

Scientific Rationale For MedAvante Assessment Services

This document provides the detailed rationale for the use of MedAvante Assessment services as the best methodology to bring certitude to patient ascertainment and assessment in clinical trials.

Executive Summary

1. Current CNS clinical trials (phase II – IV) which employ subjective diagnostic and efficacy rating instruments have an unacceptably high failure (false negative) rate.
2. There is extensive literature evidence that systematic biases are at play. In multi-site studies many patients are enrolled who do not meet appropriate diagnostic or severity inclusion criteria. In assessing efficacy, accuracy is impaired by expectancy bias (rater-patient rapport), rater drift over time, and functional unblinding due to awareness by the raters of adverse events that may be indicative of treatment assignment, among other issues.
3. There is strong literature evidence that poor assessments reduce the ability to detect whether a drug is effective or not; the effects are striking – sample sizes of less than half have been shown to be effective when ratings are done well.
4. MedAvante employs state-of-the-art video- and audio-conferencing to minimize or eliminate biases. MedAvante also hires, rigorously trains, and continuously re-assesses expert clinician raters to ensure the best possible administration of subjective diagnostic and efficacy rating instruments.
5. Results are now available from several trials. In one study of GAD, MedAvante’s methodology doubled the effect size (drug-placebo separation, controlling for standard deviation). This would allow one to maintain the same statistical power with $\frac{1}{4}$ as many subjects. In a schizophrenia study, MedAvante raters detected statistically significant efficacy in a new antipsychotic medication whereas the site raters failed.

Overview

MedAvante is the creator and currently the sole practitioner of a methodology that systematically minimizes or removes sources of bias and variability in diagnosing and assessing participants in clinical trials employing subjective, clinician-assessed instruments. As a result of its methodology, MedAvante can reduce the number of study participants, timelines and costs of clinical programs while delivering more certain results. Ultimately, the benefit is bringing critically needed

therapeutics for debilitating diseases to patients faster, reliably and more efficiently and accurately.

Background

A recent review ¹ of the FDA Data Sets including 45 trials found that 36% of these studies failed to show superiority of a standard (positive control) antidepressant over placebo. (In the case of newly FDA-approved antidepressants, the failure rate was

Scientific Rationale For MedAvante Assessment Services

52%). The critical determining factor has been the increasing difficulty for trials with known effective drugs to show signal detection with similar numbers of patients as trials conducted in the 1960s, 1970s or early 1980s. A variety of studies illustrate how factors associated with subjective clinician assessments (i.e., inter-rater reliability, rater bias and interview quality) significantly impact signal detection and study outcome. Furthermore, failed Phase 2 studies cost millions of dollars, and a failed phase 3 study can cost tens of millions. Reducing the number of failed studies will result in significant costs savings in and of itself, even before the impact on incremental revenue due to earlier launch is considered. However, this too can be substantial; even 3 months shorter time to NDA would mean an additional \$250M on a drug that sells \$1B/year. Of even greater significance would be the termination if a program with significant sales potential due to an inability of clinical studies to detect efficacy for an efficacious drug. Such a scenario seems implausible until one considers the stunning statistic that over half of active comparators (i.e. known effective, marketed drugs) fail to separate from placebo in many significant CNS indications.

A solution to the problem of failed trials reviewed above is the use of centralized raters to perform the screening and outcome measures in clinical trials. Centralized raters refer to a 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 video or teleconferencing and remotely administer the key diagnostic scales(s) and primary outcome measure(s) to

study subjects during their regularly scheduled study visits.

This document supports the case for broader deployment of the MedAvante platform across CNS development programs, in order to maximize CNS development productivity and speed treatments to market sooner. In what follows we consider these problems, and how MedAvante's methodology has empirically demonstrated solutions for them.

Why Clinical Trials Fail

The Importance of Inter-Rater Reliability

The impact of reliability on power and sample size requirements can be profound²⁻⁴. Improving the intra class correlation coefficient (ICC) from .70 to .90 would decrease sample size requirements by 22% (holding standard deviation and power constant): a 3-arm study that normally would require 333 patients would need to enroll 74 fewer patients if reliability were enhanced to the higher level. At an estimated cost of 15 thousand dollars per patient, this would save the sponsor \$1.1M. Alternatively, a drop in ICC from .90 to .70 would reduce study power from .72 to .50,

Scientific Rationale For MedAvante Assessment Services

(controlling for standard deviation and N). Given the importance of inter-rater reliability on clinical trial outcomes it is surprising that published studies rarely report⁵ reliability figures, and sponsors rarely, if ever, attempt to establish the inter-rater reliability of their raters during the course of their studies. When done, establishment of their rater cohort's ICCs is usually at the outset of a study at the Investigator Meeting during which it is established through observation and scoring of videotapes. While this has become the standard means to measure reliability it does not evaluate an individual rater's ability to conduct an interview themselves (they are scoring an interview on videotape) and eliminates the information variance that results if each rater interviewed a patient independently – which is what happens during the conduct of an actual study. The rater training that is usually done at study outset can improve interview *quality*– but not *reliability*. Demitrack et al found no evidence of improved rating performance after 6 hours of reliability training⁶. There are no reports in the literature documenting improved inter-rater reliability as a result of rater training in multicenter trials using independent interviews.⁷

Site-specific rater biases

Several factors have led to a diminishing drug effect in CNS trials over the past few decades. Because of the intense pressure to enroll patients in a timely (and profitable) manner, local raters have a strong economic incentive to overstate the patient's symptom severity and to 'miss' exclusion factors. This results in the enrollment of patients who have the wrong diagnosis and/or mild illness. After these subjects begin double-blind therapy, whether on the drug or placebo, they

may appear to have improved rapidly when their inflated scores return towards their true values in subsequent visits. In addition, patients with lower severity levels have been shown to be less drug responsive and lower scores are associated with greater change on placebo¹¹. This can both drown out the drug response and create an artificially large placebo response. Moreover, patients who should have been excluded because they don't truly meet the sponsor's protocol-specific threshold for the diagnosis or have confounding exclusionary criteria may be less likely to respond to the investigative drug treatment intended for that specific disorder. Many studies have shown that site-specific rater bias results in over-inclusion of less severely symptomatic subjects and failure to exclude a significant number of potential participants who are not eligible for participation^{18, 9, 10, 4}. To be specific, the literature and MedAvante's experience show that, for typical trials of MDD, GAD, etc., approximately 40% of the enrolled patients should have been excluded. For trials with difficult diagnoses, this rate can be as high as 80% (unpublished data).

Rater bias also can adversely affect ratings during the course of the study in the form of unconscious biases raters cannot detect or deter. Referred to as expectancy bias or therapeutic alliance, clinical raters and patients generally will expect to see improvement over time rather than no change or worsening. Rater expectancy bias appears to be magnified when a single clinician conducts all of the ratings on a specific patient in a study, a bias that may be amplified when the rater is also the treating clinician. The value of a repeated contact with a caring professional should not be underestimated; and in our experience, study

Scientific Rationale For MedAvante Assessment Services

participants not infrequently misidentify their rater as a therapist. Posternak and Zimmerman found each additional follow-up during a 6-8 week clinical trial resulted in an additional reduction of about 1 pt. on a HAM-D – for subjects on both drug and placebo. It was estimated that the therapeutic impact of a repeated assessment accounts for about 40% of the placebo response¹¹. Counter-intuitively (given the problem of controlling inter-rater reliability) a number of studies have found that data from patients rated by different raters during the trial produced significantly greater separation of drug from placebo and lower placebo response than data from patients rated by the same rater. This has been found across disorders (depression, OCD, GAD, panic, and social anxiety disorders) and across different rating scales¹²⁻¹⁵.

Interview Quality

Two recent studies have documented the critical impact of interview quality, that is, a rater's applied clinical skills in actually conducting an interview, on signal detection^{16, 17}. Without taking interview quality into account, the trials failed to separate the active drug from placebo. However, when the analysis was limited to looking at only those subjects whose baseline interviews had a mean rating of 'good' or 'excellent' (using the Rater Applied Performance Scale, RAPS^{18,9}) a large and statistically significant effect was found for the active comparator versus the placebo arm. In the two studies cited above more than half of the ratings, blindly evaluated, were of fair or poor quality on every interviewing skill dimension the RAPS scale evaluates: adherence, clarification, follow up and neutrality. Interview quality is likely related to the educational background and to the amount and quality

of prior training and clinical experience using the scale, but may also be related to the time raters allot for assessments, and raters' level of motivation.

However, the beneficial impact of high interview quality can still be vitiated by site-specific biases: enrollment bias at study outset and expectancy/therapeutic alliance biases throughout the conduct of the trial. Moreover, a study's cohort of raters, all of whom may be individually high quality interviewers, can still result in poor reliability for the study as a whole -- with its attendant negative effect on study power. That is because each site is managed independently and the sites do not continuously calibrate with each other on what are fundamentally subjective measurements. This lack of cohesion results in slight differences on how similar subjects in the study are evaluated and magnifies variability in the aggregated interview score results. Lastly, rater training before the trial does not guarantee that the standards will be maintained during the course of the trial. Thus, unless rating monitoring is frequent, an unknown proportion of assessments will be suboptimal.⁷

MedAvante: A Structural Solution to Rater Bias and Variability

A solution to the problems reviewed above is the use of centralized raters to perform the subjective screening and outcome measures in clinical trials. Centralized raters refer to a 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 video or

Scientific Rationale For MedAvante Assessment Services

teleconferencing and remotely administer the subjective screening and outcome measures to study subjects during their regularly scheduled study visit.

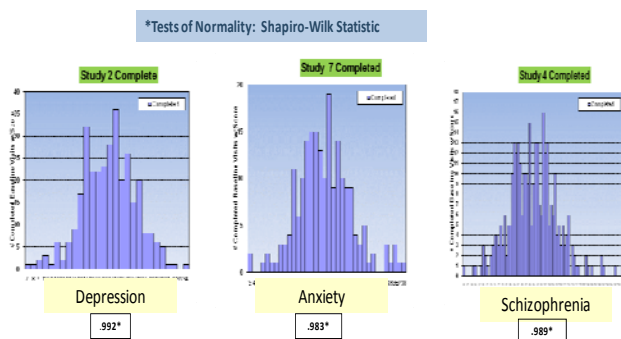
MedAvante Experience:

STUDY #	THERAPEUTIC AREA	PHASE	# MOS	# SITES	# PATIENTS	INSTRUMENTS	Pbo	Test	Act. Comp
1	MDD	POC*	36	2	62	HAMD	X	-	-
2	MDD	II	20	45	271	SIGH-ADS, HAMA	X	X	X
3	ADHD	Valid.	9	28	200	SCID, ACDS	-	-	-
4	Psychosis	Iia	10	30	306	PANSS	X	X	X
5	Schizophrenia	INV*	36	17	450	PANSS, QLS, PSP	-	-	XX
6	Schizophrenia	III	12	36	330	PANSS, PSP, CGI-S	X	X	-
7	GAD	II	6	47	204	SIGH-A (HAMA)	X	X	X
8	Neurology	Iib/III	28	32	168	SAPS	X	X	X
9	Schizophrenia	II	7	29	205	PANSS, CGI-S	X	X	-
10	Neurology	Iib/III	30	23	112	SAPS	X	X	-
11	GAD	II	10	50	441	SIGH-A (HAMA)	X	X	-
12	Psych. Dep.	III	30	30	450	SCID, BPRS, SIGH-D	X	X	-
13	Treat Resist. Dep.	I	12	5	27	SIGMA (MADRS)	X	X	-
14	Bipolar-Dep.	III	20	38	230	SIGMA (MADRS)	X	X	-
15	MDD	II	18	30	348	SCID, SIGH-D (HAMD) CGI-S	X	X	-
16	Schizophrenia	II	16	9	70	PANSS, NSA-16	-	X	X
17	Suicidality	III	21	50	1,000	SCID, GAD, PHQ, C-SSRS	-	X	X
18	MDD	II	11	19	180	GRID HAMD, SIGMA (MADRS)	X	X	X

MedAvante has completed or is in the process of supporting 18, large multi-center, phase 2 and phase 3 clinical studies across multiple psychiatric indications. This entailed setting up 275+ sites across the United States and administering and completing over 12,000+ remote sessions. The table above summarizes these.

No Enrollment Bias or Baseline Score Inflation

In every trial MedAvante conducted screening ratings, 18 trials completed and in progress, the statistically normal distribution of the baseline scores demonstrates that central raters are not subject to the same biases as the local sites when determining a subject's appropriateness for inclusion in a trial.



Independent Interviews

	ICC	P Value
17-Item HAMD ²	.93	.000
HAMA ²	.91	.000
MADRS ³	.94	.000
PANSS ⁴	.86	.002
Atypical Symptoms ²	.95	.000

²Kobak & Williams, APA, 2006. 14 raters, 70 interviews (35 pairs). ³Williams & Kobak, 2006. 6 raters, 120 interviews (60 pairs) ⁴14 interviews (7 pairs);. Note: PANSS ICC is .95 when measured by rating videos.

Highest Reliability

Centralized raters can improve reliability by simply reducing the sheer number of raters involved; for example, a 30-site multicenter trial that used 60 to 75 raters could be conducted with 8 to 10 centralized raters. More importantly, with **one cohort** of centralized raters rigorous operating and uniform methodology procedures can be put in place – which is not logistically feasible with a larger group of raters distributed among unconnected study sites - to ensure regular, ongoing monitoring of both interview quality and calibration.

For example, MedAvante Centralized raters obtained ICCs of 0.90 and higher, using **independent interviews**, measured throughout the course of a large multi-center Phase III psychosis study – not just at study outset¹⁹.

Because MedAvante's centralized raters are focused exclusively on the task of conducting clinical

Scientific Rationale For MedAvante Assessment Services

assessments they can commit the time required to

Centralized raters did not show drift over course of studies:

ICCs of Observed Interviews over course of studies	Q1	Q2	Q3	Q4
HAMA (N=100) (study 1)	.90	.96	.95	.97
HAMA (N=68) (study 2)	.98	.97	.97	.96
PANSS (N=131) (study 1)	.90	.90	.96	.88
PANSS (N=67) (study 2)	.98	.97	.98	.98

maintain high levels of accuracy and reliability.

No Site-Specific Biases

MedAvante reduces rater biases to negligible levels by removing the assessment from the investigative site, physically separating the subject and rater, and minimizing the number of times a rater evaluates a specific subject. The single most critical structural elements of the solution are that centralized raters can be 1) **blinded** to the protocol criteria for inclusion/exclusion; blinded to visit number and patient history and 2) **remote from the site** permitting them to make objective evaluations that are not subject to any conscious or unconscious biases. Bias-free assessments are impossible to achieve if the rater is proximate to the investigative site, its employees or the subject themselves.

There are subtle but absolutely critical distinctions between a research trial assessment and a therapeutic intervention assessment which is geared towards eventually finding an appropriate therapy for a patient

making regular visits to the investigative site. Multiple forms of bias are inherent in the therapeutic method, and are main contributors to the high failure rate of CNS trials. MedAvante centralized raters are hired and trained exclusively for research assessment capabilities and experience. The only task for centralized raters is the disinterested evaluation and assessment of subject appropriateness for inclusion and maintenance in trials as specified in the sponsor's protocol.

MedAvante's centralized raters will deliver the patient population exactly as specified in the sponsor's protocol and maintain the correct population throughout the trial – without the bias introduced by working for an investigative site and seeing subjects regularly. This will result in the right population being screened in and the wrong subject candidates being screened out. This results in greater assay sensitivity for the sponsor.

In summary, Centralized remote raters, blind to visit order history and independent of the site, can overcome the biases due to being local and thus significantly increase signal detection. Baseline score inflation is driven by the raters' dependency on the site for employment or simply knowledge of the impact of a subject's screen failure on their site. There is simply no effective or feasible way to overcome the local bias except for remote administration of screening.

Impact of MedAvante Patient Selection on Signal Detection

Outcomes generated to date empirically demonstrate using MedAvante scores to qualify patients for enrollment results in significant increases in signal

Scientific Rationale For MedAvante Assessment Services

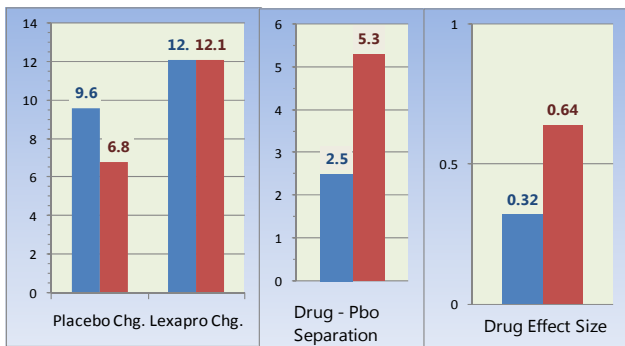
detection – in one case doubling the separation and effect size achieved with the active comparator escitalopram (Lexapro).²⁰ In this large multi-center generalized anxiety disorder study, using MedAvante to qualify patient enrollment doubled signal detection from 2.5 pts. drug-placebo chg to 5.3pts. Further analysis showed that the cohort of patients that MedAvante screen-failed (which were almost 57% of the total number the sites enrolled) contributed the placebo response that ultimately reduced the sites' signal detection to half that achieved by MedAvante central raters.

"The implications for sample size in a clinical trial are profound: if power analyses assumed these as population effect sizes (.32 vs. .64) the required sample size with centralized ratings would be reduced by greater than 50%."

Andrew C. Leon, Ph.D.
Professor of Biostatistics in Psychiatry and Professor of Public Health at Weill Medical College of Cornell University.

MedAvante provided the primary outcome measure and served as gatekeeper in two exacerbated Schizophrenia trials: the first was a proof of concept (Phase II) in which the MedAvante central raters were compared to site raters²¹. In that trial, while both site and central raters showed separation for the active comparator arm, olanzapine, versus placebo, the central raters detected signal of the test drug while the sites did not – site raters scored the test drugs no differently than the placebo. Central raters detected onset of action from the first week.

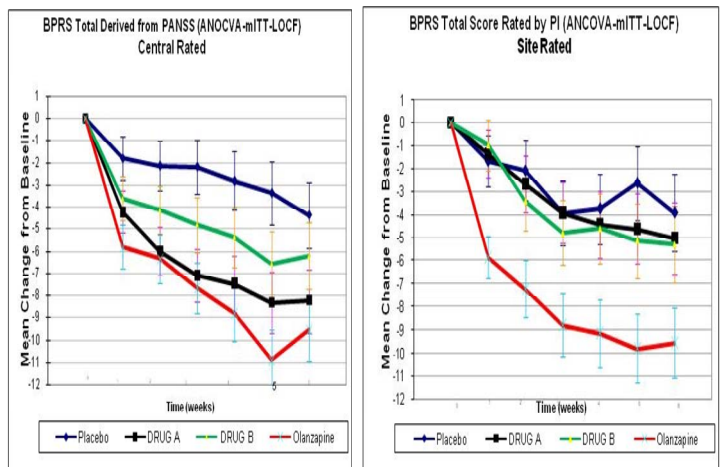
Signal Detection: Impact of Patient Selection GAD
Site vs. Central Rater Qualified



■ Site Raters Qualified & Rated ■ Central Raters Qualified & Rated

Importantly, using MedAvante to qualify and rate the patients doubled the drug effect size from 0.32 achieved by the site raters to 0.64 using central raters. This would mean, for example, for a sample size of 100 patients using site raters the sponsor could use 25 patients with MedAvante central raters and still achieve the same study power.

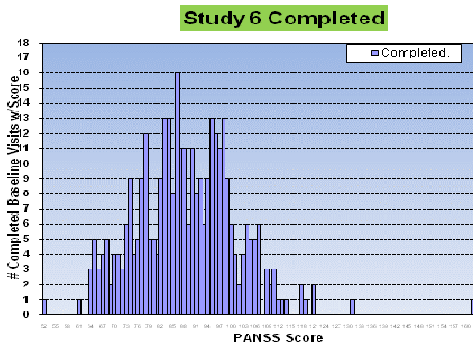
Psychosis Study: Effect of Central Raters



The second trial, a Phase III trial with over 300 patients in 35 US sites, formed the basis of a NDA submission using PANSS as the primary outcome measure and additional PSP and CGI-S ratings. The test drug showed a significant separation from the placebo arm. As in all MedAvante trials, the baseline scores exhibited

Scientific Rationale For MedAvante Assessment Services

a statistically normal distribution. At present, we are prevented by the sponsor from sharing any other data from this study.



Similarly, in a large multi-center trial of exacerbated Schizophrenia in which the investigative sites performed the diagnosis and were the gatekeepers for enrollment, MedAvante raters would have excluded 40% of the 205 patients the sites enrolled on the basis of symptom severity alone: MedAvante rated these site-enrolled patients lower on the PANSS than the bar set in the protocol for inclusion. Critically, MedAvante raters were not used for the Diagnosis: had they been, a large number of patients would have been further excluded as inappropriate despite adequate PANSS symptom severity scores (psychotic symptoms due to substance abuse, malingering etc.). This trial ultimately failed due

to extremely high placebo response that drowned out the drug-placebo difference.

MedAvante Operations Reliable & Transparent

MedAvante's operations team created standards that make the Centralized Ratings services a 'plug-in' to existing study research design. The objective is to ensure that adopting the MedAvante approach has minimal impact on other customer protocols. MedAvante assumes responsibility for delivering the assessment services at the convenience of the participating investigative sites. Depending upon the sponsor's preferences, MedAvante's system allows Operationally, the centralized ratings videoconferencing platform has performed reliably and has had greater than 99% uptime without interruption: of the remote videoconference interviews, 0.9% of these calls experienced a temporary interruption of its audio/visual quality [flutter, temporary pixilation etc.] issue but was resolved and the call was completed. In total, 0.2% of the videoconference calls ever missed or rescheduled for technical reasons.

MedAvante has had 9 sponsor audits in the past two and half years and has had virtually flawless results.

Scientific Rationale For MedAvante Assessment Services

Working with MedAvante also allows access to MedAvante staff that includes thought leaders who offer their expertise in working out details of the protocol design.

MedAvante Medical Leadership & Expertise

Michal Detke, M.D., Ph.D.	Chief Medical Officer	Board Certified Psychiatrist; 9 years of pharma experience in all phases of CNS drug development (e.g. Cymbalta; Prozac). Extensive methodology research experience. Member, ACNP.
Janet Williams, D.S.W.	Vice President Clinical Development	Co-author of Structured Clinical Interview for DSM-IV (SCID) as well as author or co-author of numerous interview guides for psychiatric ratings scales. Honorary Fellow of APA for her contributions to the field. Member, ACNP.
John Kane, M.D., Ph.D.	Chair, Scientific Advisory Board	Chairman of Psychiatry at The Zucker Hillside Hospital. Professor of Psychiatry, Neurology and Neuroscience Albert Einstein College of Medicine. Member, ACNP.
Kenneth Kobak Ph.D.	MedAvante Senior Scientist	Author of SIGMA and current IVR versions of common psychiatric scales
Scott Reines, M.D.	MedAvante Senior Scientist	Former Senior Vice president of CNS and Pain, Johnson and Johnson
Earl Giller, M.D., Ph.D.	MedAvante Senior Scientist	Former Executive Director of Psychiatry, Pfizer, Inc.
Madeline Alexander, Ph.D.	MedAvante Senior Scientist	Expert in rater training and calibration and clinical assessment of mental health using clinician administered scales.
Andrew C. Leon, Ph.D.	MedAvante Advisor	Professor of Biostatistics in Psychiatry and Professor of Public Health at Weill Medical College of Cornell University.

Scientific Rationale For MedAvante Assessment Services

References

1. Khan A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the food and drug administration database. *Journal of Clinical Psychopharmacology*. 2002;22(1):40-45.
2. Muller MJ, Szegedi A. Effects of inter-rater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *Journal of Clinical Psychopharmacology*. 2002;22(3):318-325.
3. Leon AC, Marzak PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Archives of General Psychiatry*. 1995;52(10):867-871.
4. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry*. Apr 15 2000;47(8):762-766.
5. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Inter-rater Reliability in Clinical Trials of Depressive Disorders. *Am J Psychiatry*. September 1, 2002 2002;159(9):1598-1600.
6. Demitrack MA, Faries D, Herrera JM, DeBrotta DJ, Potter WZ. The problem of measurement error in multisite clinical trials. *Psychopharmacology Bulletin*. 1998;34:19-24.
7. Kobak KA, Kane JM, Thase ME, Nierenberg AA. Why Do Clinical Trials Fail?: The Problem of Measurement Error in Clinical Trials: Time to Test New Paradigms? *J Clin Psychopharmacology*. Feb 2007;27(1):1-5.
8. 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. Paper presented at: National Institute of Mental Health, New Clinical Drug Evaluation Unit, 39th Annual Meeting, 1999; Boca Raton, FL.
9. Feltner DE, Kobak KA, Crockatt J, et al. Interactive Voice Response (IVR) for Patient Screening of Anxiety in a Clinical Drug Trial. Paper presented at: National Institute of Mental Health, New Clinical Drug Evaluation Unit, 41st Annual Meeting, 2001; Phoenix, AZ.
10. Kobak KA, Taylor LV, Bystritsky A, et al. St John's wort versus placebo in obsessive-compulsive disorder: results from a double-blind study. *Int Clin Psychopharmacology*. Nov 2005;20(6):299-304.
11. Posternak MA, Zimmerman M. The therapeutic effect of follow-up assessments on the placebo response in antidepressant efficacy trials. Paper presented at: American Psychiatric Association, 158th Annual Meeting; May, 2005, 2005; Atlanta, GA.
12. Quinn J, Moore M, Benson DF, et al. A videotaped CIBIC for dementia patients. Validity and reliability in a simulated clinical trial. *Neurology*. 2002;58:433-437.

Scientific Rationale For MedAvante Assessment Services

13. DeBrotta D, Gelwicks S, Potter W. Same rater versus different raters in depression clinical trials. Paper presented at: 42nd Annual Meeting, New Clinical Drug Evaluation Unit; June 10-13, 2002, 2002; Boca Raton, FL.
14. Smith W, Londborg P, Bielski RJ, Bourg W, Painter J, Glaudin V. Multiple Raters in Clinical Trials: A Better Research Design? Paper presented at: NCDEU, 44th Annual Meeting, 2004; Phoenix, AZ.
15. 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.
16. Feiger A, Engelhardt N, DeBrotta D, et al. Rating the raters: an evaluation of audio taped Hamilton Depression Rating Scale (HAMD) interviews. Paper presented at: National Institute of Mental Health, New Clinical Drug Evaluation Unit, 43rd Annual Meeting, 2003; Boca Raton, FL.
17. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *Am J Psychiatry*. Mar 2005;162(3):628.
18. Lipsitz J, Kobak KA, Feiger A, Sikich D, Moroz G, Engelhardt N. The Rater Applied Performance Scale (RAPS): Development and Reliability. *Psychiatry Research*. 2004;127:147-155.
19. Kane JM. Centralized Ratings: What have we learned after 21,000 Assessments? National Institute of Mental Health, New Clinical Drug Evaluation Unit, 44th Annual Meeting. Phoenix, AZ; 2008.
20. Coric V, Goddard A, Oren D, et al. A Randomized, Double-Blind, Placebo-Controlled and Active Comparator Trial of Pexacerfont, a Corticotropin Releasing Factor Receptor-1 Antagonist, in the Treatment of Generalized Anxiety Disorder. American College of Neuropsychopharmacology 47th Annual Meeting. Scottsdale, AZ; 2008.
21. Shen J, Kobak KA, Zhao Y, Alexander MM, Kane JM. Use of remote centralized raters via live 2-way video in a multicenter clinical trial for schizophrenia. *J Clin Psychopharmacology*. Dec 2008;28(6):691-693.