

# The Importance of Quality in Post-baseline Assessments in CNS Trials

Williams, JBW<sup>1,2</sup> Kobak, KA<sup>3</sup> Detke, M<sup>1,4</sup>

<sup>1</sup>MedAvante, Inc., <sup>2</sup>College of Physicians and Surgeons, Columbia University, <sup>3</sup>Center for Psychological Consultation, <sup>4</sup>Indiana University School of Medicine

## ABSTRACT

**Background.** Pressure to inflate baseline scores, functional unblinding, rater drift, and expectancy bias may all contribute to trial failure. Inappropriate subjects may be enrolled in a study when enrollment pressures cause inflated baseline severity scores. An increasing number of studies now include methods such as independent blinded raters to ensure that appropriate subjects are entered into the trial.

Post-baseline factors can also affect outcomes. Functional unblinding leading to expectancy bias can introduce error and reduce signal detection. Rater drift can introduce variability and measurement noise. Independent raters blinded to study visit can minimize functional unblinding and expectation bias that can obscure a drug-placebo difference. Rater drift can be avoided by continuously calibrating raters.

**Methods.** Studies with ratings by both site raters and uniformly blinded, independent central raters can be evaluated to see how critical continued blinding and continuous calibration post-baseline are. In a trial of acute schizophrenia, independent remote blinded raters conducted the PANSS and site raters used the BPRS on the same 313 subjects. A study of 259 subjects with Parkinson's psychosis included blinded independent central ratings in the US, and traditional site ratings in the non-US sites. Both sets of raters used subscales of the SAPS. A negative GAD trial had blinded independent raters evaluate 122 subjects admitted to the study by site raters' SIGH-A baseline evaluations.

**Results.** In the schizophrenia trial, the independent blinded raters separated on placebo, the active comparator, and one of two test arms throughout the study. Site raters separated on the active comparator, but not on the two test arms. In the Parkinson's psychosis study, blinded, independent raters separated the test drug from placebo, but site raters outside of the US did not detect a signal. In the GAD trial, blinded, independent central raters had a lower placebo response than site raters, independent of subject selection.

**Conclusions.** Data from several studies support the continued importance of rater blinding and independence, post subject selection. Precision of ratings beyond baseline can increase the sensitivity of findings in a clinical trial, decrease placebo response rates and potentially eliminate Type II errors (false negatives). Blinding raters to study protocol and visit number can decrease or eliminate expectation bias and functional unblinding in post-baseline measurements. Independence from subjects' sites minimizes expectation bias by reducing the time rating staff spends with subjects. Rater drift, even with experienced raters can be diminished only through continuous calibration of the cohort of raters.

## INTRODUCTION

Inappropriate subject selection, "functional unblinding," rater drift, and expectancy biases can all contribute to trial failure. At baseline, pressure to enroll subjects can cause either inflation of baseline severity scores or inappropriate diagnoses. An increasing number of studies now include some method, such as the use of independent blinded raters, for ensuring that the right subjects are entered into a trial.

Post-baseline factors, however, can also affect trial outcomes. Functional unblinding, rater drift and expectancy biases can obscure a drug-placebo difference or introduce Type II (false negative) errors. Familiarity with a subject over the course of a study can influence a rater's scoring and create expectancy biases. Observing adverse events may lead one to assume the subject is on the investigational drug or active comparator (functional unblinding). Measurement noise or variability is introduced by drift from standardized scoring conventions among raters over a study's duration. Expectancy bias can increase with the amount of time the subject spends at the trial site<sup>1</sup>. Independent raters, blinded to the study treatment and visit number and continuously calibrated may mitigate functional unblinding, rater drift and expectancy biases.

## KEY FOR RATING SCALES

**PANSS** = Positive and Negative Syndrome Scale    **SAPS** = Scale for the Assessment of Positive Symptoms  
**BPRS** = Brief Psychiatric Ratings Scale            **NPI** = Neuropsychiatric Inventory  
**MMSE** = Mini-mental State Examination

## METHODS

Studies with ratings by both site raters and blinded, independent central raters can be evaluated to see how critical continued blinding and continuous post-baseline calibration are. In a trial of acute schizophrenia, blinded, independent central raters conducted the PANSS and site raters used the BPRS on the same 313 subjects. A study of 287 subjects with Parkinson's psychosis included blinded independent central ratings in the US, and traditional site ratings in the sites outside of the US (OUS). Both sets of raters used subscales of the SAPS as the primary outcome measure. A negative trial of generalized anxiety disorder (GAD) had blinded independent raters evaluate 122 subjects assigned to the placebo arm who had been admitted to the study by site raters' SIGH-A baseline evaluations.

### Study #1: Acute Schizophrenia<sup>2</sup>

**Sample:** N=313 hospitalized subjects with acute schizophrenia  
**Study inclusion:** PANSS  $\geq 70$  and  $\leq 120$  by blinded centralized raters  
**Primary outcome measure:** PANSS by blinded, independent central raters (converted to a derived BPRS score)  
**Other measures:** BPRS by site raters, with access to blinded raters' PANSS scores (blinded raters always went first)  
**Study duration:** Six weeks of treatment  
**Study arms:** placebo, active comparator (olanzapine), and two experimental drug treatment arms  
**Blinded independent continuously-calibrated raters:** N=18  
**Traditional site raters:** 35 sites

### Study #2: Parkinson's Psychosis<sup>3</sup>

**Sample:** N=287 subjects with Parkinson's psychosis  
**Study inclusion:** MMSE of  $<21$ ; symptoms severe enough to warrant treatment with an antipsychotic agent as documented by items A and B of the NPI, and defined as the sum of Hallucinations (Frequency  $\times$  Severity) and Delusions (Frequency  $\times$  Severity)  $\geq$  a total score of 4  
**Primary outcome measure:** SAPS subscales for hallucinations and delusions  
**Study duration:** Six weeks of treatment  
**Study arms:** placebo, 10 mg pimavanserin, 40 mg pimavanserin  
**Blinded independent continuously-calibrated raters in US:** N=11; 50 sites in the US (no site raters in US)  
**Traditional site raters:** 36 sites outside of the US (OUS) (no blinded raters OUS)

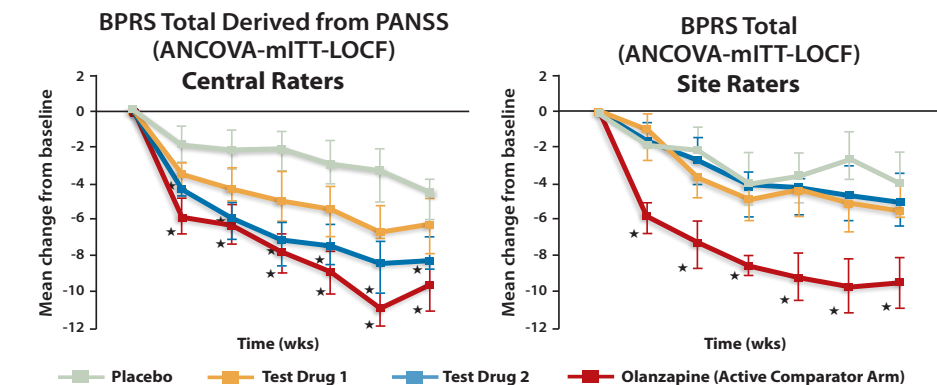
### Study #3: Generalized Anxiety Disorder<sup>4</sup>

**Sample:** N=122 subjects with generalized anxiety disorder in placebo arm  
**Study inclusion:** HAMA  $\geq 20$  at screen and baseline and  $\geq 2$  on HAMA items 1 and 2 by site raters  
**Primary outcome measure:** SIGH-A by site raters  
**Other measures:** Blinded, independent central raters interviewed at baseline and at week six; at baseline site raters always went first; at week six, rater order was counterbalanced  
**Study duration:** Six weeks of treatment  
**Study arms:** placebo, 0.9 mg/day experimental dose, and 1.5mg/day experimental dose  
**Blinded independent continuously-calibrated raters:** N=22  
**Traditional site raters:** 119 raters at 45 sites

## RESULTS

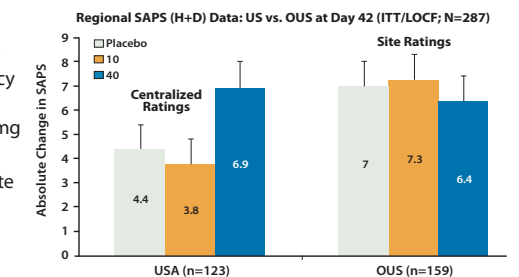
### Study #1: Acute Schizophrenia

- Blinded, independent central raters observed statistically significant efficacy at every time point in one experimental dose arm.
- Site ratings failed to show efficacy in either dose arm, at any time point.



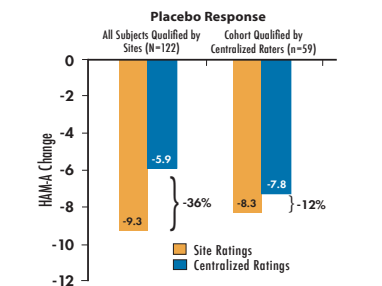
### Study #2: Parkinson's Psychosis

- Blinded, independent central raters in the US observed statistically significant efficacy at two weeks and marginally statistically significant efficacy at six weeks in the 40 mg dose arm.
- Site ratings outside the US did not separate any drug dose from placebo.



### Study #3: Generalized Anxiety Disorder

- Active treatments failed to separate by any measure; no positive control included.
- Analysis focused on placebo response.
- In overall population (N=122) placebo response was 36% lower as assessed by blinded, independent central raters. Even in the cohort included by the central raters (n=59) placebo response was 12% lower as assessed by centralized ratings.



## CONCLUSIONS

The Schizophrenia study demonstrates that post-baseline ratings performed by blinded, independent central raters did detect separation of drug and placebo in study arms where site ratings did not. Since both central ratings and site post-baseline ratings were conducted with the same subjects, the outcome improvements may be attributed to differences in post-baseline quality. In the Parkinson's psychosis study, blinded, independent central raters found differences between drug and placebo while OUS site ratings did not observe any separation. Finally, in the GAD study, blinded, independent central raters had smaller placebo response than raters based at the sites. We hypothesize that familiarity with the subject, observation of adverse events, and knowledge of visit sequence may have led to expectancy biases on the part of site raters regarding degree of change or which treatment arm a subject was enrolled in.

Data from several studies now support the importance of the accuracy of outcome assessments after subject selection, even when subject selection is performed by centralized raters. Continued vigilance and precision of ratings beyond baseline can increase the sensitivity of findings in a clinical trial and decrease placebo response rates. Blinding of raters to study protocol and visit number, independence from subjects' sites, and using different raters at consecutive visits minimizes expectancy bias; non-specific treatment effects are reduced by limiting the rating staff interactions with subjects. Rater drift, even with experienced raters, can be diminished only through continuous calibration of the cohort of raters.

### References:

- Guico-Pabia CJ, Musgnung J, Pederson R, Ninan PT: Placebo response in trials of antidepressants in patients with major depressive disorder. Poster presented at the annual conference of the National Clinical Drug Evaluation Unit 2010.
- Data on file MedAvante, Inc.
- Friedman JA, Ravina B, Mills R, Williams H, Bahr D, Peters P, Tison F, Burn D: Pimavanserin phase III PDP results. Presented at the annual conference of the American Academy of Neurology 2010.
- Williams JBW, Dunn J, Kobak KA, Giller E, Curry L, Wilson P, Detke M: Placebo response assessed by site and remote blinded centralized raters in a GAD trial. Poster presented at the annual conference of the American College of Neuropsychopharmacology 2009.

### Source of Funding:

MedAvante Research Institute, Sunovion (formerly Sepracor), and Acadia.