

Why Do Clinical Trials Fail?

The Problem of Measurement Error in Clinical Trials: Time to Test New Paradigms?

Kenneth A. Kobak, PhD, John M. Kane, MD,†‡ Michael E. Thase, MD,§||
and Andrew A. Nierenberg, MD¶#*

A recent review¹ of the Food and Drug Administration data sets including 45 trials found that 36% of these studies failed to show superiority of a standard antidepressant over placebo. (In the case of new antidepressants, the failure rate was 52%). In addition, it has become increasingly difficult 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.² This increasingly high rate of failed central nervous system trials, as well as the smaller effect sizes observed in positive studies, continues to receive much attention. We believe that poor reliability of measurement, poor interview quality (which extends to evaluation and diagnosis), and rater bias are largely to blame for this problem and that this problem can be remedied. Although clinician-administered rating scales form the building blocks upon which the entire clinical trial rests, only recently have factors associated with the clinical assessment process come under closer scrutiny as a possible source of trial failure and poor signal detection. As Fleiss has said, “The most elegant design of a clinical study will not overcome the damage caused by unreliable or imprecise measurement.”³ A variety of studies illustrate how factors associated with clinician ratings (ie, interrater reliability, interview quality, and rater bias) can significantly impact signal detection and study outcome. In what follows, we consider these problems, which have been the object of recent research. Questions around diagnostic validity and/or shifts in patient populations remain to be evaluated.

PROBLEMS

Problem 1: Poor Interrater Reliability

Several researchers have modeled statistically the impact of reliability on power and sample size requirements.⁴⁻⁶ For example, Perkins and colleagues have shown that improving the intraclass correlation coefficient (ICC) from 0.70 to 0.90 would decrease sample size requirements by 22%.⁶ A drop in ICC from 0.90 to 0.70 would reduce study power from 0.72 to 0.50. This has enormous clinical and economic implications. In addition to reducing the chance for a type II error, 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. This would result in fewer patients being exposed to placebo or experimental medication. In addition, at an estimated cost of \$15,000 per patient, this would save the sponsor \$1,100,000. The costs associated with faster time to market provide additional economic incentives (it has been estimated that each day’s delay to market results in an average of 1.3 million in lost prescription sales⁷), not to mention the more rapid availability of truly effective treatments to patients.

*MedAvante, Inc, Madison, WI; †The Zucker Hillside Hospital, Glen Oaks, NY; ‡The Albert Einstein College of Medicine, Bronx, NY; §University of Pittsburgh Medical Center; ||Western Psychiatric Institute and Clinic, Pittsburgh, PA; ¶Massachusetts General Hospital and #Harvard Medical School, Boston, MA.

Dr Kenneth Kobak is an employee of MedAvante. Dr John Kane is Chief Scientific Advisor of MedAvante. Dr Michael Thase is on the scientific advisory board of MedAvante.

Address correspondence and reprint requests to Kenneth A. Kobak, PhD, MedAvante Inc, 7601 Ganser Way, Madison, WI 53719. E-mail: kkobak@medavante.net.

Copyright © 2007 by Lippincott Williams & Wilkins

ISSN: 0271-0749/07/2701-0001

DOI: 10.1097/JCP.0b013e31802eb4b7

Given the importance of interrater reliability on clinical trial outcomes, what is the current state of interrater reliability in clinical trials? For the most part, this is largely unknown. Mulsant and colleagues⁸ found that only 3 of 63 published studies reported reliability figures. In addition, on the few occasions when reliability is established, it is usually done through observation and scoring of videotapes, which artificially inflate estimates of reliability by reducing the information variance that would result if each rater interviewed the patient independently.⁹ Even with this inflated method, positive results have been difficult to obtain; for example, Demitrack et al¹⁰ found no evidence of improved rating performance after 6 hours of reliability training. Although more recent rater training efforts have been successful in improving interview *quality* at study onset,^{11,12} there are no reports in the literature documenting improved interrater reliability as a result of rater training in multicenter trials using independent interviews.

Problem 2: 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 with a patient, on signal detection. In the first study,¹³ all baseline Hamilton Rating Scale for Depression (HAM-D) interviews in a large phase II trial were audiotaped, and a random sample of 25% were blindly evaluated for interview quality using the Rater Applied Performance Scale (RAPS).¹⁴ Without taking interview quality into account, the trial failed to separate active drug (paroxetine) from placebo (mean difference on the HAM-D total score = 0.5, $P = 0.614$). However, when the analysis was limited to only those subjects with a mean rating of "good" or "excellent" on the RAPS scale, a large and statistically significant effect was found favoring paroxetine over placebo (mean difference = 6.83, $P = 0.017$). This is particularly striking because the good or excellent ratings involved only 22 patients (10% of the total sample). The effect size of paroxetine among the subjects with the good or excellent ratings was 1.33 (by contrast, a recent meta-analysis found that the mean effect size of approved SSRIs was 0.46).¹⁵

In the second study,¹⁶ a sample of 77 tapes from 13 sites were blindly evaluated for interview quality with the RAPS scale (active drug = 42, placebo = 35). For the total sample, there was no significant difference between active drug and placebo ($p = 0.197$). Similarly, only a minority ($n = 30$, 39%) of the interviews received good or excellent ratings. And again, ratings of patients with HAM-D interviews with excellent or good clarification skills did result in separation from placebo ($p = 0.014$).

Given the critical importance of interview quality, the question becomes what is the current state of interview quality in clinical trials. Although few studies have examined this issue, in the 2 studies cited earlier,^{13,17} more than half of the ratings were of fair or poor quality on all dimensions of the RAPS scale (ie, adherence, clarification, follow up, and neutrality).¹³ These results are particularly striking considering that the individuals conducting these rating inter-

views were aware that they were being audiotaped. Forty-five percent of the interviews were done in less than 10 minutes (range, 3–35 minutes) despite Hamilton's suggestion that the interview should take at least half an hour.¹⁸

Interview quality is likely related to the educational background and to the amount and quality of prior training and clinical experience using the scale. Ideally, raters should have didactic training in psychopathology, clinical experience with the patient population being evaluated, and scale-specific expertise.¹⁹ Recent studies however have shown that 24.9% of raters have had no prior experience with the selected primary outcome measure²⁰ and that only 38% were ever observed administering the scale to patients before rating patients in the trial.²¹

Problem 3: Rater Bias

Rater bias can manifest itself both at screening and during postrandomization trial visits. Several studies have found that patient's self-ratings were discordant with clinician ratings at screening and baseline and then coalesce after randomization. The most likely explanation for this phenomenon is that clinical evaluators tend to consciously or unconsciously inflate ratings before randomization to ensure that subjects are eligible to enter the trial, whereas subjects are unlikely to do so because they are not aware of the study's entry criteria. Consistent with this explanation, in a study that required a minimum HAM-D score of 20, DeBrotta et al²² found a relatively normal distribution of self-report HAM-D scores at baseline in comparison with the skewed distribution of clinician scores. Only 4 of the 285 clinician scores were less than 20 at baseline, whereas 110 of the patients' self-report scores were less than 20. Clinician and self-report scores became more concordant once patients were participating in the randomized trial.

Similar results were found by Feltner et al²³ in a relapse prevention study of generalized anxiety disorder that used the Hamilton Rating Scale for Anxiety (HAM-A). Specifically, they found that those subjects who were above the threshold at baseline on *both* clinician and self-report HAM-As had a significantly greater relapse rate when blindly switched to placebo after an open-label phase than those who simply met criteria on the clinician HAM-A alone at baseline.

Enrollment bias can also occur in the other direction. In a multicenter study of obsessive-compulsive disorder by Kobak et al,²⁴ subjects were excluded if they scored *above* 16 on the clinician HAM-D. Thus, per protocol, all of the enrolled patients scored 16 or less on the clinician HAM-Ds. However, 27% of the patients scored 17 or higher on a self-administered, paper-and-pencil version of HAM-D²⁵ completed at the same time.

If these studies are representative, the implications are clear: rater biases result in overinclusion of less severely symptomatic subjects and failure to exclude a significant number of potential participants who are not eligible for study participation.

More recently, this discrepancy has been replicated in a study using blinded clinical ratings.²⁶ Patients were

interviewed twice at each of 3 time points: screening, baseline, and end point, once by the site rater and once remotely via videoconference by a “centralized” rater who was blind to study visit and design. The HAM-D ratings completed by the on-site evaluator were significantly higher than the centralized raters’ scores at screening and baseline, but not at end point. At screening, 36% of patients who were judged to be eligible for the study by the on-site evaluator (ie, they scored at least 17 on the HAM-D, the study’s minimum severity criterion) were rated as study ineligible (ie, HAM-D scores of 16 or lower) by a centralized rater.

Inflated baseline scores have a critical impact on signal detection, as higher pretreatment HAM-D scores are associated with greater change with antidepressants, whereas lower baseline scores are associated with greater change with placebo.¹ Including subjects with inflated scores thus results in enrolling a higher proportion of patients who are likely to be placebo responders. Moreover, as score inflation seems to dissipate rapidly after subjects begin double-blind therapy, measures of change from baseline will be distorted, which will increase error variance and reduce drug-placebo effect sizes, thus decreasing signal detection.

Rater bias also can adversely affect ratings during the course of the study. Referred to as expectancy bias, clinical raters and patients generally will expect to see improvement over time rather than no change or worsening. Rater expectancy bias seems 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. Counterintuitively (given the problem of controlling interrater 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.^{27–30} This has been found across disorders (depression, obsessive-compulsive disorder, generalized anxiety disorder, panic, and social anxiety disorders) and across different rating scales.

What is remarkable about finding an advantage for using multiple raters is that a single rater is likely to perform ratings with higher test-retest reliability than a pair of raters or multiple raters. Expectancy bias thus must be a larger adverse effect on signal detection than whatever is gained by the higher reliability of a single rater. Clearly, the use of different raters would require a well-established and well-maintained interrater reliability as a prerequisite. However, addressing these biases can have profound effects: in one of the studies reviewed earlier, the best signal detection occurred in patients who had good-to-excellent ratings at both baseline and end point performed by different evaluators.¹³

SOLUTIONS

There have been several strategies proposed to address these methodological issues, including better rater selection, training, and monitoring;¹⁹ design modifications;^{31,32} improved rating scales;³³ development of standardized assessment procedures;³⁴ and (in an article we wrote in this column

5 years ago) the use of self-report (interactive voice response; IVR) measures.^{35–37} Although each of these solutions has potential benefits for improved signal detection, they are not without costs or limitations. For example, as reviewed earlier, rater training before the trial does not guarantee that the standards will be maintained during the course of the trial. It also does not protect against the biases that occur at baseline or during the course of the study. Ongoing audiotape monitoring may help correct rater drift, but only does so “after the fact.” Thus, unless rating monitoring is frequent, an unknown proportion of assessments will be suboptimal. Self-rating (IVR) has the potential for improved screening and has recently shown equivalence to clinicians in sensitivity to change in open-label and comparison studies,^{38,39} but as yet has not shown superiority to clinicians in terms of signal detection in placebo-controlled trials.^{40–42} Moreover, self-report ratings are limited for selected symptoms/signs such as psychomotor disturbance, loss of insight, and psychosis. Results of one meta-analysis comparing the HAM-D and the Beck Depression Inventory suggest that reliance on patient ratings may come at the price of a decrease in sensitivity to detect change.⁴³

An Additional Potential Solution: Centralized Raters

The advent of new technologies brings with it the possibility of new solutions to old problems. As suggested by the findings of several studies reviewed earlier, one of these possible solutions 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 videoconferencing or teleconferencing and remotely administer the primary outcome measure to study patients during their regularly scheduled study visit.

Centralized raters can address the issues raised in the preceding summary in a number of ways, as discussed in the next section.

Reliability and Quality

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 (ie, 2 or 3 raters per site) could be conducted with 8 to 10 centralized raters. Centralized raters can be calibrated using rigorous methodological procedures that are not logistically feasible with a larger group of raters at diffuse study sites. For example, in one group of centralized raters, ICCs of 0.90 and higher were obtained on the HAM-D, HAM-A, and Montgomery-Asberg Depression Rating Scale using independent interviews.⁴⁴ Because centralized raters are focused exclusively on the task of conducting clinical assessments, they can commit the time and ongoing monitoring required to maintain high levels of accuracy. Rigorous standard operating procedures can be put in place to ensure regular ongoing monitoring of both interview quality and calibration.

Raters can be hired based primarily on their clinical assessment experience, as opposed to operational or study coordination expertise.

Bias

Because centralized raters are divorced from the study site, there is no pressure to enroll patients, thus eliminating the possibility of inflated baseline scores. Blinding to study visit, protocol requirements, and study design will likewise minimize expectancy or other biases at later visits. A centralized pool of different raters also could ensure that subjects receive a truly independent evaluation at each study visit by eliminating the tendency to rate patients based on prior ratings (vs conducting a complete, thorough, and independent evaluation at each visit).

Another advantage of using different raters is minimizing the potentially confounding therapeutic impact of repeated assessment by the same clinician. Although the process of performing a HAM-D or HAM-A scale is not inherently psychotherapeutic, the value of repeated contact with a caring professional should not be underestimated; and, in our experience, study participants not infrequently misidentify their rater as a therapist. Posternak and Zimmerman⁴⁵ found that each additional follow-up visit during a 6- to 8-week clinical trial resulted in an additional reduction of about 1 point on the HAM-D. This was true for patients on both drug and placebo. They estimate the therapeutic impact of a repeated assessment accounts for about 40% of the placebo response.⁴⁵

Before centralized raters can be used, it needs to be demonstrated that administering a scale remotely via videoconference or teleconference yields equivalent results as the same scale administered face-to-face. Studies have found equivalence between face-to-face and remote (videoconference) administration of the HAM-D,^{46–48} HAM-A,⁴⁷ Yale-Brown Obsessive-Compulsive Scale,⁴⁷ and Brief Psychiatric Rating Scale.⁴⁹ More recently, a meta-analysis found no difference in effect size or subject satisfaction between psychiatric assessments administered by video and those conducted face-to-face.⁵⁰

Similarly, studies have also found high correlations between scales administered by a clinician over the telephone and those administered face-to-face.^{51–53} The National Institute of Mental Health–sponsored Sequenced Treatment Alternatives to Relieve Depression study⁵⁴ used centralized raters who conducted HAM-D interviews over the telephone as the primary outcome measure, with site raters used for ongoing clinical management of patients (site and centralized raters were blind to each others scores). Although there has been some debate as to the incremental value of video over audio alone (particularly in the case of scales that do not involve observational items), video may create a “social presence” that enhances the interaction, improves patient satisfaction and acceptance, and—at least in principle—improves assessment of signs (as opposed to symptoms) of affective and behavioral disturbance.⁵⁵ Patient satisfaction with assessments conducted by videoconference has been generally high across a variety of scales and disorders.^{46,47,49,56–59}

Centralized raters could also be used to evaluate and recruit subjects for clinical trials from nonpsychiatric settings, such as primary care, workplace, etc. Primary care patients tend to have a lower placebo response,⁶⁰ may be more likely to be treatment naive, and provide results more generalizable to the population at large.

In conclusion, there is a growing problem with failed trials, and a number of issues associated with clinician-administered rating scales have been identified that can contribute to this problem. A new emphasis on methodological research is warranted to empirically examine a variety of potential solutions, including centralized raters, as well as other innovative approaches. As we let the data guide us, we will begin to weigh the potential merits and the potential limitations of this and other novel approaches. The time and cost associated with these efforts, although not trivial, are worth the investment considering the cost of failed trials in both human and economic terms.

REFERENCES

1. Khan A, Leventhal RM, Khan SR, et al. Severity of depression and response to antidepressants and placebo: an analysis of the food and drug administration database. *J Clin Psychopharmacol.* 2002;22:40–45.
2. Thase ME. Studying new antidepressants: if there were a light at the end of the tunnel, could we see it? *J Clin Psychiatry.* 2002;63(suppl 2) 24–28.
3. Fleiss JL. *The Design and Analysis of Clinical Experiments.* New York: Wiley; 1986.
4. Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *J Clin Psychopharmacol.* 2002;22:318–325.
5. Leon AC, Marzak PM. More reliable outcome measures can reduce sample size requirements. *Arch Gen Psychiatry.* 1995;52:867–871.
6. 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.* 2000;47:762–766.
7. Getz KA, de Bruin A. Breaking the development speed barrier: assessing successful practices of the fastest drug development countries. *Drug Inf J.* 2000;34:725–736.
8. Mulsant BH, Kastango KB, Rosen J, et al. Interrater reliability in clinical trials of depressive disorders. *Am J Psychiatry.* 2002;159:1598–1600.
9. Spitzer RL, Williams JBW. *Classification in Psychiatry.* Baltimore, MD: Williams & Wilkins; 1980.
10. Demitrack MA, Faries D, Herrera JM, et al. The problem of measurement error in multisite clinical trials. *Psychopharmacol Bull.* 1998;34:19–24.
11. Kobak KA, Engelhardt N, Lipsitz JD. Enriched rater training using Internet based technologies: a comparison to traditional rater training in a multi-site depression trial. *J Psychiatr Res.* 2006;40:192–199.
12. Kobak KA, Lipsitz JD, Williams JB, et al. A new approach to rater training and certification in a multicenter clinical trial. *J Clin Psychopharmacol.* 2005;25:407–412.
13. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *Am J Psychiatry.* 2005;162:628
14. Lipsitz J, Kobak K, Feiger A, et al. The Rater Applied Performance Scale: development and reliability. *Psychiatry Res.* 2004;127:147–155.
15. Joffe R, Sokolov S, Streiner D. Antidepressant treatment of depression: a metaanalysis. *Can J Psychiatry.* 1996;41:613–616.
16. Feiger A, Engelhardt N, DeBrotta D, et al. Rating the raters: an evaluation of audiotaped Hamilton Depression Rating Scale (HAMD) interviews. 43rd Annual Meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit; Boca Raton, FL; 2003.
17. Engelhardt N, Feiger AD, Cogger KO, et al. Rating the raters: assessing the quality of Hamilton rating scale for depression clinical interviews in

- two industry-sponsored clinical drug trials. *J Clin Psychopharmacol*. 2006;26:71–74.
18. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol*. 1967;6:278–296.
 19. Kobak KA, Engelhardt N, Williams JB, et al. Rater training in multicenter clinical trials: issues and recommendations. *J Clin Psychopharmacol*. 2004;24:113–117.
 20. Bullinger A, Targum SD. Rater experience in CNS clinical trials. 44th Annual Meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit; Phoenix, AZ; 2004.
 21. Kobak KA, Engelhardt N. Standardized training on the Hamilton Depression Scale using Internet-based technologies. 39th Annual Meeting, Drug Information Association; San Antonio, TX; 2003.
 22. DeBrotta D, Demitrack M, Landin R, et al. A comparison between interactive voice response system-administered HAM-D and clinician-administered HAM-D in patients with major depressive episode. 39th Annual Meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit; Boca Raton, FL; 1999.
 23. Feltner DE, Kobak KA, Crockett J, et al. Interactive voice response (IVR) for patient screening of anxiety in a clinical drug trial. 41st Annual Meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit; Phoenix, AZ; 2001.
 24. 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 Psychopharmacol*. 2005;20:299–304.
 25. Reynolds WM, Kobak KA. Reliability and validity of the Hamilton Depression Inventory: a paper-and-pencil version of the Hamilton Depression Rating Scale clinical interview. *Psychol Assess*. 1995;7:472–483.
 26. Kobak KA, DeBrotta DJ, Engelhardt N, et al. Site vs. centralized raters in a clinical depression trial. 46th Annual Meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit; 2006; Boca Raton, FL.
 27. 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.
 28. DeBrotta D, Gelwicks S, Potter W. Same rater versus different raters in depression clinical trials. 42nd Annual Meeting, New Clinical Drug Evaluation Unit; Boca Raton, FL; 2002.
 29. Smith W, Londborg P, Bielski RJ, et al. Multiple raters in clinical trials: a better research design? 44th Annual Meeting, New Clinical Drug Evaluation Unit; Phoenix, AZ; 2004.
 30. Glaudin V, Smith W, Ferguson J, et al. Discriminating placebo and drug in generalized anxiety disorder (GAD) trials: single vs. multiple raters. *Psychopharmacol Bull*. 1994;32:175–178.
 31. Evans KR, Sills T, Wunderlich GR, et al. Worsening of depressive symptoms prior to randomization in clinical trials: a possible screen for placebo responders? *J Psychiatr Res*. 2004;38:437–444.
 32. Landin R, De Brota DJ, DeVries A, et al. The impact of restrictive entry criterion during the placebo lead-in period. *Biometrics*. 2000;56:271–278.
 33. Bagby RM, Ryder AG, Schuller DR, et al. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *Am J Psychiatry*. 2004;161:2163–2177.
 34. Bech P, Engelhardt N, Evans K, et al. A proposal for a standardized HAMD scoring system: a collaboration among the pharmaceutical industry, academia, and government. 41st Annual Meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit; Phoenix, AZ; 2001.
 35. Greist JH, Mundt JC, Kobak K. Factors contributing to failed trials of new agents: can technology prevent some problems? *J Clin Psychiatry*. 2002;63(suppl 2):8–13.
 36. Kobak KA, Greist JH, Jefferson JW, et al. Computerized assessment of depression and anxiety over the telephone using interactive voice response. *MD Comput*. 1999;16:64–68.
 37. Kobak KA, Greist JH, Jefferson JW, et al. Computer-administered clinical rating scales. A review. *Psychopharmacology (Berl)*. 1996;127:291–301.
 38. Rush AJ, Bernstein IH, Trivedi MH, et al. An evaluation of the quick inventory of depressive symptomatology and the Hamilton rating scale for depression: a sequenced treatment alternatives to relieve depression trial report. *Biol Psychiatry*. 2006;59:493–501.
 39. Rush AJ, Trivedi MH, Carmody TJ, et al. Self-reported depressive symptom measures: sensitivity to detecting change in a randomized, controlled trial of chronically depressed, nonpsychotic outpatients. *Neuropsychopharmacology*. 2005;30:405–416.
 40. Fava M, McCall WV, Krystal A, et al. Eszopiclone co-administered with fluoxetine in patients with insomnia coexisting with major depressive disorder. *Biol Psychiatry*. 2006;59:1052–1060.
 41. Corruble E, Legrand JM, Zvenigorowski H, et al. Concordance between self-report and clinician's assessment of depression. *J Psychiatr Res*. 1999;33:457–465.
 42. GlaxoSmithKline. A multicenter, double-blind, placebo-controlled comparison of the efficacy and safety of flexible dose extended-release bupropion hydrochloride (HCl) 300–450 mg/day and placebo administered for eight weeks for the treatment of adult outpatients with major depressive disorder including symptoms of decreased energy, pleasure, and interest [GlaxoSmithKline Clinical Trials Register Web site]. September 20, 2005. Available at: http://ctr.gsk.co.uk/Summary/bupropion/IV_WELL_AK130931.pdf. Accessed November 7, 2006.
 43. Lambert MJ, Hatch DR, Kingston MD, et al. Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: a meta-analytic comparison. *J Consult Clin Psychol*. 1986;54:54–59.
 44. Kobak KA, William JBW. Development and reliability of a combined Hamilton depression, anxiety, and atypical symptoms scale. 159th Annual Meeting, American Psychiatric Association; Toronto, CA; 2006.
 45. Posternak MA, Zimmerman M. The therapeutic effect of follow-up assessments on the placebo response in antidepressant efficacy trials. 158th Annual Meeting, American Psychiatric Association; Atlanta, GA; 2005.
 46. Kobak KA. A comparison of face-to-face and videoconference administration of the Hamilton Depression Rating Scale. *J Telemed Telecare*. 2004;10:231–235.
 47. 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.
 48. Menon S, Kondapavala P, Krishna P, et al. Evaluation of a portable low-cost videophone system in the assessment of depressive symptoms and cognitive function in elderly medically ill veterans. *J Nerv Ment Dis*. 2001;189:399–401.
 49. Zarate CA Jr, Weinstock L, Cukor P, et al. Applicability of telemedicine for assessing patients with schizophrenia: acceptance and reliability. *J Clin Psychiatry*. 1997;58:22–25.
 50. Hyler SE, Gangure DP, Batchelder ST. Can telepsychiatry replace in-person psychiatric assessments? A review and meta-analysis of comparison studies. *CNS Spectr*. 2005;10:403–413.
 51. Simon GE, Revicki D, von Korff M. Telephone assessment of depression severity. *J Psychiatr Res*. 1993;27:247–252.
 52. Aneshensel CS, Frerichs RR, Clark VA, et al. Measuring depression in the community: a comparison of telephone and personal interviews. *Public Opin Q*. 1982;46:110–121.
 53. Rohde P, Lewinsohn PM, Seeley JR. Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. *Am J Psychiatry*. 1997;154:1593–1598.
 54. Rush AJ, Fava M, Wisniewski SR, et al. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials*. 2004;25:119–142.
 55. Cukor P, Baer L, Willis BS, et al. Use of videophones and low-cost standard telephone lines to provide a social presence in telepsychiatry. *Telemed J*. 1998;4:313–321.
 56. Dongier M, Tempier R, Lalinec-Michaud M, et al. Telepsychiatry: psychiatric consultation through two-way television. A controlled study. *Can J Psychiatry*. 1986;31:32–34.
 57. Frueh BC, Deitsch SE, Santos AB, et al. Procedural and methodological issues in telepsychiatry research and program development. *Psychiatr Serv*. 2000;51:1522–1527.
 58. Kennedy C, Yellowlees P. The effectiveness of telepsychiatry measured using the Health of the Nation Outcome Scale and the Mental Health Inventory. *J Telemed Telecare*. 2003;9:12–16.
 59. Dranov P. Telemedicine. *Sci Dig*. 1981;58:112–113, 120.
 60. Schweizer E, Rickels K. Placebo response in generalized anxiety: its effect on the outcome of clinical trials. *J Clin Psychiatry*. 1997;58(suppl 11):30–38.