scoring bias affecting the primary outcome measure contribute to this problem. The use of remote centralized raters may standardize assessments across raters and eliminate scoring bias. Remote raters can be blinded to protocol inclusion and exclusion criteria as well as visit number. These measures are aimed at decreasing placebo response thereby increasing signal detection. In this study of an experimental anxiolytic, site ratings were compared with remote centralized ratings. Although neither method was able to detect a difference between drug and placebo, the rate of placebo response assessed by the two rating methodologies can be compared.

**Methods:** This double-blind, randomized, placebo-controlled, multi-center study examined the efficacy and safety of two doses of an experimental compound to treat Generalized Anxiety Disorder (GAD). No active comparator was included. Site raters assessed randomized subjects 6 times over an 8 week period. The primary outcome measure was the Week 8 site rated HAM-A. In addition to site ratings, remote centralized raters independently rated subjects on the HAM-A at baseline and week 6; all raters used the SIGH-A.

The site raters randomized subjects based on MINI confirmation of a GAD diagnosis, and baseline HAM-A total score of  $\geq 20$  and a score of  $\geq 2$  on HAM-A items 1 (anxious mood) and 2 (tension). Remote raters' HAM-A evaluations followed site raters' evaluations at the baseline visit, and were counterbalanced by order with the site raters' evaluations at the 6-week visit. Site raters evaluated subjects in person, and remote raters' evaluations were conducted by telephone. There were 22 remote raters and 119 site raters across 45 sites. Remote centralized raters all received didactic and applied training, were calibrated with each other on scoring conventions and skill prior to beginning the study, and maintained high interrater reliability throughout the study. Site raters were trained at the investigator meeting with a traditional didactic session and group practice rating of a videotaped HAM-A interview.

**Results:** Site raters admitted 122 subjects to the placebo arm of the study. Of these, remote centralized raters would have admitted 59 (48%) and excluded 63 (52%), based on their HAM-A ratings. In addition, at baseline, site raters' mean item scores were all higher than the remote raters' mean item scores. The mean change from baseline in HAM-A total score in the placebo group admitted to the study by site raters was -9.3. This was significantly higher than the -5.9 point mean change on placebo as measured by the remote centralized raters in that same cohort. If one defines placebo response as a reduction in the HAM-A score of 50% or more, site raters identified 47 (39%) subjects as placebo responders, as compared to 29 (24%) subjects so identified by remote raters in the site-admitted cohort (nominal  $p=.015$ ).

At baseline, remote centralized raters' baseline scores were normally distributed despite subjects having already been qualified by the site ratings as having a HAM-A total score  $>20$ . At endpoint both sets of ratings appeared normally distributed.

The internal consistency reliability of the scale scores (Cronbach's alpha) was much higher for the remote centralized raters than the site-based ratings at baseline (0.78 vs. 0.18, respectively). At endpoint, however, the alphas were identical (0.84 for remote centralized raters and 0.84 for site raters). The mean score differences on the two required items of the HAM-A were the most significantly different of all the HAM-A items at baseline (mean difference site raters minus remote centralized raters = .41 and .44,  $t=12.46$  and  $t= 12.90$  respectively, both  $p < .001$ ). At endpoint, differences between site raters and remote centralized raters on these items were non-significant (mean difference = -.08 and .023 respectively).

**Discussion:** Site raters showed a significantly larger placebo response than remote centralized raters on subjects whom they had already rated higher at baseline. In addition, at screen and baseline the site raters' internal consistency was very low, especially on the two inclusion scale items, but rose to equal the remote centralized raters' alphas at endpoint.

These data are suggestive of and consistent with the potential for score inflation at baseline when rated by sites. Remote centralized raters may improve signal detection in clinical trials by decreasing placebo response as a result of appropriate baseline scoring.

Janet B.W. Williams<sup>1,3</sup> Judith Dunn<sup>5</sup> Kenneth A. Kobak<sup>1</sup> Earl Giller<sup>1</sup> Lisa Curry<sup>2</sup> Phebe Wilson<sup>2</sup> Michael Detke<sup>1,4</sup>  
<sup>1</sup>MedAvante Research Institute, Hamilton, NJ <sup>2</sup>Sepracor, Inc., Marlborough, MA <sup>3</sup>Columbia University College of Physicians and Surgeons <sup>4</sup>Indiana University School of Medicine; Harvard Medical School <sup>5</sup>Former Employee Sepracor Inc, Marlborough, MA

**Introduction:** The placebo response rate in anxiety disorder studies is growing, leading to an increasing concern regarding the rate of failed trials (Khan et al, 2005). Many possible explanations have been proposed, including lack of standardization across sites and raters, poor inter-rater reliability, and possible scoring biases affecting the primary outcome measure. Efforts to address this problem have included increasing the number of patients in a trial to achieve statistical significance, selectively picking sites that have performed well in previous studies, and increasing the quality of rater training at start-up meetings. Despite these efforts, however, the failure rate of trials remains high.

The use of remote centralized raters offer a solution that may address several identified problems. Remote centralized raters standardize and calibrate assessments across raters so that high inter-rater reliability can be initially obtained and continuously maintained, minimizing noise due to inconsistent assessments. In addition, remote raters have no incentive to enroll patients in a trial, and are blinded to protocol inclusion and exclusion criteria and visit number. Because they are remotely located, subjects are interviewed over videoconferencing or telephone and the amount of interaction with the subject is limited to the content of the assessment instrument. These measures are aimed at appropriate patient selection that could lead to decreasing placebo response by eliminating potential sources of bias such as enrollment bias (deBrotta et al, Feltner et al), therapeutic alliance bias (Posternak et al), and expectancy bias (Glaudin et al), and thereby increasing signal detection. This study of an experimental anxiolytic included both site ratings and remote centralized ratings. The rate of placebo response assessed by the two rating methodologies was compared.

**Method:** This double-blind, randomized, placebo-controlled, multi-center study examined the efficacy and safety of two doses of an experimental compound to treat Generalized Anxiety Disorder (GAD). No active comparator was included. Site raters assessed randomized subjects 6 times over an 8 week period. The primary outcome measure was the Week 8 site-rated Hamilton Anxiety Scale (HAM-A). In addition to site ratings, remote centralized raters independently rated subjects on the HAM-A at baseline and week 6. Both sets of raters used the Structured Interview Guide for the Hamilton Anxiety Scale (SIGH-A) (Williams, 1996). Altogether, there were 22 remote raters, and 119 site raters who qualified to rate across the 45 sites that randomized subjects.

The site raters randomized subjects based on MINI (Sheehan et al, 1998) confirmation of a GAD diagnosis, and screen and baseline HAM-A total scores of  $\geq 20$  as well as a score of  $\geq 2$  on HAM-A items 1 (anxious mood) and 2 (tension). Remote raters assessed only subjects who met these randomization criteria, as determined by the sites. Remote raters' HAM-A evaluations followed site raters' evaluations at the baseline visit, and were counterbalanced by order with the site raters' evaluations at the 6-week visit. Site raters evaluated subjects in person, and remote raters' evaluations were conducted by telephone. Studies have demonstrated the equivalence of face-to-face and telephone interviews for anxiety and depression scales (Kobak et al, 2008; Baer et al, 1995).

Site raters were required to have five years of clinical experience with GAD, and at least 2 years of experience using the HAM-A scale to participate in the rater qualification program conducted by United Biosource Corporation (UBC). The training and assessment component of the qualification program that took place during the investigators' meeting and during the in-study period consisted of both didactic and experiential training on the HAM-A. At the investigator meeting, potential raters viewed didactic presentations on the administration and scoring conventions of the HAM-A as well as a didactic presentation on interview skills, and had the opportunity to discuss any issues related to usage and administration of the scale. After this discussion, a videotaped patient assessment utilizing the HAM-A was shown to the potential raters for training and discussion purposes. After this discussion, a HAM-A qualification video was shown and potential raters' scores for qualification were collected.

An acceptable score range was established for each interview item. Those potential raters who performed within the acceptable range were considered to be "qualified", while potential raters who scored outside of the range were targeted for remediation and retesting. All raters who met the Sepracor experience and credential criteria and who successfully demonstrated adequate rating proficiency received a qualification memo and were able to participate in the study.

Remote centralized raters had a minimum of 2 years of clinical experience. All received extensive didactic and applied training on an individual basis, were calibrated with each other on scoring conventions and skill prior to beginning the study, and maintained high inter-rater reliability throughout the study with quarterly group calibrations and regular observations by trainers. By design, a different remote rater was used at each visit so that subjects were never interviewed by the same remote rater at both visits. This method eliminates expectancy bias that occurs with the expectation of improvement over time, potentially decreasing signal detection.

**Results:** There was no active comparator, and neither method was able to detect a significant difference between drug and placebo [data not shown]. However, the rate of placebo response assessed by the two rating methods can be compared. Site raters admitted 122 subjects to the placebo arm of the study. Of these, remote centralized raters would have admitted 59 (48%) and excluded 63 (52%), based on their HAM-A ratings. It should be noted that both of these distributions present scores on the same subjects. There was no significant age or gender difference between the subject groups admitted by the site raters and those admitted by the central raters.

Across all treatment groups, site raters' scores at baseline were truncated at the inclusion score of 20, as expected, since site raters were the gatekeepers for the study [Fig. 1]. Remote blinded centralized raters' baseline scores, however, were normally distributed despite subjects having already been qualified by the site ratings as having a HAM-A total score  $>20$ . At endpoint, however, site raters' and remote raters' scores were similarly distributed [Fig. 2]. At baseline, site raters' mean score on the HAM-A was 24.08 (SD, 3.508; N=437) as compared to remote raters' mean score, which was 19.77 (SD, 5.99; N=437). In addition, site raters' mean scores were higher than remote raters' on all of the individual HAM-A items at baseline. At endpoint, however, remote raters' scores were similar to site raters' (14.03 vs. 14.58), with several items scored higher than site raters in this placebo arm. The mean length of interviews of the remote centralized raters was 37 minutes at baseline and 32 minutes at week 6. Site raters' visit lengths were not tracked.

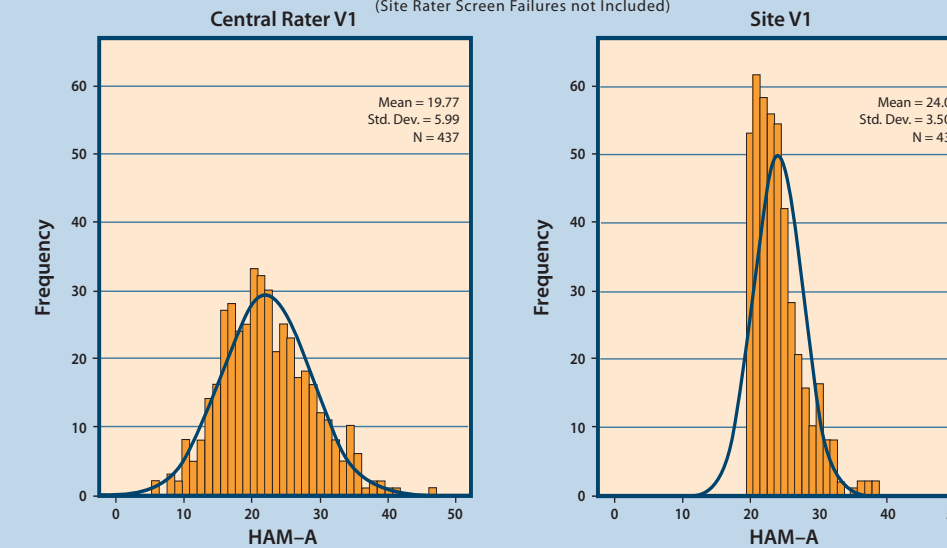
Cronbach's alpha scores (a measure of internal consistency) were much higher for the remote centralized raters than the site-based ratings at baseline (0.78 vs. 0.18, respectively). At endpoint, however, the alphas were identical (0.84 for remote centralized raters and 0.84 for site raters) [Fig. 3].\* The mean score differences on the two required items of the HAM-A were the most different of all the HAM-A items at baseline (mean difference of .41 points for 'anxious mood' and mean difference of .44 points for "tension"). At endpoint, differences between site raters and remote centralized raters on these two items were .08 and .023, respectively.

Placebo response is assessed by calculating the mean change from baseline to endpoint in an outcome measure. In this study, exploratory analyses found the mean change from baseline in HAM-A total score in the placebo group admitted to the study by site raters was -9.3 [Fig. 4]. This was significantly higher than the -5.9 point mean change on placebo as measured by the remote centralized raters in that same cohort. For the subgroup excluded by the remote raters, site raters measured more than twice the amount of placebo response detected by remote raters. In the cohort of subjects that were "qualified" by remote raters to enter the study, placebo response was just slightly lower than the site raters.

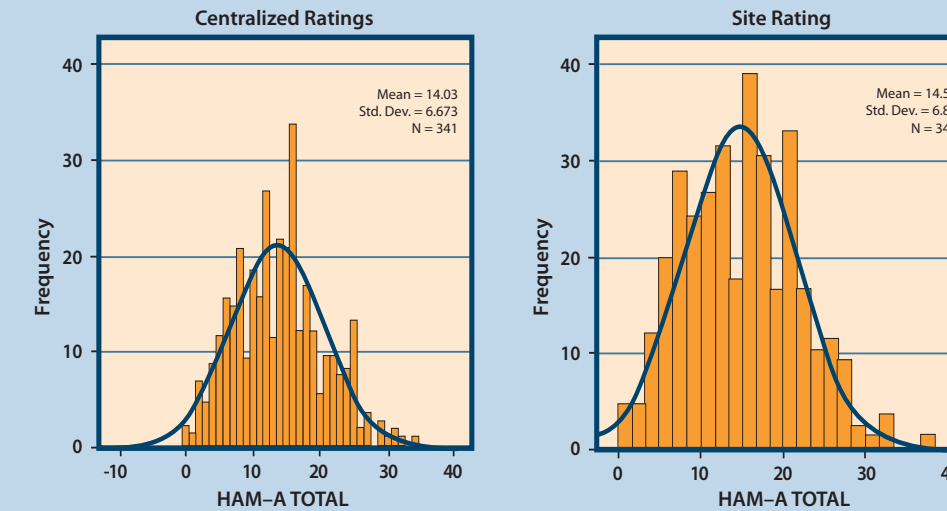
If one defines positive placebo response as a reduction in the HAM-A score of 50% or more, site raters identified 47 (39%) subjects as placebo responders, as compared to 29 (24%) subjects so classified by remote raters in the site-admitted cohort (nominal  $p=.015$ ).

\* It should be noted that truncation of range can artificially reduce Cronbach's alpha.

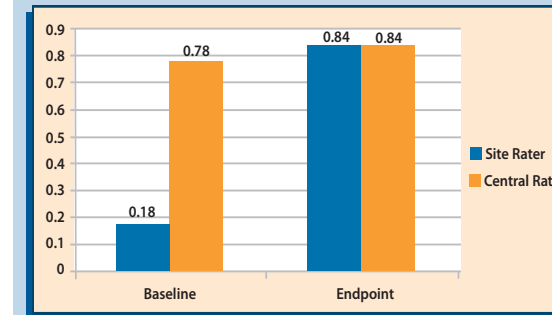
**Figure 1. BASELINE VISIT**  
Frequency Distribution of Baseline Visit Scores of Subjects Included in the Study by Site Raters (Site Rater Screen Failures not Included)



**Figure 2. ENDPOINT VISIT (week 6)**



**Figure 3. Internal Consistency Reliability (Coefficient Alpha) Site and Central Raters**



**Figure 4. Subjects Randomized to Placebo**

Site Raters as Gatekeeper	Subjects Randomized to Placebo			
	All Subjects Enrolled ("Qualified") By Sites	Cohort "Qualified" by Remote Centralized Raters (RCR)	Cohort "Excluded" by Remote Centralized Raters (RCR)	
Placebo Response (HAM-A Change from Baseline to Endpoint)	Site Ratings (n=122) -9.3 ± 6.2 (N=122)	RCR Ratings (n=59) -5.9 ± 5.6 (N=59)	Site Ratings (n=59) -8.3 ± 6.5 (N=59)	RCR Ratings (n=59) -7.3 ± 6.1 (N=59)
	Site Ratings (n=63) -10.3 ± 5.7 (N=63)	RCR Ratings (n=63) -4.6 ± 4.7 (N=63)		

**Discussion:** This study replicates at least two previous studies showing that blinded remote raters show a reduced placebo response compared to site raters (Kobak et al, 2009, and other poster). Several factors may contribute to this result. At baseline, blinded raters have no incentive to score subjects above an inclusion cut-off score. Therefore, as expected, the distribution of their baseline scores is highly normalized. There is evidence that site raters, who generally are incentivized to admit subjects to studies, may (consciously or unconsciously) inflate the baseline scores of subjects in order to include them in the study. This is supported by our finding that site raters' baseline scores on the two required HAM-A items were much higher than blinded raters' scores on those items at baseline but not at endpoint, when minimum scores on those items are no longer required. In addition, the finding of discrepant alphas at baseline (with site raters being low) but not at endpoint, together with the fact that the alphas were consistent for remote raters but highly inconsistent for site raters, suggests that some factor is influencing site raters' scores at study entry.

Site raters' mean scores at baseline were higher than remote raters' on all of the HAM-A individual items. At endpoint, however, remote raters scored half of the items higher than site raters. In addition, site raters showed a significantly larger placebo response than remote centralized raters on subjects whom they had already rated higher at baseline and who would have been excluded by central raters. Remote centralized raters may improve signal detection in clinical trials by decreasing placebo response as a result of appropriate baseline scoring. Remote centralized raters do not evaluate subjects at consecutive visits, and they are, in fact, blinded to study visit. Both of these factors may reduce expectancy bias in which raters expect to see change over time. In the absence of a positive control, however, it is not clear what the magnitude of change would be in an active treatment arm. Without knowing the ratio of drug to placebo change and variability, the impact on effect size is unknown.

Traditional rater training at an investigator meeting may align site raters in their scoring in the short term. However, such training has been shown to lack lasting effect through a study (Kobak et al). Other forces, such as rater drift, come into play and can significantly affect the fidelity of the way a scale is administered. Consistent administration and scoring of a rating scale requires a high quality of administration skill as well as consistent application of scale conventions and valid scoring of items (Lipsitz et al). Such skills can only be initially obtained and then maintained throughout a study with experienced raters and continual training and monitoring (Kobak et al, 2007).

Baer L, Cukor P, Jenike MA, et al. Pilot studies of telemedicine for patients with obsessive-compulsive disorder. *Am J Psychiatry* 1995;153:1383-1385  
 DeBrotta D, Demitrack M, Landin R, Kobak KA, Greist JH, Potter W. A Comparison Between Interactive Voice Response System-Administered HAM-D and Clinician-Administered HAM-D in Patients with Major Depressive Episode. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 39th Annual Meeting, Boca Raton, FL, 1999  
 Feltner DE, Kobak KA, Crockatt J, Haber H, Kavoussi R, Pandey A, Greist JH. Interactive Voice Response (IVR) for Patient Screening of Anxiety in a Clinical Drug Trial. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 41st Annual Meeting, Phoenix, AZ, 2001  
 Glaudin V, Smith W, Ferguson J, DuBoff E, Rosenthal M, Mee-Lee D. Discriminating placebo and drug in generalized anxiety disorder (GAD) trials: single vs. multiple raters. *Psychopharmacology Bulletin* 1994;32(2):175-178  
 Hamilton M. The assessment of anxiety states by rating. *Br J Med Psychol* 1959;32:50-55  
 Khan A, Kolts RL, Rapaport MH, Krishnan KR, Brodhead AE, Browns WA. Magnitude of placebo response and drug-placebo differences across psychiatric disorders. *Psychol Med*. 2005;35(5):743-9  
 Kobak KA, Lipsitz JD, Williams JBW, Engelhardt N, Jeglic E, Bellow KM. Are effects of rater training sustainable? Results from a multi-center clinical trial. *J Clin Psychopharmacol* 2007;27:534-535  
 Kobak KA, Williams JBW, Jeglic E, Salucci D, Sharp I. Face-to-face vs. remote administration of the Montgomery-Asberg Depression Rating Scale (MADRS) using videoconferencing and telephone. *Depress Anxiety* 2008;25:913-9  
 Lipsitz J, Kobak K, Feiger A, Sikich D, Moroz G, Engelhardt N. The Rater Applied Performance Scale (RAPS): Development and Reliability. *Psychiatry Res* 2004;127:147-155  
 Lynneham HJ, Rapee RM. Agreement between telephone and in-person delivery of a structured interview for anxiety disorders in children. *J Am Acad Child Adolesc Psychiatry* 2005;44(3):274-82  
 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  
 Sheehan DV, Lecrubier Y, Harnett-Sheehan K et al. The Mini International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998;59(suppl 20):22-33  
 Williams JBW. Structured Interview Guide for the Hamilton Anxiety Rating Scale (SIGH-A). 1996, Biometrics Research Department, New York State Psychiatric Institute, New York, New York  
**ACKNOWLEDGEMENTS**  
 The authors would like to acknowledge the help of Donna Salvucci, Christopher Klunkner, Matt Webster, and Melinda Snyder.  
**Author Disclosure Information:** 1. Williams, MedAvante, Part 1; MedAvante, Part 2; MedAvante, Part 3; MedAvante, Part 5; J. Dunn, (former) Sepracor, Part 1; K. Kobak, MedAvante, Part 1; MedAvante, NIH, Part 2; MedAvante, Part 5; E. Giller, Pfizer, Wyeth, Memory Pharmaceuticals, Marinus Pharmaceuticals, MedAvante, Part 1; MedAvante, Part 2; MedAvante, Part 5; L. Curry, Sepracor, Part 1; Sepracor, Part 2; Sepracor, Part 3; Sepracor, Part 5; P. Wilson, Sepracor, Part 1; Sepracor, Part 2; Sepracor, Part 3; Sepracor, Part 5; M. Detke, MedAvante, Eli Lilly, Part 1; MedAvante, Eli Lilly, Part 2; MedAvante, Eli Lilly, Part 3; MedAvante, Part 5