

Can Improved Signal Detection in International Schizophrenia Trials Be Attained with Blinded Independent Calibrated Raters?

April 3, 2011

Janet B.W. Williams, DSW

**VP, Clinical Development
MedAvante**

**Professor Emerita
College of Physicians and Surgeons
Columbia University**

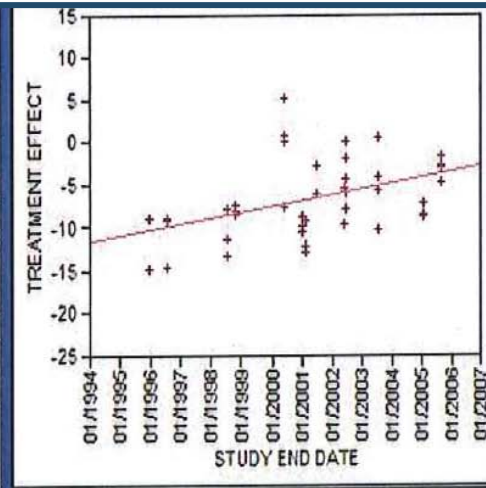
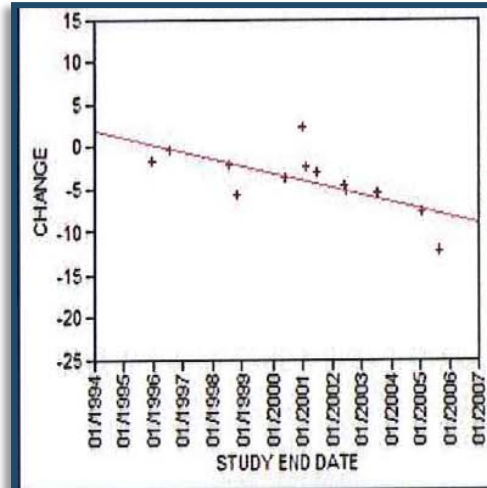
FDA Sees Trends In U.S. And Overseas Trials, 1994–2007¹: Increasing Placebo Effect, Decreasing Drug Effect

Time Plot in Schizophrenia

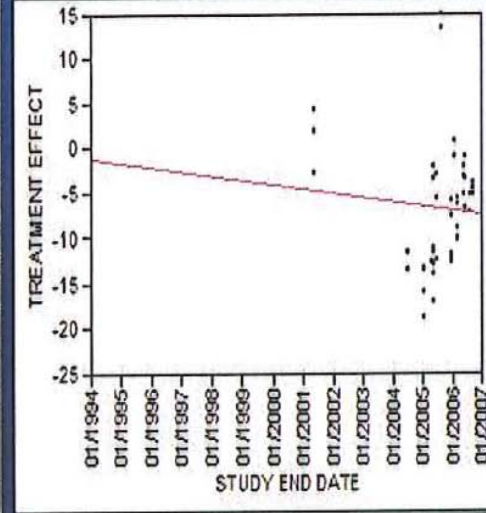
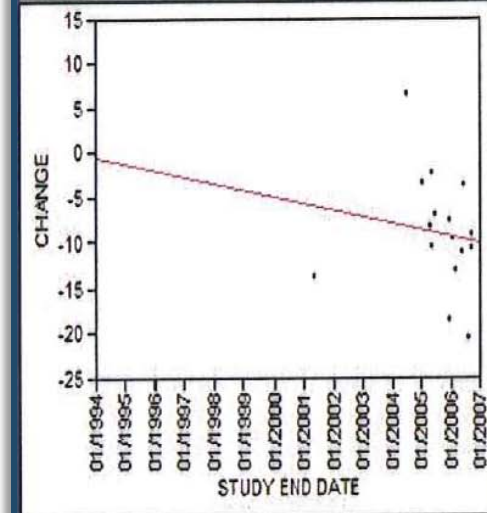
Placebo Response

Treatment Effect

U.S.



Non-U.S.



¹ Khin N A MD “Update on Regulatory and Scientific Issues Regarding the Reliance of Efficacy Data from Foreign Sites to Support New Drug Applications and Supplements”, NCDEU FDA Symposium, July 2, 2009

Some Major Threats To Signal Detection

- **Inappropriate subject selection**
- **Expectation bias**
- **“Functional unblinding”**
- **Relationship bias**
- **Poor interrater reliability**

Clinical Trial Assessments

Two major sources of trial failure

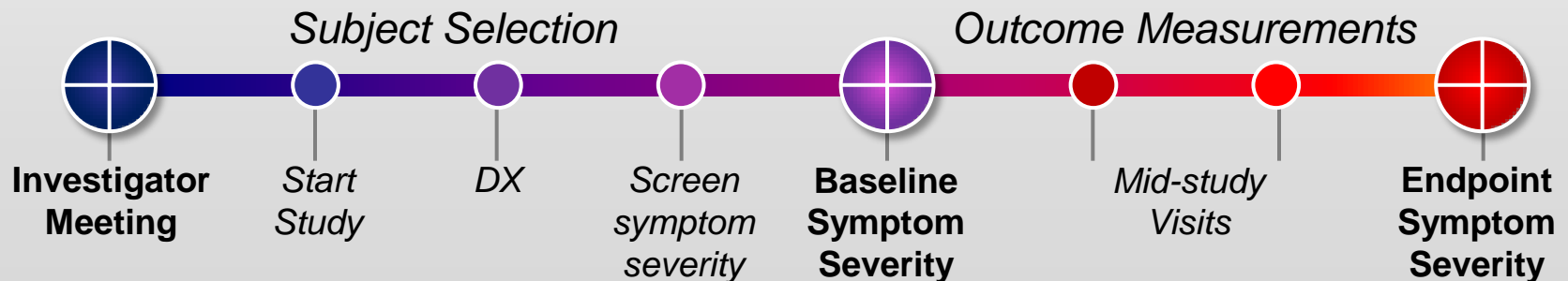
1. Inappropriate Subject Selection

- Enrollment biases
- Score inflation

2. Inaccurate/Variable Outcomes

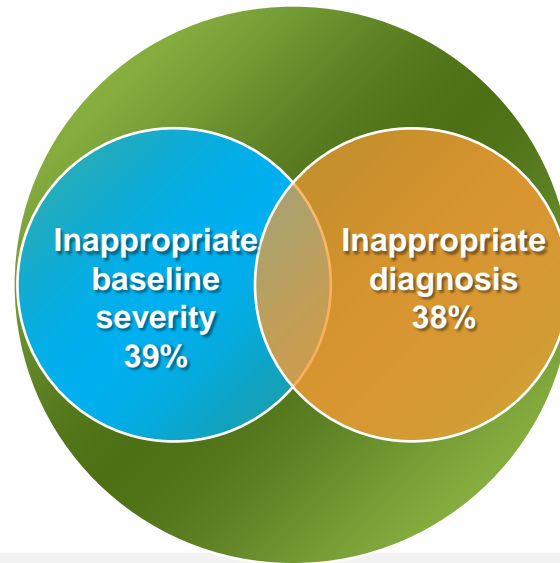
- Expectancy biases: priming
- Interviewer variability

Multiple entry points for bias and variability

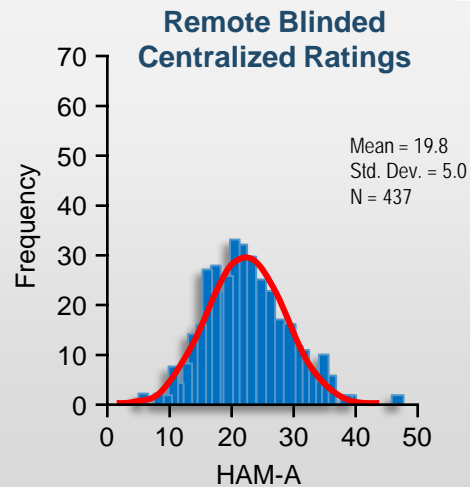
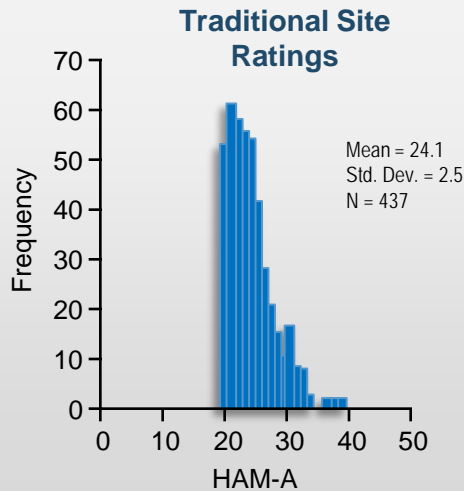


Inappropriate Subject Selection

- As many as two out of five subjects enrolled are ineligible according to symptom severity¹
- 38% are inappropriately diagnosed (not mutually exclusive)



Comparison of independent raters' scoring of the same subjects scored by site raters at baseline



- Proposed 'solutions' to CNS trial failure must exclude inappropriate subjects

1. MedAvante data on file

Reasons For Inappropriate Subject Selection

- **Lack of blinding**
- **Enrollment pressure**
- **Relationship bias**
- **Lack of diagnostic sophistication**

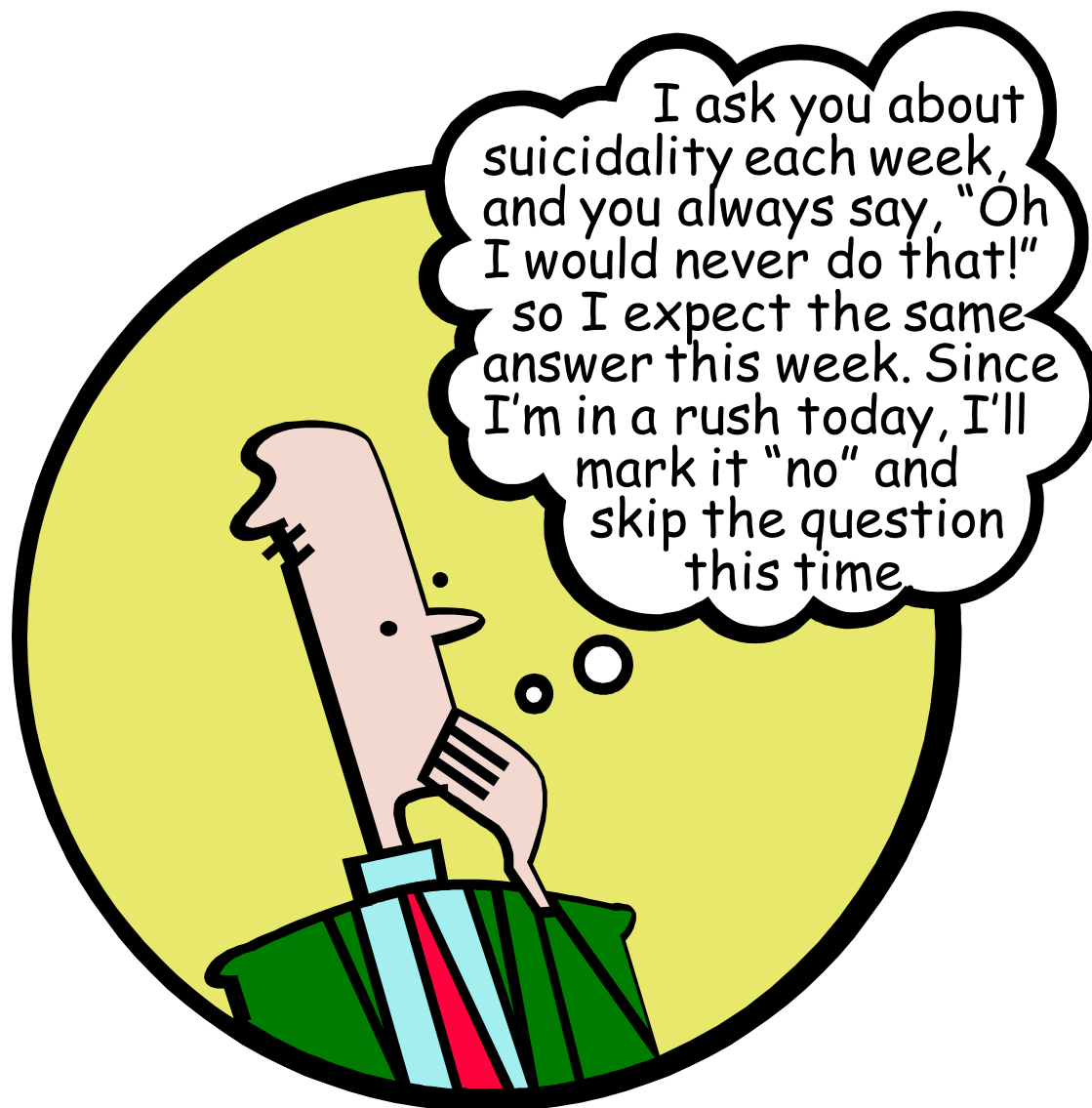
Some Major Threats To Signal Detection

- Inappropriate subject selection
- **Expectation bias**
- “Functional unblinding”
- Relationship bias
- Poor interrater reliability

Expectation Bias

- **Occurs when an individual's expectations about an outcome influence one's perceptions of one's own or others' behavior**
- **Both raters and subjects may enter a trial with expectations for the outcome**
- **When expectations (usually of improvement) influence raters' scores and subjects' reports, signal detection can be impacted**
- **Examples abound in everyday life and all fields of medicine**

Expectation Bias In Raters



Expectation Bias In Subjects



Some Major Threats To Signal Detection

- Inappropriate subject selection
- Expectation bias
- **“Functional unblinding”**
- Relationship bias
- Poor interrater reliability

“Functional Unblinding”

- **Occurs when a test drug or active comparator is associated with adverse events that unblind raters as to treatment allocation**
- **Examples:**
 - ▶ **Antipsychotics causing sedation, dry mouth, weight gain**
 - ▶ **SSRI/SNRI causing nausea, sexual dysfunction**
- **Functional unblinding leading to expectation bias can introduce error and reduce signal detection**
- **May disproportionately penalize novel treatments with milder side effect profiles**

Some Major Threats To Signal Detection

- Inappropriate subject selection
- Expectation bias
- “Functional unblinding”
- **Relationship bias**
- Poor interrater reliability

Relationship Bias

- **Occurs when a person's relationship with another person influences one's own or others' behavior**
- **The effect of therapeutic alliance is amplified as the number of follow-up visits increases and may contribute to increased placebo response and decreased signal detection**
- **Examples:**
 - ▶ **Posternak and Zimmerman, 2007 (as the number of follow-up visits increases, placebo response also increases)**
 - ▶ **Guico-Pabia, Musgnung, Pedersen, and Ninan, 2010 (as the number of assessments per visit increases, effect sizes decrease)**

Some Major Threats To Signal Detection

- Inappropriate subject selection
- Expectation bias
- “Functional unblinding”
- Relationship bias
- **Poor interrater reliability**

Poor Interrater Reliability

- **Interrater reliability rarely measured in clinical trials, except maybe at investigator meetings (and generally runs in “fair” to “good” range)**
- **Interrater reliability affects study power, sample size, and signal detection**
- **One-time training at investigator meeting does not improve ongoing reliability (*Demitrack et al., 1998*)**
- **Increases with number of raters/sites/countries**
 - ▶ **Rater turnover**
 - ▶ **Cultures, languages, translations**
 - ▶ **Rater drift over time**

The Danger Of Rater Drift

- **In mood disorders:**

“The results indicate that although rater training can successfully improve raters’ applied clinical skills, these skills can erode over the course of a trial.”

Kobak KA, Lipsitz J, Williams JBW, Engelhardt N, Jeglic E, Bellew K: Are Effects of Rater Training Sustainable? Results from a Multi-Center Clinical Trial. *Journal of Clinical Psychopharmacology* 27:5, 534-535, 2007. (Quote on page 534.)

- **In psychosis:**

“...drift is a critical concern for quality control in longitudinal studies.”

“...rater bias and drift have been well documented as methodologic problems in clinical assessment.”

Ventura J, Green MF, Shaner A, Liberman RP. Training and quality assurance on the Brief Psychiatric Rating Scale: the “drift busters”. *International Journal of Methods in Psychiatric Research* 1993;3:221–224. (Quotes on page 222.)

A Need For Change

- **High proportion of trial failures points to the need for methodological solutions to:**
 - ▶ **Improve identification of appropriate subjects**
 - ▶ **Increase reliability and accuracy of outcomes measurement**
- **Many strategies have been proposed to improve the accuracy of subject selection and outcome measurement**
- **There is substantial evidence that the use of *expert blinded, independent, calibrated clinicians* is an effective way to reduce placebo response and improve signal detection**

Centralization In Other Therapeutic Areas

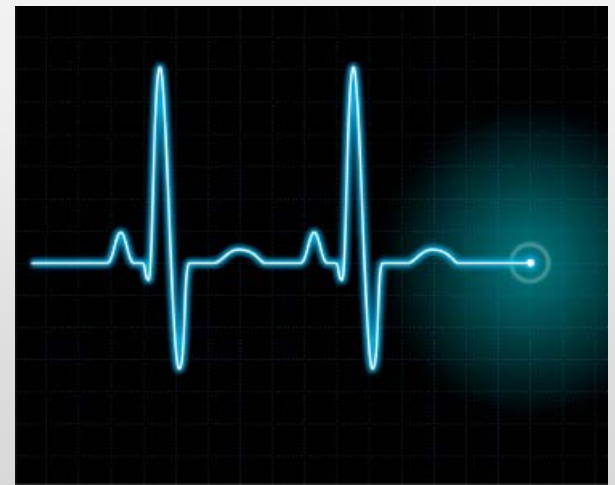
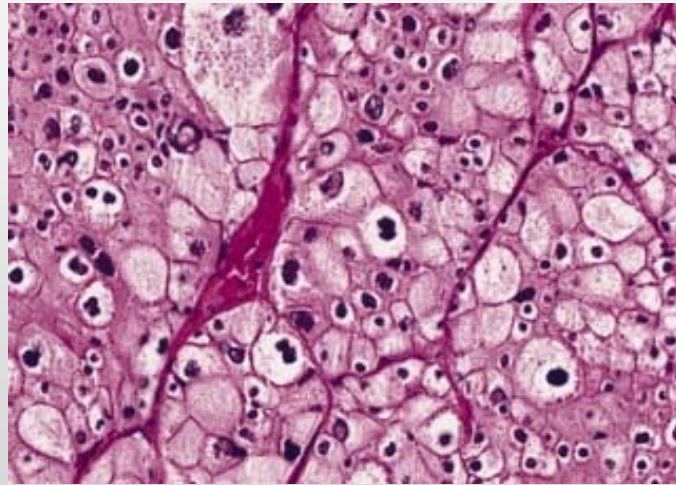
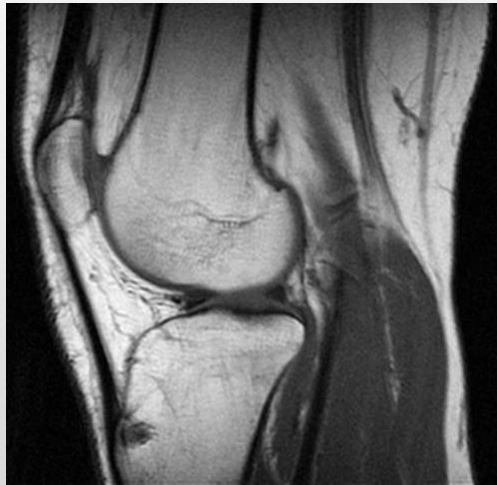
FDA guidance documents and EMA guidelines emphasize the value of centralization

- **“Centralised blinded review is needed in order to establish progression.”**

(from Appendix 2 to the Guideline on the evaluation of Anticancer Medicinal Products in Man (CPMP/EWP/205/95 Rev. 3) on Confirmatory studies in Haematological Malignancies, EMA)

- **“We recommend that blinded image evaluations by multiple independent readers be performed in the phase 3 efficacy studies.”**

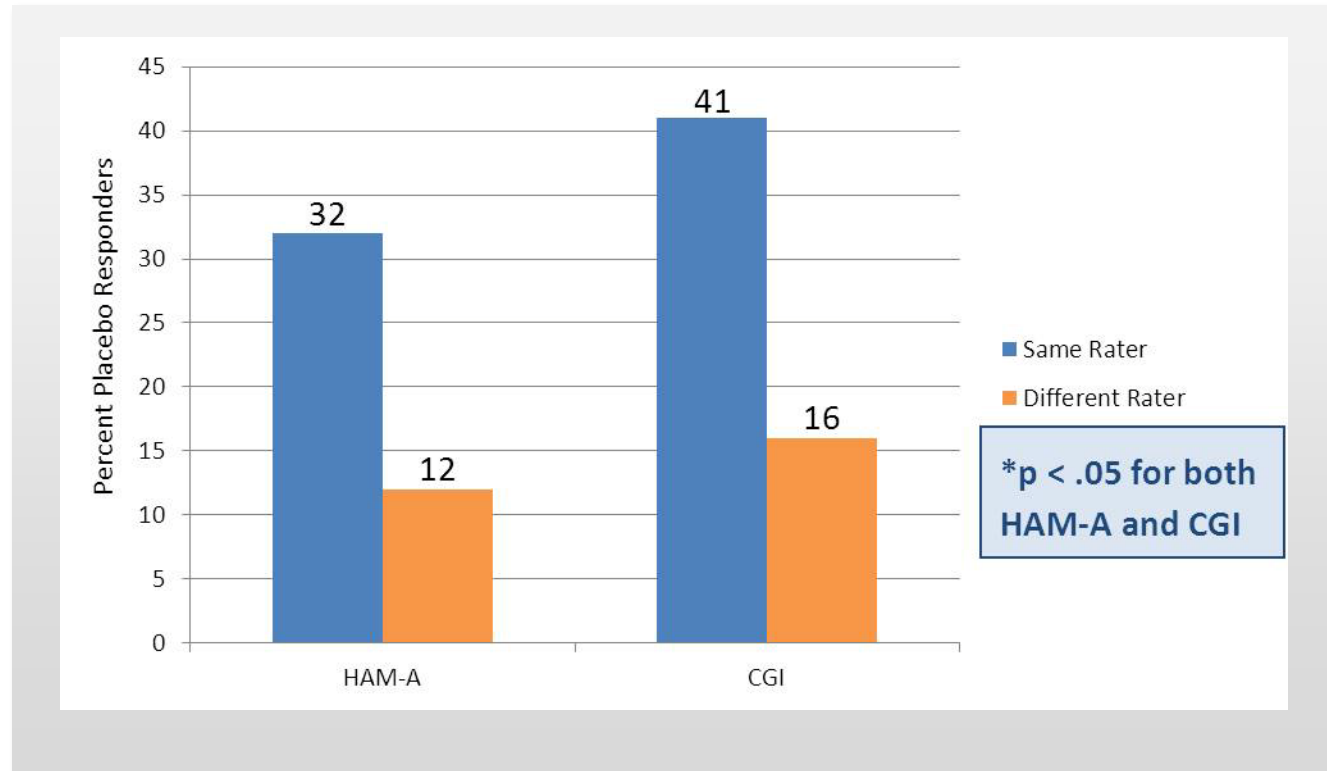
(from Guidance for Industry Developing Medical Imaging Drug and Biological Products Part 3: Design, Analysis, and Interpretation of Clinical Studies, FDA)



A New Model For Clinical Trial Assessments: Centralized Raters

- **Expert**
- **Blinded**
- **Independent**
- **Calibrated**

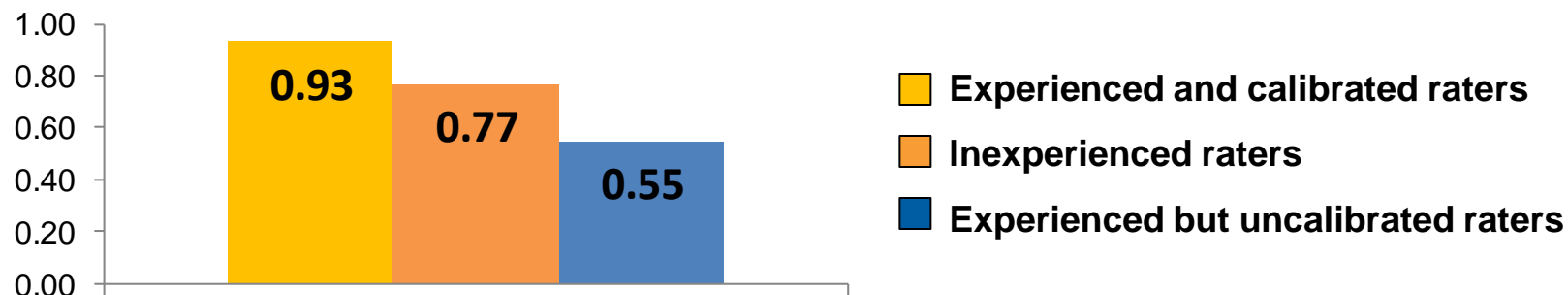
Using Multiple Raters Can Avoid Many Sources Of Bias



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. *Psychopharmacol Bull* 32: 175-8 (1994)

Experience Doesn't Help If Raters Are Not Calibrated

Study of ICC values for three groups of raters¹



¹Kobak, K.A., Brown, B., Sharp, I., Levy-Mach, H., Wells, K., Okun, F., and Williams, J. B. W.(2009). Sources of Unreliability in Depression Ratings. *Journal of Clinical Psychopharmacology*, 29, 82-85.

- **Experienced raters without cohort calibration had the lowest interrater reliability (ICC)**
- **Raters exposed to different scoring conventions, as a result of their experience, may make interrater reliability worse**
- **Need rigorous calibration to other study raters to minimize variability**

Continuously Calibrated Central Rater Cohort Can Sustain Quality And Accuracy

- ICCs of central rater cohort remains high over duration of a study
 - ▶ ICC of interviewer and expert rater observing interview real-time

	Q1	Q2	Q3	Q4
HAM-A (N=100) (Study 1)*	0.90	0.96	0.95	0.97
HAM-A (N=68) (Study 2)*	0.98	0.97	0.97	0.96
PANSS (N=131) (Study 3)*	0.90	0.90	0.96	0.88
PANSS (N=67) (Study 4)*	0.98	0.97	0.98	0.98

* MedAvante study data on file and made available to sponsor during study.

Consequence Of Appropriate Subject Selection Is Reduced Placebo Response

MDD Study¹
Placebo Response:



p = 0.001 **n = 51**
Standard Deviation:
Site Raters (n = 13) 5.99
Central Raters (n = 8) 6.60

1. Kobak, K.A. et al. (2010, April). Site versus Centralized Raters in a Clinical Depression Trial: Impact on Patient Selection and Placebo Response. *Journal of Clinical Psychopharmacology*, 30 (2) 193-1972.

GAD Study²
Placebo Response:

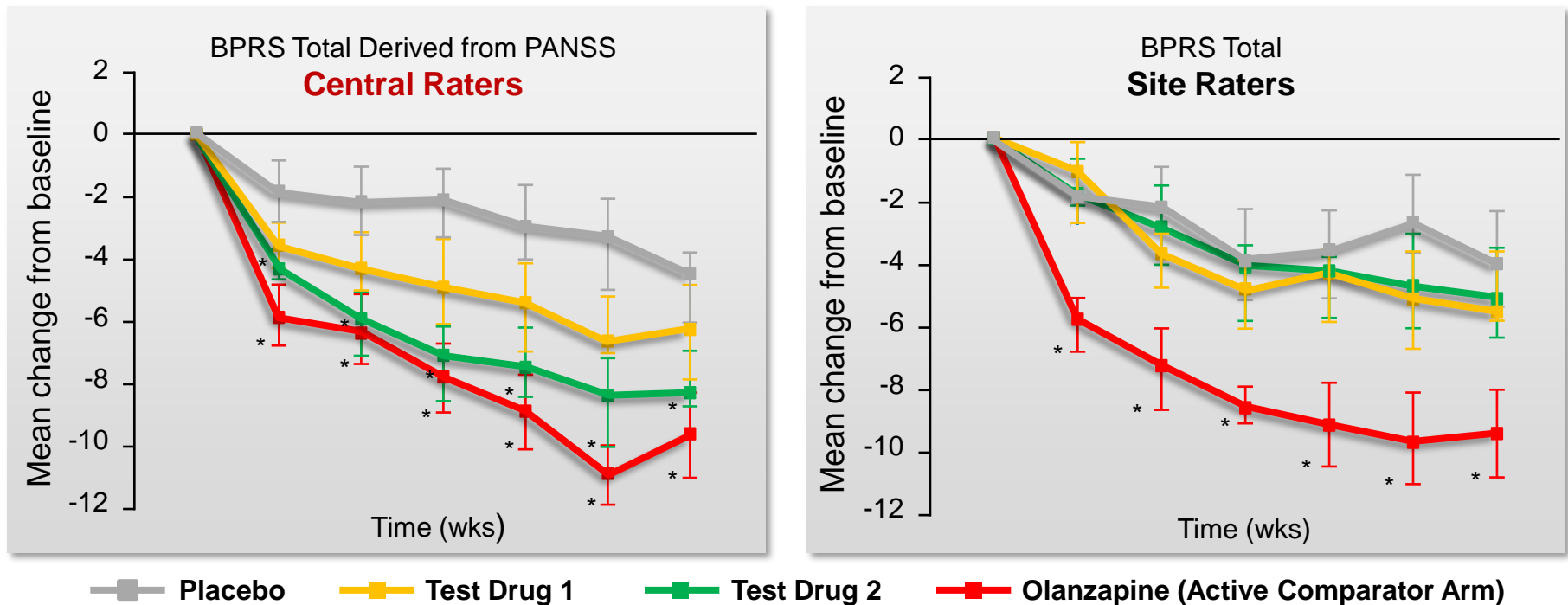


p < 0.001 **n = 220**
Standard Deviation:
Site Raters (n = 122) 6.20
Central Raters (n = 22) 5.60

2. Williams, J.B.W. et al. (2010). Placebo Response Assessed by Site and Remote Blinded Centralized Raters in a GAD Trial. Presented at NCDEU Annual Meeting, Boca Raton, FL.

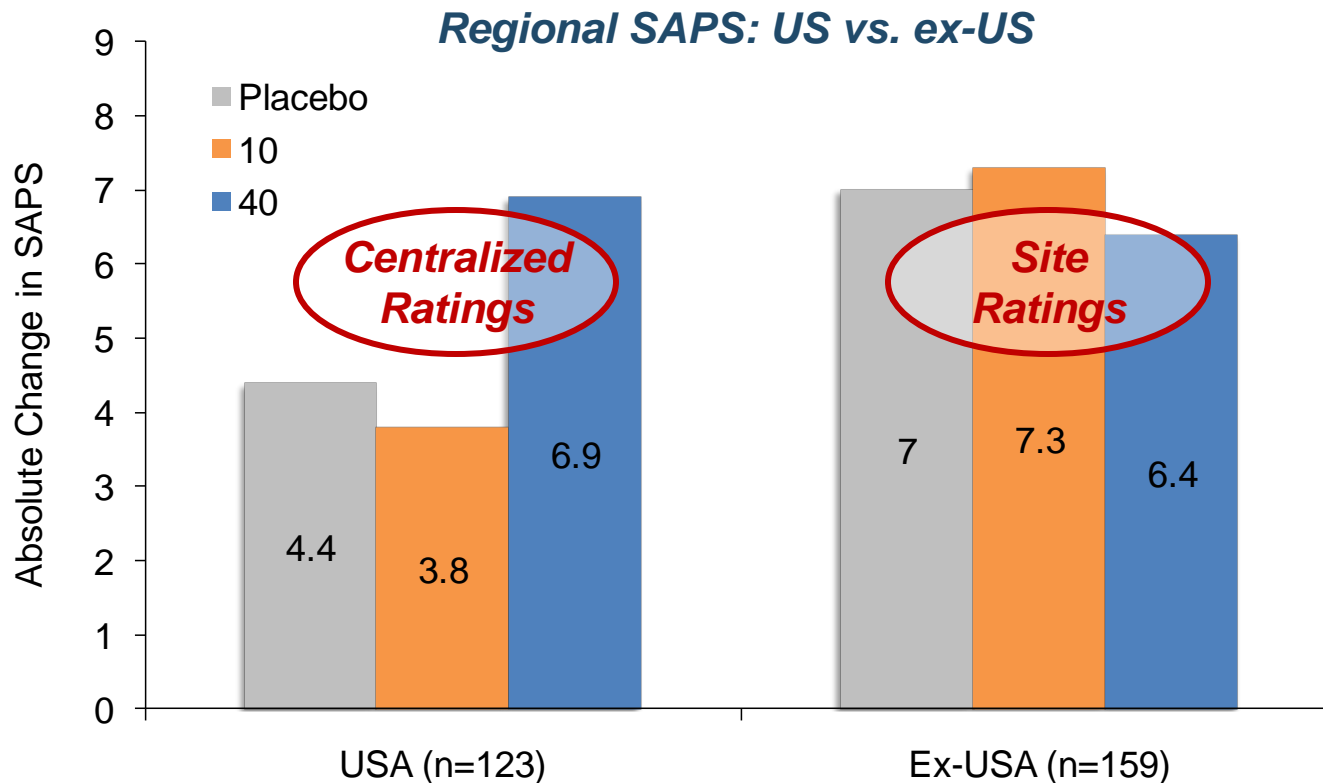
Value Of Blinded Independent Raters In Schizophrenia

- Blinded independent raters and site raters evaluated same subject sample
 - ▶ Central raters determined inclusion by severity
- Blinded independent (central) raters detected signal for test drug – unblinded sites did not
 - ▶ Site raters may have been “functionally unblinded” to olanzapine



Parkinson's Psychosis: Elimination Of False Negative

- Blinded independent central ratings in U.S. separated test drug from placebo
 - ▶ Site ratings outside U.S. did not separate



Friedman, J.H. (2009, April). Pimavanserin Phase III PDP Results ACP-103-012. Presented at AAN 2010 Conference Toronto, Canada.

Summary: Ways To Avoid Threats To Signal Detection

- **Independence**

- ▶ Inappropriate subject selection
- ▶ Expectation bias
- ▶ Relationship bias

- **Blinding**

- ▶ Inappropriate subject selection
- ▶ Expectation bias
- ▶ “Functional unblinding”
- ▶ Relationship bias

- **Multiple Calibrated Raters**

- ▶ Expectation bias
- ▶ “Functional unblinding”
- ▶ Relationship bias
- ▶ Poor inter-rater reliability