Published in final edited form as:

J Clin Psychopharmacol. 2009 February ; 29(1): 56–64. doi:10.1097/JCP.0b013e318192e2fa.

Neuropsychiatric Clinical Trials: Should They Accommodate Real-World Practices or Set Standards for Clinical Practices?

Robert E. Becker, MD, CM and Nigel H. Greig, PhD

Drug Design and Development Section, Laboratory of Neurosciences, Intramural Research Program, National Institute on Aging, National Institutes of Health, Baltimore, MD

Abstract

The fundamental tenet of Evidence-Based Medicine (EBM) is to “integrate the best research evidence with clinical expertise and patient values,”¹(p1) a commitment accepted in neuropsychiatry.^{2,3} The EBM group recognizes various factors that undermine the quality and use of evidence generated in research, “three limitations...to science and medicine—shortage of coherent evidence, difficulties applying evidence in care, and barriers to quality practice—and further impediments to EBM practice—practitioners lacking skills evaluating evidence sources, having limited time, and being unaware of support for EBM working, thus failing to follow its practices.”¹(p7) Other risks to validity are less widely acknowledged.

Clinical trials (CTs), especially randomized controlled trials (RCTs), and summary reviews of results from more than 1 RCT provide EBM’s gold standard sources for sound evidence.¹(pp105–144) Sackett et al¹ and other authors suggest subjecting RCTs and reviews of RCTs to specific tests of validity before the practitioner uses the evidence. We recently compiled additional threats to validity of the neuropsychiatric evidence base,^{4,5} a list already incomplete in view of recent concerns with industry influence evidenced by ghost authorships⁶ and selective reporting.^{7,8} Each of the factors we compiled potentially affects the reliability and therefore the validity of the RCT evidence base, is not addressed systematically in EBM guidance on how to develop and use the research literature, and potentially impacts neuropsychiatric research by allowing drugs to fail because of the factor functioning as a methodological weakness in clinical studies.⁵

In this article, we (1) cull from the literature factors that methodologically put clinical research and the evidence base at risk, (2) uncover assumptions that may account for these factors going unnoticed as risks to medicine’s evidence base, and (3) suggest steps to increase the effectiveness of neuropsychiatric drug developments, CTs, and validity and use of the evidence base for practitioners. Specifically, we provide evidence that problems of unreliability caused by human errors and biases currently undermine the validity of psychiatric research. We suggest revisions of some assumptions behind research methods and practices as part of an effort to protect research from these errors and biases.⁴

Keywords

clinical trials; errors; biases; real world conditions; protocol standards

Reprints & correspondence: Robert E. Becker, MD, CM, (rebecker2008@comcast.net) or Nigel H. Greig (Greign@grc.nia.nih.gov), Drug Design and Development Section, Laboratory of Neurosciences, 5th floor BRC Bldg, 251 Bayview Blvd, Baltimore, MD 21224.

The authors have received no other funding for this work and have no financial conflicts of interest. The views expressed within this article are those of the authors and may not represent those of the National Institute on Aging, National Institutes of Health.

BACKGROUND

Grounds for Suspecting Flaws in Clinical Neuropsychiatric Research Practices

Pangalos et al ⁹ recently focused on how factors in early stages of drug development contribute to drug attrition and undermine the magnitude of clinical importance detected in CTs. We and others raise how unreliability associated with methodologies and practices possibly contributes to CT failures.^{4,5,10,11} Becker and Markwell ¹² found random measurement errors, with two of the most used cognitive rating scales in Alzheimer disease (AD) exceeding the magnitudes of average expected drug effects and annual rates of decline caused by the illness. Engelhardt et al ¹⁰ reported that over half of the raters in 2 multicenter antidepressant trials could not detect drug effects in actively treated patient subjects. Cogger ¹³ demonstrated how data generated by inaccurate raters reduce the effect sizes in antidepressant CTs. Kobak et al ¹⁴ showed that investigators in an antidepressant CT exaggerated the pathology for potential patient subjects compared with blinded off-site raters. Kahn et al ¹⁵ found high failure rates in antidepressant CTs. Becker et al ⁵ found, in reports of clinical research with AD drug candidates, little and often no evidence that investigators specifically addressed these and other methodological sources of unreliability able to invalidate clinical studies.

A full understanding of the scope of these problems is blocked by lack of access to relevant data. As various authors have noted for decades, many clinical research studies go unpublished, especially if results from the study are negative. Over a decade ago, authors ^{16,17} called nonpublication unscientific and biased. Recently, Cochrane review authors ¹⁸ noted failures of pharmaceutical firms to release data for reviews, and Becker et al ⁵ found that 75% of announced AD drug developments lacked published results about development or reasons for abandonment. Turner et al ¹⁹ found 31 out of 74 antidepressant CTs unpublished. When Turner et al ¹⁹ took unpublished studies into account, the impression from the published literature that 94% of studies were positive declined to only 51% positive.

Clinical trial registrations seem not to have their intended effects of increasing access to information about drug study outcomes. Neither do they reveal how methods may be associated with drug failures in clinical stages. Becker et al ⁵ found little information relative to possible reasons for trial failures in CT registrations. Because of the resulting lack of information about factors that may affect the validity of unpublished AD CTs ⁵ and distort reporting of some published CTs,^{6,7,19} we conclude that full discovery of the methodological problems in neuropsychiatric CTs is not possible. This lack of access to human research outcomes contributes to evidence of unreliability in neuropsychiatric CTs being indirect, at times anecdotal, fragmentary, and difficult to assess systematically. At present, we must work with what is known and hope that further research and improved standards governing human research will increase our understanding of the specific problems facing CT researchers.⁴

SOURCES FOR UNRELIABILITY IN NEUROPSYCHIATRIC CTs

Increasingly during the last decade, investigators have become concerned with the integrity of clinical rating methodologies used for neuropsychiatric disorders. Investigators find that unreliability interferes significantly with the power of CTs to detect differences between active drug and placebo conditions.^{10–13,20–26} A wide range of potential sources of unreliability has been suggested (Table 1).

Subject Samples

Pangalos et al⁹ describe how heterogeneity among subjects in CTs adds to variance in data. They describe how, by introducing subjects incapable of responding to treatment because of misdiagnosis or genetics, heterogeneity reduces the actively treated group response to drugs undermining the power in critical early clinical phase studies and in CTs. Researchers have identified 3 problematic sources for heterogeneity in subject samples: failures of diagnostic or other CT criteria, deviations from selection criteria and study protocols at research sites, and excessive sizes of CTs.

Visser et al,²⁷ using the patient sample in their AD clinic, showed how entry criteria used in 6 different studies of mild cognitive impairment (MCI) selected 6 different groups of subjects. Visser et al²⁷ applied the MCI enrollment criteria from 6 studies to 150 patients in their own clinic. Using the different study criteria, they made MCI diagnoses in 51% to 17% of their patients. Becker²⁸ and Cogger¹³ calculated how imprecision and inaccuracy introduce errors into scale values used to quantify severity at entry and outcomes from treatment. Both imprecision and inaccuracy, to the degree that they allow deviations from rating ranges of severity required by the study protocol, can introduce heterogeneity into study samples.

Kobak et al¹⁴ demonstrated how flawed implementations of study protocols can become a source for heterogeneity and samples being nonrepresentative of the disorder being studied. They found that clinicians inflated ratings of depressive pathology, presumably to qualify subjects for CTs. One would expect inflated ratings to lead to heterogeneity in subject samples, to samples not reflecting study criteria, and to the responses of these subjects not being representative of what a clinician can expect for her subjects who do match study criteria.

The large number of subjects needed in CTs necessitates research sites to advertise to fill quotas and complete studies in a timely manner. Subjects who do not volunteer for CTs can have different characteristics from subjects who do volunteer, a matter not often taken into account in designing CTs.²⁹ It seems reasonable to expect comparable differences between subjects who answer advertisements and those who do not.

Further difficulties arise generalizing from CT subject and investigator samples to applications in clinical practice. Samples are not random but selected by convenience, necessitating that clinicians determine whether their current patient matches the research sample in the study they will use. This convenience compromise is unavoidable: total populations of affected persons, sites, and ultimate practitioner users are not known and available to be sampled; sites and investigators self-select often based on economic considerations; and study patient samples are self-selected by persons willing to give informed consent to participation. A failed demonstration of efficacy with the study sample may lead to abandonment of development efforts for the candidate drug because the investigators assume that any other similarly selected sample from the populations of subjects or investigators will probably respond as did the study samples. If a truly random sample has been drawn, this reasoning may be sound at the indicated level of statistical significance for the outcome. This conclusion is potentially misguided if, as Raschetti et al³⁰ and Kobak et al¹⁴ find, the study sample has been biased by methods of selection, as Worrall³¹ emphasizes, by unobserved chance differences, or by methodological lapses. Selected not randomly but by convenience, subjects, investigators, and research sites may or may not be representative of community skills and practices, and existing clinical skills and practices may or may not be adequate to take full advantage of a drug dependent on conditions of use for its efficacy.

Given that trained raters exhibit imprecision, inaccuracy, and bias evaluating subjects using instruments on which they received training, clinicians, most of whom will not use standardized ratings, can expect difficulties ensuring that their patients have the degrees of severity studied in the CT. The current inability of CTs to provide clinical decision rules for practitioners, a logical and ethical aim for CTs, may also undermine applications of CT-demonstrated drug efficacy and safety to clinical practice.^{4,32–35} In sum, because of sampling deficiencies, CTs may not be methodologically adequate to evidence the efficacy possible with a drug and may fail clinicians because of limitations when generalizing from the CT to the clinic.

Adverse Effects From Unreliability on CT Design and Implementation

When inaccuracies, in the form of reduced ability of raters to detect drug effects,¹⁰ and imprecision, excessive random measurement errors leading to increased variance,²⁸ are present, then the power of a proposed CT to detect significant differences will be reduced. Becker²⁸ and Cogger¹³ calculated how imprecision and inaccuracy in outcome ratings, by compromising power, may cause CTs to fail, to underestimate treatment effects in the presence of efficacy, and to require excessive numbers of subjects and sites to provide subjects. These increased numbers of sites and investigators introduce further opportunities for compromises to the integrity of the research.

To increase power in CTs, investigators add numbers of subjects and thus tuck the problems with design and execution of the CT into the denominator of the statistical test of significance.³⁶ This increase in size of the CT is supported by recommendations from statisticians who argue that larger numbers of subjects increase the ability of a study to detect differences that are present.³⁷(p273) Various problems arise from this assumption. First, a difference caused merely by the size of the study sample may well have no clinical importance.³⁶ Second, AD researchers have begun to realize that as sites are added to provide subjects, larger numbers of subjects create unrealistic training and monitoring demands, increase intersite variance entered into the CT statistical analysis, and in turn compromise power, leading to possible Type II errors or failures of a CT to detect or accurately quantify efficacy when it is present.^{4,15,26,38–41} Third, because of the time, costs, and demands on clinical investigators involved, large CTs cannot be practically replicated. Replication of results is a core feature for claiming scientific validity.

In addition, for many neuropsychiatric conditions, larger numbers of subjects combine with abilities of patients to obtain treatment from already approved drugs to shrink the numbers of persons seeking out or willing to participate in CTs. This necessitates investigators advertising for potential subjects with the risks to validity we have already discussed. Butcher reports Aisen and Vellas regarding “large degrees of variance among sites,” the need for “70 to 80 sites around the world,” sites contributing only “2 to 5 patients,” and noncompliance of sites with CT protocols as difficulties that contribute to CTs failing; yet they do not elaborate on these difficulties or suggest routes to overcoming these risks in upcoming trials.³⁸

Muller and Szegedi⁴² and Becker and Greig⁴ emphasize how unreliability risks false-negative reports from CTs and the ethical consequences of patients being exposed to inadequately designed and controlled research conditions. We reason that the imprecision that we found in AD rating scales,¹² failures of clinical raters of depression to follow protocols for uses of rating scales,¹⁰ the inaccuracy of up to 50% of trained raters in 2 multicenter antidepressant trials, and unreliability produced by these failures could be, until proven otherwise, considered risks for all clinician-rated neuropsychiatric outcome measures.⁴³ We fear that errors, bias, and faulty practices found in depression, AD,

posttraumatic stress disorder (PTSD),⁴⁴ and other neuropsychiatric disorder studies potentially impugn all neuropsychiatric clinical studies until proven otherwise.

The Institute of Medicine,^{44(pp137-139)} in a review of CT evidence claimed to ground treatments of PTSD, found numerous problems with sample size, blinding, independence of investigators, the need to exclude studies that were uncontrolled case studies and case series, high dropout rates capable of introducing bias, potential influence from widespread pharmaceutical funding, publication bias, issues with population selections, and other areas of concern. The committee authoring the Institute of Medicine report concluded that for PTSD, “significant gaps in the evidence...made it impossible to reach conclusions establishing the efficacy of most treatment modalities.”^{44(p ix)} Faced by the “scientific inadequacy” of available CTs, the committee recommended “the latest and most rigorous methods for designing and executing study protocols.” Unfortunately, these authors do not discuss the extent to which methods and practices need to be governed by protocols.^{44(p x)} Engelhardt et al,¹⁰ Raschetti et al,³⁰ and others agree on the need for protocols to control clinical research practices.

Protocol Deviations

Kelloff et al⁴⁵ describe how cancer drugs fail by testing inappropriate populations and settings where an end point cannot be reached. Pincus and Stein⁴⁶ report similar problems in rheumatoid arthritis. These authors emphasize the importance of establishing criteria for practices associated with CTs. Unlike Engelhardt et al,¹⁰ who followed the protocol for administration of the Hamilton Depression Scale by William,⁴³ they do not develop protocols able to systematize, control, replicate, and evaluate conditions and practices across investigators in centers. In their review of MCI CTs, Raschetti et al³⁰ found that failures to adequately operationalize diagnostic criteria and rating methods resulted in wide differences in the diagnostic instruments and rating scales used in various studies. They concluded that quite different populations had been recruited to the various MCI studies in the literature.

From investigators’ failures to report how protocols govern the administration of clinical rating scales in AD CTs,⁵ we infer that most investigators at research sites depend on earlier clinical skills, experiences with rating scales, and training for the CT. Research shows that most of the clinical raters in CTs are often clinically unfamiliar with the disorder being studied and previously inexperienced with the rating instruments being used.⁴⁷ This and the failures of some raters to respond adequately to training suggest the need for protocols for qualifying and training raters and not just protocols for administering rating scales.^{48,49} Because most CT reports lack documentation of training, investigators’ skills, and monitoring, it is not possible to determine if the guidance provided by research protocols is sufficient to ensure quality in clinical ratings and execution of the CT in accordance with the proposal for the CT.⁵ The work we cite does not give us reasons to be sanguine about whether, even for situations where protocols are available, clinical research complies with plans and clinical equivalents of good laboratory practices.

Levels of Rater Training and Possible Absence of Needed Skills Among Some Rater Candidates

Demitrack et al⁴⁸ and Kobak et al⁴⁹ found that extended training of raters did not improve the performance reliability of some clinical raters in CTs, and investigators remain undecided about how to overcome problems of rater unreliability.^{26,50} Engelhardt et al,¹⁰ Cogger,¹³ and others regard lack of experience with rating scales and with diagnoses as important factors affecting rater performance. Limitations that raters bring to studies may make unavoidable unreliability, inaccuracy, lapses leading to compromises of research protocols dependent on human estimations, and other errors of omission and commission.

PLACEBO GROUP IMPROVEMENTS

Placebo group improvements occur frequently in antidepressant CTs—

Exaggeration of pathology to qualify subjects for clinical research, as reported by Kobak et al,¹⁴ has been suggested as a possible source for some placebo group improvements. Hick and Feldman⁵¹ explain placebo group improvements in psoriasis as caused by similar eligibility creep-patients having higher measured severities at initial assessments. Kirsch et al,⁵² in a meta-analysis of all CT data submitted to the Food and Drug Administration (FDA) for 4 new-generation antidepressants, found that decreased responsiveness to placebo among very severely depressed patients accounted for the reported increased benefits from antidepressants in this group. They reported mean active drug-placebo differences of 1.8 points on the Hamilton Rating Scale for Depression. These differences are only 23% of the mean placebo response of 7.8, supporting the marginal evidence for efficacy of antidepressants emphasized recently by Turner et al.¹⁹ Given the magnitude of placebo responses in depression, FDA regulators and others conclude that comparisons with active drugs already shown efficacious cannot be substituted for placebo.^{53,54}

Ferris⁵⁵ and others have argued that placebo group improvements are not increased in AD compared with decades earlier. Our personal experiences as consultants for studies that pharmaceutical companies did not publish suggest that placebo group improvements occur with inordinate frequency if one accepts as a standard for AD patient declines in earlier published placebo studies and untreated decline rates for AD.^{5,28} Walsh and Sysko⁵⁴ report an increased rate of placebo group improvements in major depression CTs in the last 2 decades of the 20th century. Given the high rates of failure to publish negative clinical studies, it is not possible to exclude that even greater numbers of placebo group improvements occur in CTs but go undetected.

The recent report on drug prevention of seasonal affective disorder (SAD), drawn to our attention by Steven Markwell (oral communication, September 25, 2005), illustrates another not sufficiently discussed implication of high placebo response rates.⁵⁶ Sixty-nine percent to 79% of placebo-treated patients and 81% to 89% of active drug patients were depression free during the CT. This placebo response does not discriminate whether the data are modeling the natural course of SAD or are influences on active drug- and placebo-treated patients from factors other than the active treatment. For example, Andrews⁵⁷ discusses, as one explanation for placebo improvements in depression, the benefits from good clinical care-attention to subjects and to their problems of living. Andrews⁵⁷ reports difficulties persuading other than general practitioners to assess and resolve patient problems and to encourage patients to resume positive activities, matters that were part of our AD Center patient care.⁵⁸ This aspect of our AD patient care may account for the anecdotal support we reported for benefits from counseling AD patients.⁵⁸ It is possible that in the SAD study, only 13% of subjects required drug to avoid SAD episodes and that treatment recommendations drawn from this study should be broader than the easily drawn implication that the drug provides a preferred line of prevention among SAD patients. A wide range of possible sources affecting patient clinical courses can complicate, compromise, or invalidate CT data and data interpretations.

Hazards Suspected but Not Yet Documented

Although not extensively discussed in the literature, other factors-research sites falsifying data, lack of adherence of raters and administrators to protocols, lax applications of diagnostic criteria to secure subjects, abuses of advertising for patients for a CT, and so forth-occurred in our experience and generated anecdotal comments at meetings and in published reports from investigators.³²

RISKS OF TYPE II ERRORS IN NEUROPSYCHIATRIC CLINICAL RESEARCH

Do Flawed CT Methodologies Fail Drugs or Ineffective Drugs Fail in CTs?

Earlier we asked whether, at least in part, investigators, in choosing how they develop drugs, currently fail drugs by not providing fair tests for efficacy and adequate safeguards against conditions that undermine safety.^{4,28} Against the background of possible error and bias intrusions into studies and concerns with accountability at research sites implementing neuropsychiatric research protocols, we concluded that limitations in drug candidates do not necessarily account for all failures of CTs to confirm efficacy or accurately quantify efficacy.^{4,28,38,59,60} Failures of studies caused by the conditions of experimentation are basic risks in all research. Methodological sources for failure can produce Type II errors in interpreting CT outcomes. Given the range of potential compromises to the power of CTs, Type II error risk may, in some areas of neuropsychiatric drug developments and CTs, be more prevalent than is appreciated.⁹

Can Type II Error Risks Be Better Controlled?

Pangalos et al⁹ and Becker and Greig⁴ propose modifications in drug development and CT practices designed to minimize, not eliminate, Type II errors. These steps include a wide range of changes in practices (Table 2).

In addition to the interventions listed in Table 2, more research is needed into effects from unavoidable random measurement errors and inaccuracies, rater skills, required levels of training and monitoring at research sites; into subject selection and inclusion in CTs; into relations between rated outcome measuring instruments and uses of biomarkers and surrogate markers in research and patient care; and into human limitations using rating instruments with precision, accuracy, and without bias.^{10,12,47} Additional studies are needed to determine the adequacy of scientifically unsystematized and uncontrolled patient evaluations made by patient care practitioners. Even with this research, we are not sanguine that currently available methodological practices and recommendations to modify these practices will be adequate.^{4,9} For example, Demitrack et al⁴⁸ and Kobak et al⁴⁹ found that some raters cannot achieve, even with training, the skills required to rate a patient's clinical status. Leadership from clinical pharmacologists will be needed to improve scientific standards for clinical phases of drug development and prescription in clinical practice to the levels of Good Laboratory Practices used in preclinical research.

THE EFFECTS FROM CONTEXT ON CTs

Is There Currently Sufficient Evidence to Warrant Redirection of CTs or Is More Research Required?

We have presented evidence that CTs may be biased by investigators qualifying subjects who do not meet entry criteria,¹⁴ by large numbers of raters who by failing to accurately detect changes in patient status undermine effect sizes,¹⁰ by rating scales lacking sufficient precision to distinguish subtle but potentially important changes in patient clinical status,¹² and by other forms of error and bias. If this is the case, then neuropsychiatric drug developments can be severely handicapped and patients are potentially deprived of treatments that could offer benefits.^{4,9} From the literature and our own research, we cannot adequately catalog the full range of the problems, fully describe each potential problem we are able to identify, and determine the depth of the problems individually or their consequences for CTs. Clearly, more research is needed if methodological problems of neuropsychiatric CTs are to be addressed adequately.

Based on the evidence available to us, we consider neuropsychiatric CTs at increased risk from undetected Type II errors.⁵ Drugs capable of efficacy may not achieve regulatory approval and patient care may suffer. Drugs with proven efficacy may not be used with optimal effectiveness in the clinic if practitioners lack guidance about effect sizes possible with the drug, conditions required for drugs to show efficacy, and rules to govern proper uses of drugs with patients.³³ Most phase III investigators justify this environment as unavoidable, as a consequence from prescribing drugs under real-world conditions.

We ask, “Is it in the best interest of patients for neuropsychiatrists to accept ‘real world of practice conditions’ for CTs and clinical practice?” We suggest replacing real world practice conditions for phases I, II, and III with standards equivalent to the Good Laboratory Practices that currently govern preclinical phases of drug development. In section 3, we consider what framework for research and clinical practice would implement Good Clinical Research Practices once developed. We now question some current assumptions that allow investigators to tolerate real-world conditions and practices affecting clinical research.

Insights From Asking “What Does a CT Test?”

Investigators assume that CTs test the efficacy of the active drug. Investigators fail to give adequate consideration that the CT combines tests of the research hypothesis of drug efficacy and interrelated tests of the assumptions, conditions, and methods used. A failed experiment, similar to the many failed AD and antidepressant CTs, poses a dilemma for the investigator.^{5,19,57} Has the drug failed or have the methods failed the drug by undermining detection of evidence witnessing to its effectiveness?²⁸ This dilemma is especially challenging if a strongly supported theory is at risk for disconfirmation by the CT.

During an eclipse of the sun, a first telescope confirmed but a second smaller telescope failed to confirm predictions from Einstein’s theory that large masses curve space. Astrophysicists accepted evidence from the larger telescope and dismissed the data from the smaller telescope as caused by limitations in that telescope. In their interpretations, both the research hypothesis and auxiliary hypotheses about the validity of methods and conditions received consideration. Pangalos et al.⁹ and others describe similar methodological uncertainties when interpreting investigational outcomes (Table 3).^{4,49,50}

How does an investigator interpret a failed CT when no methodological errors are obvious? A number of cholinesterase inhibitors (ChEIs) have shown efficacy in AD as have multiple antidepressants of particular classes. An investigator of a new ChEI might choose to ascribe a failed AD CT to methods or molecular properties unrelated to the intended action of the ChEI because it is of the same class as many other drugs with efficacy. In contrast, for a new muscarinic agonist in AD, where many drugs of the class have failed, an investigator would seek out specific justifications to invoke failure as caused by the molecule or trial methods.⁵ When a CT fails, before accepting the null hypothesis of no treatment effects with the compound, investigators must consider how conditions of the experiment could account for the failure. Similarly, practitioners need to consider whether a patient’s failure to respond or degree of response to a treatment with proven efficacy stems from a correctable flaw in her conditions of drug use, such as inaccuracy in assessments, mistaken dosing, patient noncompliance, and so forth.

As Duhem^{61(pp190-195)} describes for the nonlife sciences, we must assume that a clinical investigation tests a research hypothesis of efficacy, multiple auxiliary hypotheses about the sources of the research hypothesis, and the adequacy of our methods and conditions used in the CT. The investigator cannot view a statistically significant outcome with clinical importance or a failure to show efficacy as a conclusion supported by CT data until the investigator accepts the reasoning and methodology as sound. Similarly, a practitioner

cannot safely translate findings from a research setting into the clinic without taking into consideration how specific practice standards may be required to achieve the efficacy demonstrated in the CT. Additional critical assumptions are always required even if one uses science's soundest practices and EBM's best evidence. Conclusions relating to efficacy are vulnerable in many ways other than the often cited risks of Types I and II errors.

Do Scientific Standards Require That CTs Prescribe to the Real World the Specific Conditions Required to Optimize Uses of the Drug?

Many difficulties we and others report with neuropsychiatric CTs and drug developments can be interpreted in 2 different ways. The most prevalent interpretation is that compromises to CT investigative conditions represent features of the real world and that the drug must demonstrate efficacy under these real-world practice conditions. To correct or mitigate compromises to CTs from real-world practice conditions, we offer an alternative viewpoint. We take the evidence we cite here and elsewhere^{4,5} as supporting the interpretation that current neuropsychiatric drug development research and clinical practices, by imposing too few stringent conditions on CT research sites and practitioners, in important ways do not meet standards of practice appropriate to real-world medicine becoming science based.

Our review of the literature and our own work have identified a wide range of factors that operate to make research and clinical practices inconsistent with medicine practiced in compliance with scientific standards^{1,4,9} (Tables 1–3). Consequently, we propose that investigators emphasize the CT as a test of the complex hypotheses of (1) drug efficacy with (2) the efficacy test conducted under CT protocol-specified conditions of use of the drug. In this view, CT tests for efficacy are accepted as valid or invalid conditionally, in the context of the CT conditions of drug use. In clinical practice, without evidence supporting a clinician's personal choices how to use a drug, efficacy can be regarded as optimized for patients only under CT-proven conditions of use of the drug. This viewpoint charges CT investigators to determine, by protocol, the conditions under which the drug is used with optimal effectiveness in the real world. Good clinical research practices prevent real-world conditions from negatively affecting the efficacy identified for the drug.

OPTIMIZING EFFICACY IN CTs

Biases Introduced by Real-World Conditions and Intention-to-Treat Analyses

Current practices call for intention-to-treat (ITT)-based analyses of CT data.^{62–66} Unfortunately, this standard can allow methodological shortcomings in the CT design or real-world compromises to CT protocols to cause drugs to fail or CTs to misrepresent efficacy.^{4,13,28} Intention-to-treat analyses include unreliable data produced by investigators or sites displaying carelessness, haste, inattention to protocols, inaccuracy and imprecision using rating instruments, or other practices. This tolerance for design flaws, protocol deviations, real-world conditions, and other defects opens CT interpretations to Type II errors. These ITT requirements conservatively protect regulatory approvals of drugs from selection bias possible if investigators can control the inclusion of data in analyses. On the other hand, this regulatory required standard of analysis may have led to the suppression of other analyses needed to differentiate whether a drug fails because of methodology or to its inherent lack of efficacy.

Intention-to-treat assumptions require investigators to accept effects from errors, biases, carelessness, and other factors as real-world conditions. Intention-to-treat analyses of data may more reflect flawed standards of practice in the community than efficacy or lack of efficacy of the treatment being tested.⁵ We propose that additional study design and statistical analysis research are needed. Aims of this research would be to preserve the high standards imposed by ITT analyses and provide additional models for analyses better able

than ITT-based analyses and interpretations to quantify the influence of real-world factors in the CT. We propose these exploratory analyses as essential steps toward imposing more rigorous scientific standards on interpretations of CT outcomes.

ITT Analyses

Modified ITT analyses aim to minimize the risks from bias shaping outcomes from CT analyses. All patients who receive at least 1 dose of the active study drug and receive at least 1 outcome assessment are included in the efficacy analysis. These stringent limits are imposed to avoid analyses of CTs based on subsets of the entire randomized group, such as including only patients who complete a CT. Censoring of data runs the risk of introducing biases not controlled by unbiased blind assignments to treatment arms.^{31,67} The ITT analyses may be flawed and disposed to generating Type II errors; but as state of the art for controlling bias using blinding, ITT analyses are mandated by the FDA.⁶⁴ The ITT analyses aim to estimate the impact of treatments in subsequent clinical practices. Users of ITT analyses assume that the practice settings used for the CT are representative of the settings where the drug will be used and that these real-world conditions are unavoidable conditions in future uses of the drug once approved for prescription.⁶⁸ In Table 4, we list risks for CTs from real-world conditions.

Per-Protocol Analyses

We propose, as a supplement to current ITT analyses, the statistical analysis of data for subjects whose conditions of treatment during the CT complied with the research protocol conditions. The aim of this analysis is to reveal real-world conditions for what they are and what effects they have on the efficacy possible with the compound under study. Real-world conditions are currently assumed as necessary conditions for the CT to predict drug efficacy in clinical practice. This assumption requires investigators to use widely scattered research sites, sites with not always adequately experienced and trained personnel, and sites monitored only in limited ways and often not disposed to comply with CT protocol procedures.^{4,5,38} The proposed per-protocol analyses will hopefully, by excluding influences from these currently tolerated real-world conditions, provide indications of efficacy less biased by methodological lapses, errors caused by human factors, biases, and noncompliance of sites and investigators with protocols. We are not proposing that per-protocol analyses replace ITT analyses as the standard for regulatory approvals. We see per-protocol analyses as indicating levels of efficacy achievable under controlled and systematized good practice conditions and thus challenging investigators to conduct CTs in ways that enable efficacy in ITT analyses.

We foresee CT analyses able both to preserve the protections against bias in ITT analyses and to recognize CTs as tests of the (1) efficacy of a compound (2) used under CT protocol-specified conditions for use of the drug. Condition (2)-that a drug is used under specified protocol conditions allowing the CT to evidence standards for subsequent practice-requires the exclusion from the per-protocol statistical analysis of instances of uses of the drug under other than protocol conditions. If an aim of a CT is to pharmacologically demonstrate the available effects from a drug in the human species, then the testing must not be compromised by scientifically inadmissible practices. To reduce risks of Type II errors unfairly terminating drug developments and to set the conditions of drug use to optimize effectiveness for patients in later practice per-protocol analyses will be most conservative. We expect these per-protocol analyses to render drug developments more effective by reducing Type II errors, to increase the accuracy of effect sizes estimated, and to provide drug approval regulatory agencies opportunities to approve features of use implied by condition (2)-that the drug has demonstrated efficacy and safety, not under unspecified real-world conditions, but under protocol-specified conditions of use.⁶⁹

DISCUSSION

Authority to support the recommended changes may not exist under the present FDA regulatory enabling legislation and may be resisted by pharmaceutical firms. On the other hand, the needs for methodological research into CT conditions and how they affect the validity of CTs give academic clinical pharmacologists a framework to advance the roles of science in drug development and patient care. The aims of the changes we propose are that scientifically verified standards of care become well evidenced by CTs for latter applications in clinical practice and that these advanced standards be taken seriously in patient care. As Schwartz and Lellouch⁶⁹ note, both explanatory and practical approaches are needed according to whether the aim of the CT is to answer research and clinical application questions. We have tried to specify more clearly the function and roles of each approach.

Protocol standards for qualifying data for entry into per-protocol CT analyses will not necessarily compromise the randomization or blinding of a CT if steps to exclude biases occur before identification of subjects as active or placebo-drug treated. Supplemental per-protocol analyses would not change the restrictions against using subject noncompliance to exclude subjects.^{70,71} We seek to specify with CTs the conditions required to exclude factors able to interfere with drug efficacy and its assessment. For example, we suggest protocols to specify how investigators and practitioners can best ensure subject compliance. Both the research subject and patient must comply with dosing to provide a fair test of the candidate compound's efficacy. In this perspective, failure of a subject to comply is an indicator of possible ineffectiveness of CT or care methodologies. Failure of a subject to comply becomes an outcome measure testing the drug administration protocol's abilities to ensure that the efficacy of the treatment is optimized for patients. Just as the research adequately tests the drug for efficacy only if dosing is fully optimized in phase II, so too the research tests for drug efficacy only if methods to ensure compliance and other conditions needed for efficacy are determined in phase II and effectively implemented in phase III. Per-protocol analyses aim to assess both how well investigators complied with protocol conditions and whether phase II determinations of conditions required for optimal efficacy are confirmed in phase III. A CT may fail because phase II conditions for efficacy were not adequately tested. This possibility emphasizes the importance of adequate phase II and earlier studies of drugs-requirements often not met in current drug developments as sponsor rush into clinical studies. Our intention recommending per-protocol analyses is to uncover methodological or practice factors that lead to drug failures and to identify more specifically the optimizing conditions a practitioner can seek to best ensure drug efficacy in his or her patient. A sounder framework for neuropsychiatric drug development must uncover nonadherence to or weaknesses in research methodologies, practices, and protocols as lurking possible causes for drug failures in research and in later practice applications to patients.^{72,73}

Treatments have use in medicine to the degree that they benefit patients. Currently, clinical pharmacological practices leave too much room for confusion about whether some failed or compromised drug developments may be caused by methodological deficiencies in CT designs-to failures of researchers at various sites to comply with research protocols, inadequate earlier preparations for setting CT conditions needed for efficacy and safety-or to drugs themselves. A wide range of supports can benefit CTs and patient care (Table 5). Without these and other supports-training, specialized experiences with disorders of interest, monitoring for quality, and so forth-researchers and clinical practitioners can be expected to respond at times to patient observations as if they were true and not reflections of error components in the clinical data. We hope that as a result of advances in CT methods and practices, practitioners committed to EBM standards for practice and supported by more informative CTs will be able to access well-evidenced protocols able to guarantee science-

based patient care. To implement this quality standard in EBM applications, we advocate greater regard for the CT as both establishing efficacy and as creating the conditions of accuracy, precision, freedom from bias, and so forth that allow efficacy to emerge. The CTs inseparably establish efficacy and conditions of accuracy, precision, freedom from bias, and other methodological conditions important for efficacy and its effective applications in patient care. Scientific neuropsychiatric research and clinical practices cannot subscribe to real-world conditions without risking unnecessarily failed drugs and patient treatments.

Acknowledgments

The authors are supported in part by the Intramural Research Program of the National Institute on Aging, National Institutes of Health.

The authors thank reviewers whose constructive criticisms greatly guided the authors in revising the manuscript, in organization and presentation of evidence, and in arguments for reframing the conditions of CT.

References

1. Sackett, DL.; Strauss, SE.; Richardson, WS., et al. Evidence-Based-Medicine: How to Practice and Teach EBM. New York, NY: Churchill-Livingstone; 2000.
2. Gray GE, Pinson LA. Evidence-based medicine and psychiatric practice. *Psychiatr Q.* 2003; 74:387–399. [PubMed: 14686461]
3. Gupta M. Does evidence-based medicine apply to psychiatry? *Theor Med Bioeth.* 2007; 28:103–120. [PubMed: 17503219]
4. Becker R, Greig NH. Alzheimer's disease drug development in 2008 and beyond: problems and opportunities. *Curr Alz Res.* 2008; 5:346–357.
5. Becker RE, Greig NH, Giacobini E. Why do so many drugs for Alzheimer's disease fail in development? Time for new methods and new practices. *J Alzheimers Dis.* 2008; 15:303–325. [PubMed: 18953116]
6. Ross JS, Hill KP, Egilman DS, et al. Guest authorship and ghostwriting in publications related to rofecoxib: a case study of industry documents from rofecoxib litigation. *JAMA.* 2008; 299:1800–1812. [PubMed: 18413874]
7. Psaty BM, Kronmal RA. Reporting mortality findings in trials of rofecoxib for Alzheimer disease or cognitive impairment: a case study based on documents from rofecoxib litigation. *JAMA.* 2008; 299:1813–1817. [PubMed: 18413875]
8. DeAngelis CD, Fontanarosa PB. Impugning the integrity of medical science: the adverse effects of industry influence. *JAMA.* 2008; 299:1833–1835. [PubMed: 18413880]
9. Pangalos MN, Schechter LE, Hurko O. Drug development for CNS disorders: strategies for balancing risk and reducing attrition. *Nat Rev Drug Discov.* 2007; 6:521–532. [PubMed: 17599084]
10. 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. [PubMed: 16415710]
11. Kobak KA, Kane JM, Thase ME, et al. Why do clinical trials fail? The problem of measurement error in clinical trials: time to test new paradigms? *J Clin Psychopharmacol.* 2007; 27:1–5. [PubMed: 17224705]
12. Becker RE, Markwell S. Problems arising from generalizing of treatment efficacy from clinical trials in Alzheimer's disease. *Clin Drug Invest.* 2000; 19:33–41.
13. Cogger KO. Rating rater improvement: a method for estimating increased effect size and reduction of clinical trial costs. *J Clin Psychopharmacol.* 2007; 27:418–420. [PubMed: 17632240]
14. Kobak, KA.; DeBrot, DJ.; Engelhardt, N., et al. Site vs. centralized raters in a clinical depression trial. National Institute of Mental Health, New Clinical Drug Evaluation Unit. 46th Annual Meeting; Boca Raton, FL. June 2006;

15. Kahn A, Leventhal RM, Kahn 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. [PubMed: 11799341]
16. Chalmers I. Underreporting research is scientific misconduct. *JAMA*. 1990; 263:1405–1408. [PubMed: 2304220]
17. Esterbrook PJ, Berlin JA, Gopalan R, et al. Publication bias in clinical research. *Lancet*. 1991; 337:867–872. [PubMed: 1672966]
18. Frampton M, Harvey RJ, Kirchner V. Propentofylline for dementia. *Cochrane Database Syst Rev*. 2003; 2:CD002853. [PubMed: 12804440]
19. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*. 2008; 358:252–260. [PubMed: 18199864]
20. Becker RE, Colliver JA, Markwell SJ, et al. A double-blind, placebo-controlled study of metrifonate, an acetylcholinesterase inhibitor, for Alzheimer's disease. *Alzheimer Dis Assoc Disord*. 1996; 10:124–131. [PubMed: 8876775]
21. Becker RE, Colliver JA, Markwell SJ, et al. A six month study of the effects of metrifonate on cognitive decline in Alzheimer's disease. *Alzheimer Dis Assoc Disord*. 1998; 12:54–57. [PubMed: 9539412]
22. Becker RE. Lessons from Darwin: evolutionary biology implications for Alzheimer's disease research and patient care. *Curr Alz Res*. 2005; 2:319–326.
23. 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. [PubMed: 10773186]
24. Lachin JM. The role of measurement reliability in clinical trials. *Clin Trials*. 2004; 1:553–566. [PubMed: 16279296]
25. Woods SW, Gueorguieva RV, Baker CB, et al. Control group bias in randomized atypical antipsychotic medication trials for schizophrenia. *Arch Gen Psychiatry*. 2005; 62:961–970. [PubMed: 16143728]
26. Kobak KA, Kane JM, Thase TE, et al. Reply to comments by Drs Rickels and Robinson. *J Clin Psychopharmacol*. 2007; 27:421.
27. Visser PJ, Scheltens P, Verhey FR. Do MCI criteria in drug trials accurately identify subjects with predementia Alzheimer's disease? *J Neurol Neurosurg Psychiatry*. 2005; 76:1348–1354. [PubMed: 16170074]
28. Becker RE. Lessons from Darwin: 21st century designs for clinical trials. *Curr Alz Res*. 2007; 4:458–467.
29. Schmoor C, Olschewski M, Schumacher M. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Stat Med*. 1996; 15:263–271. [PubMed: 8643884]
30. Raschetti R, Albanese E, Vanacore N, et al. Cholinesterase inhibitors in mild cognitive impairment: a systematic review of randomized trials. *PloS Med*. 2007; 4:1818–1828.
31. Worrall J. Why there's no cause to randomize. *Br J Phil Sci*. 2007; 58:451–488.
32. Kralawish J. Alzheimer's disease-clinical trials and logic of clinical purpose. *N Engl J Med*. 2006; 355:1604–1605. [PubMed: 17035654]
33. Becker RE. Focusing pharmaceutical research on patient care: using clinical trials to develop clinical decision rules. *Clin Drug Invest*. 2002; 22:269–278.
34. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA*. 2003; 290:1624–1632. [PubMed: 14506122]
35. Rich MW. From clinical trials to clinical practice. *JAMA*. 2002; 287:1321–1323. [PubMed: 11886324]
36. Cohen J. Things I have learned (so far). *Am Psychol*. 1990; 45:1304–1312.
37. Hays, WL. *Statistics for Psychologists*. New York, NY: Holt, Rinehardt and Winston; 1963.
38. Butcher J. The hunt for drugs to modify Alzheimer's disease. *Lancet Neurol*. 2007; 6:1038–1039. [PubMed: 18031701]

39. Becker RE. Treating Alzheimer's disease: do clinical trial methods produce unnecessary practitioner dilemmas? *Am J Alzheimers Dis Other Dement*. 2006; 21:81–84. [PubMed: 16634462]
40. 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. [PubMed: 15453011]
41. Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Arch Gen Psychiatry*. 1995; 52:867–871. [PubMed: 7575107]
42. Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessments on power and sample size calculations in clinical trials. *J Clin Psychopharmacol*. 2002; 22:318–325. [PubMed: 12006903]
43. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry*. 1988; 45:742–747. [PubMed: 3395203]
44. Institute of Medicine. Committee on Treatment of Post Traumatic Stress Disorder: Treatment of Post Traumatic Stress Disorder. An Assessment of the Evidence. Washington, DC: The National Academies Press; 2007.
45. Kelloff GJ, Bast RC, Coffey DS, et al. Biomarkers, surrogate end points, and the acceleration of drug development for cancer prevention and treatment: an update. *Clin Cancer Res*. 2004; 10:3881–3884. [PubMed: 15173097]
46. Pincus T, Stein CM. Why randomized controlled clinical trials do not depict accurately long-term outcomes in rheumatoid arthritis: some explanations and suggestions for future studies. *Clin Exp Rheumatol*. 1997; 15:S27–S38. [PubMed: 9266130]
47. Kobak KA, Engelhardt N, Lipsitz JD. Enriched rater training using Internet-based technologies: a comparison to traditional rater training. *J Psychiatr Res*. 2006; 3:192–199. [PubMed: 16197959]
48. Demitrack MA, Faries D, Herrera JM, et al. The problem of measurement error in multisite clinical trials. *Psychopharmacol Bull*. 1998; 34:19–24. [PubMed: 9564194]
49. 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. [PubMed: 16160614]
50. Rickels K, Robinson DS. Why do clinical trials fail? *J Clin Psychopharmacol*. 2007; 27:420–421. [PubMed: 17632242]
51. Hick J, Feldman SR. Eligibility creep: a cause for placebo group improvement in controlled trials of psoriasis treatments. *J Am Acad Dermatol*. 2007; 57:972–976. [PubMed: 17884244]
52. Kirsch I, Deacon BJ, Huedo-Medina TB, et al. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med*. 2008; 5:e45. [PubMed: 18303940]
53. Walsh BT, Seidman SN, Sysko R, et al. Placebo response in studies of major depression: variable, substantial, and growing. *JAMA*. 2002; 287:1840–1847. [PubMed: 11939870]
54. Walsh BT, Sysko R. Placebo control groups in trials of major depressive disorder among older patients. *J Clin Psychopharmacol*. 2005; 25:S29–S33. [PubMed: 16027558]
55. Ferris SH. Are placebo controls still placebo controls? *Neurobiol Aging*. 2006; 27:S1–S7. [PubMed: 16570374]
56. Modell JG, Rosenthal NE, Harriett AE, et al. Seasonal affective disorder and its prevention by anticipatory treatment with bupropion XL. *Biol Psychiatry*. 2005; 58:658–667. [PubMed: 16271314]
57. Andrews G. Placebo response in depression: bane of research, boon to therapy. *Br J Psychiatry*. 2001; 178:192–194. [PubMed: 11230026]
58. Becker RE, Vicari S. Management of Alzheimer's disease: community based care for patients and families. *Home Health Care Consultant*. 1998; 5:11–17.
59. Ulrich CM, Potter JD. Folate and cancer-timing is everything. *JAMA*. 2007; 297:2408–2409. [PubMed: 17551134]
60. Kaurah P, MacMillan A, Boyd N, et al. Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer. *JAMA*. 2007; 297:2360–2372. [PubMed: 17545690]
61. Duhem, P. The Aim and Structure of Physical Theory. New York, NY: Atheneum; 1962.
62. Mani RB. The evaluation of disease modifying therapies in Alzheimer's disease: a regulatory viewpoint. *Statist Med*. 2004; 23:305–314.

63. Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? *JAMA*. 1999; 282:790–795. [PubMed: 10463719]
64. Katz R. FDA evidentiary standards for drug development and approval. *NeuroRx*. 2004; 1:307–316. [PubMed: 15717032]
65. Nordic Council on Medicines. Good Clinical Trial Practice: Nordic Guidelines, NLN Publication No. 28. Uppsala, Sweden: 1989.
66. Food and Drug Administration. Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications. Rockville, MD: FDA, US Department of Health and Human Services; 1988.
67. Hill, AB. Principles of Medical Statistics. New York, NY: Oxford University Press; 1971.
68. Lewis JA, Machin D. Intention to treat: who should use ITT? *Br J Cancer*. 1993; 68:647–650. [PubMed: 8398686]
69. Schwartz D, Lellouch J. Explanatory and pragmatic attribution in therapeutical trials. *J Chron Dis*. 1967; 20:637–648. [PubMed: 4860352]
70. Lee YJ, Ellenberg JH, Hirtz DG, et al. Analysis of clinical trials by treatment actually received: is it really an option? *Stat Med*. 1991; 10:1595–1605. [PubMed: 1947515]
71. Efron B, Feldman D. Compliance as an explanatory variable in clinical trials. *J Am Stat Assn*. 1991; 86:9–17.
72. Wright CC, Sim J. Intention-to-treat approach to data from randomized controlled trials: a sensitivity analysis. *J Clin Epidemiol*. 2003; 56:833–842. [PubMed: 14505767]
73. Bellamy SL, Lin JY, Have TRT. An introduction to causal modeling in clinical trials. *Clin Trials*. 2007; 4:58–73. [PubMed: 17327246]

TABLE 1**Not Widely Acknowledged Potential Compromises to AD Drug Developments**

Sources of CT Unreliability	Representative
CT subject samples	Heterogeneity
	Nonrepresentative sampling
	Sensitivity and specificity of diagnostic criteria
	Sponsors and investigators relax criteria
CT design and implementation	Inaccuracies in observations
	Imprecision in observations
	Excessively large numbers of subjects and sites to provide subjects
	Inadequate training and monitoring at research sites
Protocol deviations	Lack of protocols to control and systematize practices
	Diagnostic criteria disregarded
	Clinical ratings without protocols to control administrations of scales
Inadequate clinical instruments and rater skills	Clinical raters not extensively experienced with the disorder under study
	No prior use of clinically rated outcome measures by the investigator
	Inaccurate ratings
	Imprecise ratings
Placebo group improvements	Careless mistakes
	Possible bias by raters
	Nonspecific responses of subjects to attention and other aspects of CT participation
Undocumented hazards	Falsified data
	Protocol violations

TABLE 2**Representative Changes Implied as Needed to Meet Current Methodological Risks to AD Drug Developments**

More extensive in vitro laboratory, animal, and phases I and II human dose-finding studies before phase III CTs
Ongoing interactions of basic and clinical researchers to support human phases of drug development
Design and execution of clinical studies to provide maximum power without excessive numbers of subjects
Uses of phases I and II to train and qualify raters for participation in phase III
Test raters' abilities using outcome measures for reliability
Recruit, evaluate, and as appropriate to ensure protocol compliance and reliability, eliminate raters and sites
Adapted from Becker and Greig. ⁴

TABLE 3**Representative Risks to CT Integrity**

Drugs may be given
in inappropriate doses
for too short a period
Patients may be noncompliant
In humans, one often cannot know whether the drug has not reached a brain target at adequate concentrations to be effective or is itself ineffective
Will a drug, ineffective with clinical outcomes when used alone, become effective in combinations against complex neuropsychiatric diseases?
Does the timing of an intervention in relation to disease progression affect efficacy?
In a preclinical/clinical difference of response to drug
has the animal model failed to express the target?
is the intended target invalid?
is a failure caused by target differences between bench and animal models and humans?
is the failure an undetected pharmacokinetic factor?

TABLE 4

Possible Criteria for Excluding Subjects From Per-Protocol Analyses Based on the Presence of Methodological Risks to the Integrity of the CT

Subjects with data from raters
who rate placebo-treated patients as improving during the double-blind
with excessive random measurement errors in data
who do not meet “gold standards” of accuracy
Subjects with data from CT imaging and biochemical marker evaluations that do not comply with
protocols for collection of samples
Good Laboratory Practices
do not meet standards using blind analyses of control samples
Subjects with data from sites that
do not comply with CT protocol
do not participate in training
do not meet protocol standards during monitoring for quality of site performance
do not preserve blinding