Factors Contributing to Failed Trials of New Agents: Can Technology Prevent Some Problems?

John H. Greist, M.D.; James C. Mundt, Ph.D.; and Kenneth Kobak, Ph.D.

For psychiatry to contribute to the development of the next generation of antidepressant pharmacotherapies, effective use of clinical trial methods is as critical as innovation in neurochemical research. Results from clinical trials on the efficacy of a new drug can be obscured by methodological problems. Accurate diagnosis and precise measurement of the clinical symptoms during conduct of the clinical trials are crucial to obtaining interpretable outcomes. As tools that reliably diagnose disorders and assess symptoms become available, computer administration of rating instruments may improve the accuracy of clinical trial results. This article describes methodological factors that can confound study outcomes and discusses the potential for interactive voice response (IVR) technology to address some of these problems.
(J Clin Psychiatry 2002;63[suppl 2]:8–13)

New antidepressants are needed because people respond in different ways to the same drug: some respond favorably, some respond but experience intolerable adverse events, and others fail to respond at all. Too often, clinical trials of new agents produce inconsistent and uninterpretable findings. Results from clinical trials on the efficacy of new drugs can be obscured by various methodological problems. Consequently, patients may have to suffer through delays in the development of new treatments, researchers and physicians may become doubtful regarding drug efficacy, and pharmaceutical companies may lose potential revenue that could be used to further other drug development or manufacture. Therefore, for the field of psychiatry to move into the next generation of antidepressant pharmacotherapy, the use of careful clinical trial methodology is as critical as innovation in neurochemical research. This article describes methodological factors that can confound study outcomes and discusses how the use of interactive voice response (IVR) technology might prevent many of these methodological problems.

From Healthcare Technology Systems, Inc., Madison, Wisc.
Presented at the scientific symposium "New Antidepressants: Light at the End of the Tunnel?" which was held May 10, 2001, at the 154th annual meeting of the American Psychiatric Association in New Orleans, La., and sponsored by an unrestricted educational grant from Eli Lilly and Company.
Reprint requests to: John H. Greist, M.D., Healthcare Technology Systems, Inc., 7617 Mineral Point, Suite 300, Madison, WI 53717 (e-mail: greij@healthtechsys.com).

FACTORS CONTRIBUTING TO FAILED TRIALS

Many factors may contribute to trials that do not separate investigational medicines from placebo treatment, one being use of an ineffective medication. In this instance, trial failure to separate active drug from placebo would not reflect an inaccurate trial outcome. However, even an effective medication may not demonstrate therapeutic potency if the trial has methodological problems. Many types of methodological design flaws can cause inaccurate trial results (Table 1).

Unreliable Rating Instruments

Using clinical instruments with reliable measurement properties is of foremost importance in a trial of a new agent. Accuracy of diagnostic assessment is crucial, as is measurement precision of the clinical symptoms that will define the efficacy of the medication being tested. Standardized diagnostic assessments that can be easily and reliably administered in multisite clinical trial settings have become generally available, such as the Primary Care Evaluation of Mental Disorders (PRIME-MD).[1]

Poor Subject Selection

Selection of an appropriate treatment sample is another factor that is crucial to the success or failure of a trial. Transportation, medication, and other costs are often paid to trial participants, so certain people make a career of participating in studies, which can influence the generalizability of results. Advertising to recruit subjects can become akin to "rounding up the usual suspects." Entry criteria should be met legitimately. Careful screening should be

Table 1. Factors Contributing to Failed Trials

Unreliable rating instruments
Poor subject selection (treatment resistance vs placebo response)
High dropout rates
Inadequate medication dose
Inadequate trial duration
Functional unblinding

done before allowing people to enter a study, despite the pressures that investigational sites may feel to enroll subjects quickly. Unfortunately, competition between sites can bias investigators (consciously or subconsciously) toward allowing inappropriate subjects into a trial. Inappropriate subjects may also be allowed into a trial because it is difficult for a caring clinician to exclude a suffering patient, knowing that trial participation may be that person's only means to receive treatment. The clinician's compassion may permit comorbidity or treatment resistance to be overlooked. Such patients may produce a lesser response to effective treatments than would be found with a homogeneous sample of patients meeting protocol inclusion and exclusion criteria.

Woggon[2] reminded investigators that, in a trial of the efficacy of a new antidepressant, only patients with recognized depressive disorder of at least moderate severity should be included but that efficacy can best be demonstrated in non–treatment-resistant patients. Patients with very mild forms of illness may respond to placebo, making separation of a treatment effective for more severe depression more difficult. Subjects with mild depression may also find the frequent clinical interaction associated with the trial to be therapeutic because it makes them feel special (the Hawthorne effect), or they want to please the clinician who is trying to help them. When this happens, their response is difficult to ascribe to the drug. Schatzberg and Kraemer[3] noted that concern has arisen recently regarding placebo response rates in antidepressant studies, which may erode clinicians' confidence in the drugs being studied.

High Dropout Rates

A high dropout rate can also contribute to poor clinical trial results. There should be enough subjects in the treatment sample to reliably infer statistical significance (i.e., provide adequate statistical power). Even with trials using endpoint analyses, a high dropout rate results in fewer patients receiving an adequate dose or duration of an effective active treatment, reducing drug-placebo separation. If test procedures are overly demanding, subjects may drop out due to demands placed on them. Another reason subjects may drop out is a subjectively perceived failure to respond to the treatment being provided, leading them to conclude they must be receiving placebo. This can occur especially when subjects are allowed to enter the study with baseline ratings that are too high, which may be related to treatment resistance. Subjects with very low baseline ratings may drop out because they spontaneously remit and decide additional treatment is unnecessary. Others will drop out due to medication side effects if clinicians are not careful to educate patients about the expected duration and severity of these effects.

Inadequate Treatment Dose or Duration

Inadequate medication dose or duration are other factors that can result in negative trial results when investigating effective compounds. Studies are often set up with resources and funding for only a specific time period; 4 to 6 weeks has been typical. However, a longer duration may be necessary for certain subpopulations of patients to respond to medications.[4] Also, the study may have a set medication dose that patients receive, but some patients may need either higher doses than others to achieve worthwhile benefit or lower doses to achieve efficacy without bothersome side effects. Turri and Stein[5] noted that often Phase 2 clinical trials have failed to determine the dose regimens to be used both in later clinical trials and in clinical practice, although this determination should be an important aim of these trials.

Functional Unblinding

During a blinded trial, clinicians may develop personal beliefs about the treatment assignment of a subject—active drug or placebo—based on perceived symptom changes. Trial investigators may rate symptomatic patients as improved because they desire to see improvements, either for the sake of the patient or the success of the trial. These subconscious investigator desires can affect their interpretation of what the patients actually say when being interviewed about their symptoms. This confirmation bias—seeing what we believe to be present because we want to see it—was described by Yogi Berra, who reputedly said, "If I hadn't believed it, I wouldn't have seen it." Although Hamilton[6] advised clinicians not to attend to anything except what is observed during the patient interview and to avoid reviewing previous scores while making ratings, such practices often occur. Besides noting symptom change, we also tend to notice side effects, which can contribute to assumptions that patients are receiving an active drug. The clinician may recognize side effects consistent with profiles of effective drug treatments, accurately perceive that the patient is taking the drug and not placebo, and infer that the patient is (or should be) getting better. Clinicians also listen to staff comments such as "John seems to really be better this week." All of these influences can have the effect of functionally unblinding the clinician. This functional unblinding of the clinician can mask or magnify the true results of the study because the patients are rated incorrectly—using information from the clinician's knowledge and beliefs rather than strict reliance on observations.

IMPROVING TRIAL METHODS THROUGH TECHNOLOGY

How IVR Works

Some methodological problems in clinical trials may be addressed by using interactive computer versions of reliable rating scales during both subject qualification screening and symptom change assessment. IVR technology uses a touch-tone telephone to interact with a computer. Telephones are accessible to the vast majority of the American population, and almost everyone knows how to operate a telephone. Patients call a designated number and answer questions related to their symptoms; their answers are stored by the computer. The use of the telephone to ask questions via computer has advantages over using a computer terminal alone in that the patient does not have to physically be present at the computer, and the person's reading ability is not a critical factor.

Capabilities of IVR include feedback loops during the call, transfers to or from a call center, pages to the clinician, and reports sent by fax or e-mail. Patients can be transferred to a call center for communication with a human or from a person to the IVR system for further interaction. If a patient in a clinical trial has a score on a rating scale that indicates that he or she has attempted to commit suicide, the computer can page the clinician. This has happened, and in one instance, the responding doctor sent police to the aid of a patient who could not be contacted by telephone, and the patient was found in a comatose state. Finally, because IVR delivers data to a computer, automatic reports can be sent to the clinician via fax or e-mail.

IVR is available 24 hours per day, 7 days per week. Patients considering a trial can call anytime, learn about the site nearest them, and either have the phone number of the site given to them or leave their name, telephone number, and the best time to be contacted.

Our research group validated IVR in a study with the PRIME-MD in 1997.[7] We found that IVR allows increased availability of psychopathology assessment and provides physicians with information that will increase the quality of care without additional physician time or expense. The computer-administered telephone scale used was the PRIME-MD, and the same group of subjects was also interviewed by clinicians using either the PRIME-MD face-to-face or the Structured Clinical Interview for DSM-IV Diagnosis by telephone. Thirteen screening questions asked about mood, anxiety, alcohol abuse, or eating disorders. If a person entered a positive response to any question, the IVR system branched to more questions for that module. All questions were in a yes-or-no format. The prevalence rates for any psychiatric disorder were found to be similar between diagnoses made by the computer and those made by clinicians (60.0% vs. 58.5%). Prevalence rates for individual diagnoses were also generally similar across interviews.

Table 2. Benefits of IVR Over Clinician Interviews

Captures patient testimony without clinician's biases
Uses standardized administration
Is available 24 hours/day, 7 days/week
Automatically checks errors
Reduces embarrassment about sensitive information
Stores data electronically

Advantages of IVR

Human interviewers do have some advantages over the computer, such as flexibility, educability, recognition of nonverbal cues, and affective relationships with the patient. However, IVR technology provides many benefits that clinician interviews cannot (Table 2).

With IVR, the patient's personal perspective of an experience—rather than the clinician's interpretation of that perspective—is recorded. Clinicians filter patients' experiences through themselves, but because IVR is standardized, it assesses each patient using consistent and objectively identical processes to obtain information. The electronic record of the patient's responses throughout the trial can be shared with the patient to anchor the patient's global impression of his or her improvement. Sometimes patients do not remember how they felt 8 weeks before when the trial began, but IVR can record their individualized symptom report in their own voice, with their inflections and affect. In this way, patients can hear themselves describe the effect of their symptoms on their functioning several weeks earlier and compare that with their current functioning. They can then tell us whether they are better, worse, or unchanged since that recording was made. This use of previously recorded evaluations is called Memory Enhanced Retrospective Evaluation of Treatment.

Also, because IVR is constantly available, it provides additional assessment possibilities. Before the trial begins, IVR can be used to qualify and enroll patients by having them call a number to be interviewed at any time. After the trial is underway, the constant availability of IVR permits clinicians to look for early treatment onset more easily than by increasing the burden of study participation by having patients come to the clinic more frequently. Patients can call using their cellular phones, or the computer can initiate the call to the patient to collect data by IVR.

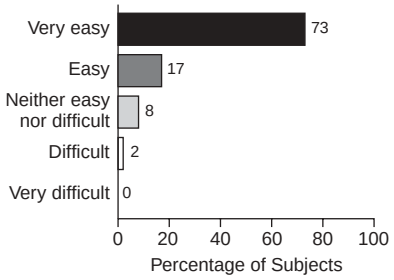
Error checking is done immediately with IVR because patients press keys that must be within the range of permissible responses. Patients respond by pressing a number corresponding to their answer on any touch-tone telephone. The consistency of responses can be verified against previous responses. The risk of incorrectly applying the scoring algorithm to determine diagnoses based on patient replies is eliminated when a computer administers the interview instead of a clinician.

Also, patients are often more comfortable disclosing sensitive information to a computer than to a clinician.

Table 3. Assessments That Have IVR Adaptations

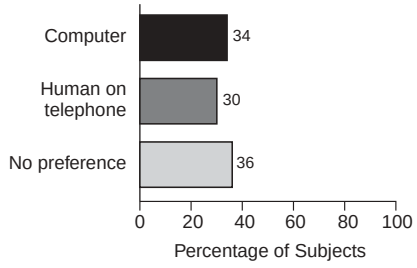
Central Nervous System Rating Scales
Brief Social Phobia Scale (BSPS)
Davidson Trauma Scale (DTS)
Hamilton Rating Scale for Anxiety (HAM-A)
Hamilton Rating Scale for Depression (HAM-D) (including French and Spanish versions)
Liebowitz Social Anxiety Scale (LSAS)
Patient Global Impression (PGI)
Yale-Brown Obsessive Compulsive Scale (YBOCS)
Patient Monitoring Evaluations
Asthma Symptom Diary
Headache Severity Scale
Hypertension Monitoring Scale
Irritable Bowel Syndrome Monitoring System
Life Chart (Bipolar Disorder Symptom Diary)
Pneumonia Monitoring System
Short Musculoskeletal Functional Assessment
Side Effects Monitoring Scale
Sleep Diary
Quality of Life, Work Productivity, and Functional Impairment Assessments
Asthma Quality of Life Scale
Endicott Work Productivity Scale
Memory Enhanced Retrospective Evaluation of Treatment (MERET)
Patient Global Impression of Improvement
Quality of Life Enjoyment and Satisfaction Questionnaire
Work and Social Adjustment Scale
Work Productivity and Activity Impairment Questionnaire
SF-12, SF-20, and SF-36 scales

Figure 1. Subjects' (N = 874) Opinions on Ease of Use for Computer Interview[a]



[a]Subjects had Hamilton Rating Scale for Depression and Liebowitz Social Anxiety Scale assessments.

Figure 2. Subjects' (N = 874) Preferences on Format of Interview



With IVR, patients are assured confidentiality through an assigned code and a self-selected password.

IVR does not appear to be better than clinicians at detecting drug versus placebo when a characteristic side effect profile is presented. However, several studies[8,9] have found larger effect sizes with drugs with more prominent side effects, suggesting possible rating bias due to unblinding. Others[10] have found that self-reports (as a proxy for blinded ratings) produce smaller antidepressant treatment effect sizes than clinician ratings, but not enough to invalidate study results. Unpublished data (J.H.G., 2001) from a clinical trial investigating a compound without substantial side effects found that IVR assessments did separate drug from placebo, while clinicians did not.

Therefore, IVR has many benefits when used both before the trial to enroll subjects and during the trial to monitor symptom change and assess quality of life. Many rating scales have been converted for use with telephones or computer terminals or both (Table 3).

Patient Comfort With Computers

Trial data have shown that patients are as comfortable with rating scale interviews by telephone and computer as with those conducted by clinician, if not more so. During the screening for a study of patients with social anxiety disorder,[11] 90% of 874 patients who participated in 2 telephone interviews, first with a clinician and then with a computer, rated the computer interview as "very easy" or "easy" (Figure 1). There was no significant difference in interview format preference (Figure 2).

Our research group[12] examined the reliability and validity of IVR versions of both the Hamilton Rating Scale for Depression (HAM-D) and the Hamilton Rating Scale for Anxiety (HAM-A). The correlations between the IVR and the clinician versions of these instruments were high. Subjects completed the IVR HAM-D in 12.23 minutes versus 15.21 minutes for the clinician version, and subjects completed the IVR HAM-A in 11.27 minutes versus 15.33 minutes for the clinician version. Subjects rated the computerized version as causing significantly less embarrassment, although they felt that they could describe and explain their feelings better to a human. Sensitive topics—such as alcohol or drug abuse, suicidal tendencies, and sexual behavior or symptoms—can often be addressed more easily with a computer than with a clinician.[13]

Kobak et al.[14] examined the use of computer-administered rating scales in a population for whom face-to-face interviews might be difficult: patients with social anxiety disorder. Correlations between the desktop computer and the original versions of the Liebowitz Social Anxiety Scale (LSAS), Brief Social Phobia Scale, and the Fear Questionnaire were high at baseline and remained high throughout the study. Katzelnick et al.[15] also compared clinician-rated versus computer-rated versions of

the LSAS. There was no significant difference between scores in their sample, but the majority of the sample preferred to be interviewed by computer.

Error Checking

A depression study[16] set the patient entry criterion at a score of 20 or greater on the 17-item HAM-D. Patients were rated at what clinicians believed was screening and baseline, and nearly every patient scored at least 20. However, many IVR assessments were below 20. This trial failed, possibly in part due to baseline score creep.

The IVR validation study of PRIME-MD[7] found that clinicians had made 12 (10%) incorrect diagnoses, in which the patients should or should not have received diagnoses according to their answers.

Dropout Rate

Another way to avoid confusing trial results is to lessen dropout rate. One study[17] using the HAM-A compared clinician-rated entry scores with those measured by IVR. The question was whether entry rating scores affected the dropout rate. The protocol required a HAM-A score of 20 or greater at baseline. We found that among subjects who were rated by the clinicians as having a HAM-A score of 20 or greater, 52.4% had IVR-rated HAM-A scores of less than 20. Among patients with a HAM-A score of 20 or greater as rated by both the clinician and IVR, the dropout rate was 16%. If the IVR-rated score was 16 to 19, 19% dropped out. In patients whose IVR score was 6 to 15, 46% terminated the study early, which comprised 54% of all early terminations.

Speed of Enrollment

A trial[18] used IVR technology to administer a 17-item HAM-D for trial qualification. After the IVR interview took place, the site received a fax within 2 minutes indicating that the patient had completed the assessment and whether the patient's score was acceptable for study enrollment or not. More than 1500 patients were enrolled in 16 weeks, which was half the anticipated enrollment time. IVR excluded 23% of the potential subjects. Among those accepted, 77% completed the 8-week protocol, compared with 72% and 60% rates of completion for drug and placebo patients, respectively, in 2 previous trials that did not use IVR screening.

CONCLUSIONS

IVR technology offers diagnostic and assessment instruments and models for incorporation into clinical trial methodology that can improve current practices. If used to qualify subjects for a trial, IVR technology may reduce clinician inclusion bias. IVR screening objectively selects subjects who are qualified to enter the study. More consistent, homogeneous samples are more likely to separate effective treatments from placebo. The use of IVR can also reduce enrollment time through more efficient screening of large numbers of potential subjects, permitting studies a better chance to complete on time.

Once the trial is underway, IVR assessment may contribute to reduced study dropout rates. Appropriately selected patients appear to be less likely to withdraw from the study early. Whether this reflects reduced spontaneous remission (placebo response), diminished treatment resistance (through reduced comorbidity), or some other mechanisms is not clear at present. Additionally, reducing study task demands, such as providing more convenient assessment methods via home telephone, can permit increased data collection procedures without increased patient distress.

Many central nervous system rating scales have been developed and validated for IVR delivery. Patient monitoring systems (diary log methods) and assessments of quality of life, work productivity, and functional impairment are also readily available. Use of these scales during trials can lessen the functional unblinding of clinicians by distancing them from the symptom monitoring process.

Thus, IVR assessments are a consistent, standardized method to make diagnoses before entering patients into clinical trials and can be used to conveniently record symptom change directly from the patients receiving treatments. In the end, it is not a case of humans versus computers; rather, combining computer interviews with clinician care provides patients with the best possible care and provides researchers with the most objective, and hopefully informative, trial results.

REFERENCES

1. Spitzer RL, Williams JB, Kroenke K, et al. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. JAMA 1994;272:1749–1756
2. Woggon B. Methodology of measuring the efficacy of antidepressants: European viewpoint. Psychopharmacology (Berl) 1992;106(suppl): S90–S92
3. Schatzberg AF, Kraemer HC. Use of placebo control groups in evaluating efficacy of treatment of unipolar major depression. Biol Psychiatry 2000;47:736–744
4. Quitkin FM. Methodology of measuring the efficacy of antidepressants. Psychopharmacology (Berl) 1992;106(suppl):S87–S89
5. Turri M, Stein G. The determination of practically useful doses of new drugs: some methodological considerations. Stat Med 1986;5:449–457
6. Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967;6:278–296
7. Kobak KA, Taylor LvH, Dottl SL, et al. A computer-administered telephone interview to identify mental disorders. JAMA 1997;278: 905–910
8. Greenburg RP, Bornstein FR, Greenburg MD, et al. A meta-analysis for antidepressant outcome under "blinder" conditions. J Consult Clin Psychol 1992;60:664–669
9. Greenburg RP, Bornstein FR, Zborowski MJ, et al. A meta-analysis for fluoxetine outcome in the treatment of depression. J Nerv Ment Disord 1994;182:547–551
10. Petkova E, Quitkin FM, McGrath PJ, et al. A method to quantify rater bias in antidepressant trials. Neuropsychopharmacology 2000;22:559–565
11. Katzelnick DJ, Kobak KA, De Leire T, et al. Impact of generalized social anxiety disorder in managed care. Am J Psychiatry 2001;158:1999–2007

12. 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
13. Kobak KA, Greist JH, Jefferson JW, et al. Computer-administered clinical rating scales: a review. Psychopharmacology (Berl) 1996;127:291–301
14. Kobak KA, Schaettle SC, Greist JH, et al. Computer-administered rating scales for social anxiety in a clinical drug trial. Depress Anxiety 1998; 7:97–104
15. Katzelnick DJ, Kobak KA, Greist JH, et al. Sertraline for social phobia: a double-blind, placebo-controlled crossover study. Am J Psychiatry 1995; 152:1368–1371
16. DeBrota DJ, Demitrack MA, 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. Presented at the 39th annual meeting of the New Clinical Drug Evaluation Unit; June 1999; Boca Raton, Fla
17. Feltner DE, Kobak KA, Crockatt J, et al. Interactive voice response (IVR) of patient screening in a clinical drug trial. Presented at the 41st annual meeting of the New Clinical Drug Evaluation Unit; May 28–31, 2001; Phoenix, Ariz
18. Kobak KA, Jeffers CR. Accelerating Patient Recruitment for Clinical Trials: Combining a Centralized Call Center and Computer-Administered Telephone Screening [white paper]. Madison, Wisc: Healthcare Technology Systems; 2001