

Same Versus Different Raters and Rater Quality in a Clinical Trial of Major Depressive Disorder: Impact on Placebo Response and Signal Detection

Kenneth A. Kobak, PhD¹ Joshua D. Lipsitz, PhD¹ Alan D. Feiger, MD² Michael J. Detke, MD, PhD¹
¹MedAvante, Inc., ²University of Colorado Depression Center

ABSTRACT

Introduction: Minimizing error variance in clinical trials has traditionally been accomplished by using same rater at each visit. However, recent data suggest using different raters may enhance signal detection.

Method: We retrospectively compared patients who had the same rater at baseline and endpoint (N=163) to patients who had a different rater at baseline and endpoint (N=53) in a multisite (N=20) depression study. The study sponsor provided data from the active comparator (paroxetine; N=109) and placebo (N=107) cells.

Results: Subjects with the *same* interviewer at baseline and endpoint had a smaller mean HAMD change (drug minus placebo) vs. those with different raters (+0.56 vs. -3.76 respectively). The greatest paroxetine-placebo difference was found with good interview quality¹ and different raters at baseline and endpoint (mean HAMD change = -15.5, N=5, p=.008). The small cell sizes make the risk for Type I error great, and thus this result should be interpreted with caution.

Discussion: In the present study, different raters at baseline and endpoint was associated with larger drug-placebo differences. Limitations include the use of retrospective data, and a small number of subjects per cell. It is unclear if these findings generalize to other disorders.

Introduction: Minimizing error variance in clinical trials has traditionally been accomplished by using same rater at each visit. However, clinicians (and patients) typically expect to see improvement over time rather than no change or worsening. Thus, knowing the study visit may subtly bias clinicians' ratings. In addition, seeing the same patient week after week can result in less thorough or independent probing. This figures most prominently in clinical trials when a patient is followed over time by a single clinician (especially if the rater thinks he or she knows which treatment the patient is receiving). Using different raters may tend to mitigate this, as well as the potentially confounding impact of the 'therapeutic alliance' (vs. change due to study drug).

Method: We retrospectively compared patients who had the same rater at baseline and endpoint (N=163) to patients who had a different rater at baseline and endpoint (N=53) in a multisite (N=20) depression study. The study sponsor provided data from the active comparator (paroxetine; N=109) and placebo (N=107) cells.

Results: Subjects who had the *same* interviewer at baseline and endpoint had a mean Hamilton Depression Scale (HAMD) change (drug change minus placebo change) of +0.56 (mean change on paroxetine = 9.1, mean change on placebo = 9.7), $t(161) = -4.89$, $p = .625$. Those with *different* raters at baseline and endpoint had a mean HAMD change (drug minus placebo) of -3.76 (mean change paroxetine = 11.5, mean change placebo 7.7), $t(51) = 1.884$, $p = .065$.

The difference between drug minus placebo change for same raters and different raters was not statistically significant, $p = .062$. While the small cell size for different raters may have precluded statistical significance, results indicate a trend in this direction.

The same study also examined the impact of interview quality on signal detection. All baseline interviews were audiotaped and a random sample of 25% (N=56) were rated for interview quality using the Rater Applied Performance Scale (RAPS) scale. As previously reported¹, while overall paroxetine failed to separate from placebo, those whose baseline ratings were rated 'good' or 'excellent' did achieve a significant separation between change found for paroxetine (11.61) and placebo (4.78), $p = .02$, while those rated 'fair' or 'poor' did not (paroxetine = 7.56, placebo = 10.44), $p = 0.27$. The greatest paroxetine-placebo difference was found with good interview quality and different raters at baseline and endpoint (mean HAMD change = -15.5, N=5, $p = .008$) (none of the other comparisons were significant). The smallest difference was same raters and poor interview quality (mean HAMD change = 2.87, N=25). The small cell sizes make the risk for Type I error great, and thus this result should be interpreted with caution.

Discussion: One possible reason for the inferior signal detection with same raters is expectancy bias. Clinicians (and patients) typically expect to see improvement over time rather than no change or worsening. Thus, simply knowing the study visit may subtly bias a clinician's ratings. This figures most prominently in clinical trials when a patient is followed over time by a single clinician (especially if the rater thinks he or she knows which treatment the patient is receiving). When screening patients, the clinician is focused on finding and quantifying symptoms, in order to determine if the symptoms meet the threshold for study entrance. However, in subsequent visits, the focus is on determining if the patient's symptoms are decreasing, i.e., if the patient is improving. To some extent, this sets up an inherent expectation bias in the clinician to 'find what you are looking for'. In addition, a clinician who sees the same patient week after week can become accustomed to the patient's previous responses, and not probe as thoroughly or independently. Demand characteristics (such as the tendency for the patient to report improvement to please the physician) may also play a factor. Using different raters may tend to mitigate this, as well as mitigating the potentially confounding impact that the 'therapeutic alliance' may have on patient improvement (vs. change due to study drug). Limitations to the current study include the use of retrospective data. To fully test the hypothesis one would need to prospectively randomize subjects to same or different rater conditions. In addition, it is unclear if these findings generalize to other disorders, such as schizophrenia or other psychotic disorders, where use of the same rater may facilitate patient disclosure.

Reference: 1. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *Am J Psychiatry*. Mar 2005;162(3):628.

Figure 1.

Change from Baseline at Endpoint (Paroxetine Minus Placebo) on Hamilton Depression Scale (HAMD) Scores for Subjects with Same (N=163) vs. Different (N=53) Raters at Baseline and Endpoint

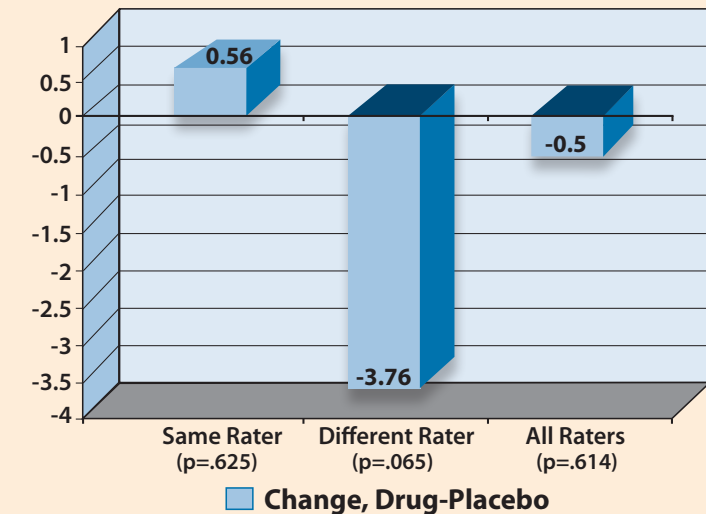


Figure 2.

Change (Paroxetine Minus Placebo) from Baseline to Endpoint on Hamilton Depression Scale Scores: Interaction Between Interview Quality and Same vs. Different Rater

	Good or Excellent Interview Quality	Fair or Poor Interview Quality
Same Rater	-4.27 (N = 17)	+2.87 (N = 25)
Different Rater	-15.50 (N = 5)*	+3.50 (N = 9)

*p = .008

Note: Sample size limited to the 56 subjects with ratings on interview quality

Reference: 1. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *Am J Psychiatry*. Mar 2005;162(3):628.

Kenneth A. Kobak, PhD, Joshua D. Lipsitz, PhD and Michael J. Detke, MD, PhD, are employees of MedAvante, which provides centralized ratings for CNS trials. Michael J. Detke, MD, PhD, is a major stockholder at Eli Lilly and MedAvante. Alan D. Feiger, MD, is a consultant to Bristol Myers Squibb and Neuronetics.

Author Disclosure Information: K. Kobak, MedAvante, Part 1; MedAvante, NIMH, Part 2; MedAvante, Part 5; J. Lipsitz, MedAvante, Part 1; MedAvante, Part 2; A. Feiger, Bristol Myers Squibb, Neuronetics, Part 1; Bristol Myers Squibb, Neuronetics, Part 2; Bristol Myers Squibb, Neuronetics, Part 3; Bristol Myers Squibb, Neuronetics, Part 4; M. Detke, MedAvante, Eli Lilly, Part 1; MedAvante, Eli Lilly, Part 2; MedAvante, Eli Lilly, Part 3; MedAvante, Part 5.

