Neuropsychopharmacology (2010) 35, 2491–2501

www.neuropsychopharmacology.org

Perspective

What Have We Learned about Trial Design From NIMH-Funded Pragmatic Trials?

John March[1], Helena C Kraemer[2], Madhukar Trivedi[3], John Csernansky[4], John Davis[5], Terence A Ketter[6] and Ira D Glick*[,6]

[1]Division of Neurosciences Medicine, Clinical Research Institute, Duke University, Durham, NC, USA; [2]Department of Psychiatry and Behavioral Sciences (Emerita), Stanford University; Department of Psychiatry, University of Pittsburgh; [3]Department of Psychiatry, University of Texas Southwestern Medical Center at Dallas, Dallas, TX, USA; [4]Department of Psychiatry, Northwestern Feinberg School of Medicine, Chicago, IL, USA; [5]Department of Psychiatry, University of Illlinois at Chicago, Chicago, IL, USA; [6]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA

At the 2008 annual meeting of the American College of Neuropsychopharmacology (ACNP), a symposium was devoted to the following question: 'what have we learned about the design of pragmatic clinical trials (PCTs) from the recent costly long-term, large-scale trials of psychiatric treatments?' in order to inform the design of future trials. In all, 10 recommendations were generated placing emphasis on (1) appropriate conduct of pragmatic trials; (2) clinical, rather than, merely statistical significance; (3) sampling from the population clinicians are called upon to treat; (4) clinical outcomes of patients, rather than, on outcome measures; (5) use of stratification, controlling, or adjusting when necessary and not otherwise; (6) appropriate consideration of site differences in multisite studies; (7) encouragement of '*post hoc*' exploration to generate (not test) hypotheses; (8) precise articulation of the treatment strategy to be tested and use of the corresponding appropriate design; (9) expanded opportunity for training of researchers and reviewers in RCT principles; and (10) greater emphasis on data sharing.

Neuropsychopharmacology (2010) **35,** 2491–2501; doi:10.1038/npp.2010.115; published online 25 August 2010

Keywords: clinical pharmacology/clinical trials; drug discovery/development; psychopharmacology; schizophrenia/antipsychotics; depression, unipolar/bipolar; pragmatic design; mood disorder

INTRODUCTION

In the Learning Health Care System, the Institute of Medicine defines evidence-based medicine, thus: 'the decisions that shape the health and health care of Americans should be grounded on a reliable evidence base, will account appropriately for individual variation in patient needs, and will support the generation of new insights on clinical effectiveness' (Olsen *et al*, 2007). To support evidence-based practices, the field is turning to comparative effectiveness research (CER; Johnston and Hauser, 2009; Wang *et al*, 2009), which is the conduct and synthesis of research comparing the benefits and harms of different interventions and strategies to prevent, diagnose, treat, and monitor health conditions in 'real world' settings (Avorn, 2009; Demaria, 2009; Lauer, 2009; Tunis *et al*, 2009). CER consists of: (1) prospective clinical studies, including PCTs; (2) retrospective studies using administrative datasets; (3) decision models; and (4) systematic reviews. As pointed out in a recent article by Philip Wang, the National Institute of Mental Health (NIMH) Deputy Director, CER is a key strategy for guiding healthcare decision-making on the pathway to personalized medicine in mental illness (Wang *et al*, 2009).

In approaching CER, it is useful to first acknowledge that randomized controlled trials can be categorized as having either a pragmatic or an explanatory attitude, with the former required in CER (Armitage, 1998; Thorpe *et al*, 2009; Zwarenstein *et al*, 2008). In contrast to explanatory trials, which ask the question 'can this intervention work under ideal conditions?', PCTs seek to answer the question, 'does this intervention work under usual conditions?' Accordingly, trials with an explanatory attitude are developed specifically to evaluate the efficacy of an intervention (maximizing signal detection) and by a desire to understand the mechanisms by which the intervention is associated with benefits or harms. Conversely, trials with a pragmatic aim (frequently called effectiveness trials in psychiatry) are developed specifically to answer a question faced by decision makers at one or more levels of the health care system from patients and doctors to third-party payers to public policy makers and to identify markers that support

*Correspondence: Dr ID Glick, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, 401 Quarry Road Room 2122, Stanford, CA 94305, USA, Tel: +6 50 799 1583, Fax: +6 50 723 2507, E-mail: iraglick@stanford.edu
Received 8 May 2010; revised 28 June 2010; accepted 29 June 2010

risk stratification (March *et al*, 2005; Tunis *et al*, 2003; Zwarenstein and Treweek, 2009). In this context, the ultimate goal of a pragmatic trial is to reveal which treatment or treatment strategy is best for particular patient subgroups or stratified medicine (Trusheim *et al*, 2007) and, ultimately, for each individual patient or personalized medicine (Garber and Tunis, 2009; Lee and Mudaliar, 2009).

Funded by the NIMH as contracts, the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE; Lieberman, 2007), the Multimodal Treatment Study of Children and Adolescents with Attention-deficit/hyperactivity disorder (MTA; MTA, 1999), Sequenced Treatment Alternatives to Relieve Depression (STAR*D; Rush, 2007), Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD; Sachs *et al*, 2003), and Treatment of Adolescents with Depression Study (TADS; March and Vitiello, 2009) were intended to be pragmatic trials (PCTs) that addressed the comparative effectiveness of widely used treatments in a clinical context where important public health questions remained unanswered and where clear answers might improve public health. At the 2008 annual meeting of the ACNP, a symposium was devoted to addressing the following question: 'what have we learned about the design of PCTs from these long-term, large-scale trials of psychiatric treatments?' While it is possible to consider each of these PCTs separately, eg, (Kraemer *et al*, 2009), the interest of symposium participants was not to dissect each PCT, but rather to understand what has been learned from this set of large PCTs (hereafter referred to collectively as target-PCTs or T-PCTs) in order to inform the cost-effective design of future trials.

In what follows, we will address issues related to conceptualization and articulation of the primary research hypothesis in Section Ensuring that a trial is fully pragmatic, and then discuss the problems engendered by exclusive emphasis on statistical rather than clinical significance in Section Statistical *vs* clinical significance. Choice of relevant populations and of relevant primary outcomes is covered in Sections The population of interest and A clinically relevant outcome. Issues related to personalization, ie, identification of which patients will or will not respond well to a particular treatment, are next discussed in Section Heterogeneity and personalization. Next, we discuss the merits of studies of treatment in sequence as opposed to in combination in Section Treatments in combination or sequence. We conclude by discussing a pragmatic trials infrastructure likely to provide definitive results, at lesser time and lower cost, and comment on the importance of data sharing in Sections Building a sustainable infrastructure for PCTs and Data sharing respectively.

ENSURING THAT A TRIAL IS FULLY PRAGMATIC

It is critically important to assure that PCTs are indeed pragmatic in intent and implementation. Each of the T-PCTs had a pragmatic intent and hewed to most, if not all, of the specific criteria listed for PCTs by March *et al*, (2005) who stated that PCTs should (1) address a straightforward clinically relevant question, (2) have a representative sample of patients and practice settings, (3) provide sufficient power to identify clinically significant effects, (4) use randomization to protect against bias, but not necessarily concealed allocation, (5) have clinical uncertainty regarding the outcome of treatment at the patient level, (6) include assessment and treatment protocols that enact best clinical practices, (7) utilize simple and clinically relevant outcomes, and (8) entail limited patient and investigator burden. On the other hand, simply listing the characteristics of PCTs without clearly stating what is not a PCT allowed trials to include explanatory elements to such a degree that explanatory trials were being labeled as PCTs merely because they broadened the sampling frame.

To address this issue and to help trialists to scale the explanatory to pragmatic continuum, Thorpe *et al* (2009) developed the Pragmatic-Explanatory Continuum Indicators Summary (PRECIS) tool. As shown in Table 1, which presents the PRECIS indicators and relates them to the T-PCTs, these trials included many explanatory elements and, as such, were not fully pragmatic in experimental design or implementation. For example, virtually all of the PCTs addressed multiple different, if related, research questions. For example, the TADS included many cognitive measures that, in part, were conceptualized both as endpoints and as mediators (TADS, 2003) and CATIE included an extensive neuropsychological test battery for much the same purpose (Keefe *et al*, 2006).

It is easy enough to design a PCT optimally to address one research question, for each decision concerning sampling, measurement and design can specifically optimize the answer to that research question. With multiple populations sampled, multiple treatments (separately, in combination, or in sequence), multiple randomizations, and multiple outcome measures, as was true in all the T-PCTs, the power to obtain clear and unambiguous answers to any one research question can be diluted to such an extent that the trial may fail to answer unambiguously any of its research questions. All of the T-PCTs attempted to answer multiple research questions, not one straightforward clinically relevant question.

In a PCT, extended follow-up is usually implemented using survey-research methodology or, better, via extracting clinical data from an electronic health record (Thorpe *et al*, 2009; Treweek *et al*, 2006). None of the T-PCTs met this criterion. Rather, each implemented extended follow-up by the research teams at the cost of many millions of dollars. For some trials, such as TADS, evaluating the added value with respect to remission and minimizing relapse was a primary goal (March and Vitiello, 2009). For others, such as the MTA, extended follow-up yielded an inception cohort that was followed into young adulthood (Molina *et al*, 2009) with limited value with respect to evaluating initial randomized treatments (Hazell, 2009).

As pointed out by Baigent, (1997) 'complexity is rarely a virtue whereas simplicity can lead to the randomization of very large numbers of patients and to results which may lead to worldwide changes in practice within very short periods of time'. On the operational side, some of the T-PCTs—the MTA, TADS, and CATIE in particular—included intensive quality assurance mechanisms intended to maximize treatment and assessment fidelity. Although the measurement-based care model in STAR*D likely improved outcomes by insuring adequate dosing and reducing practice heterogeneity (Trivedi *et al*, 2007), such a level of control

Table 1 Criteria for Pragmatic Clinical Trails and Experience with The T-PCTs

Characteristic	Criterion	T-PCTs
Participant eligibility criteria	All willing participants who have the condition of interest are enrolled, regardless of their anticipated risk, responsiveness, comorbidities, or past compliance, excluding only those who are known to be harmed by one of the treatments.	While all PCTs imposed restrictive entry criteria-based safety and ethical considerations, each PCT used a broad sampling frame.
Experimental intervention flexibility	Instructions on how to apply the experimental intervention are flexible, offering practitioners the type of leeway they would have in ordinary practice in deciding how to formulate and apply it.	While each PCT provided for flexible titration based on clinical endpoints, the titration schedules were constrained to the protocol some highly so, c.f. MTA and STAR*D.
Experimental intervention practitioner expertize	The experimental intervention typically is applied by the full range of practitioners and in the full range of clinical settings, regardless of their expertize, with only ordinary attention to dose setting and side effects.	The MTA, TADS, and CATIE used primary academic clinicians whereas STAR*D and STEP-BD used both academic- and community-based clinicians. All were trained to high quality practice standards.
Comparison intervention	'Usual practice' or the best available alternative management strategy, offering practitioners considerable leeway in deciding how to apply it.	Only the MTA provided a 'treatment as usual' condition.
Comparison intervention practitioner expertize	The comparison intervention typically is applied by the full range of practitioners, and in the full range of clinical interest, regardless of their expertize, with only ordinary attention to their training, experience, and performance.	With the exception of the MTA TAU condition, treatments were provided by a mix of community and academic provides with the aim of testing high quality treatments.
Follow-up intensity	No formal follow-up visits of study individuals at all. Instead, administrative databases (such as clinical databases and mortality registries) are searched for the detection of outcomes.	Each protocol required extended follow-up. For some trials, such as TADS, extended treatment was a primary aim. For others, such as the MTA, extended follow-up resulted in an inception cohort that is of uncertain value.
Primary trial outcome	The primary outcome is an objectively measured, clinically meaningful outcome to the study participants. The outcome does not rely on central adjudication and is one that can be assessed under usual conditions: eg, special tests or training are not required.	Except for STAR*D, which used the clinician-friendly QIDS, and CATIE, which used all cause discontinuation, each PCT used research instruments for primary endpoints.
Participant compliance with 'prescribed' intervention	There is unobtrusive (or no) measurement of compliance and no special strategies to maintain or improve compliance are used.	As exemplified by the use of Adjunctive Services and Attrition Prevention (ASAP) protocols in the MTA and TADS, each PCT implemented research-specific compliance enhancing strategies.
Practitioner adherence to study protocol	There is unobtrusive (or no) measurement of practitioner adherence and no special strategies to maintain or improve it are used.	Each PCT included quality assurance procedures for both treatment and assessment.
Analysis of primary outcome	The analysis includes all patients regardless of compliance, eligibility, and others (the 'intention-to-treat' analysis).	Analyses were uniformly based in the intention-to-treat principle.

slants toward an explanatory attitude in that it limits clinician flexibility, a key element in fully pragmatic trials.

Even now, the definition of PCTs continues to evolve. In this evolution, it is critically important to understand that any protocol-driven action that would or could not be reproduced in clinical practice limits the value of the study with respect to influencing clinical practice.

Recommendation no. 1

Although very few studies can or should be completely devoid of explanatory elements, PCTs in the future will be well served by systematically evaluating the design and implementation of PCTs against criteria for PCTs such as those articulated by March *et al* (2005) and, from an implementation perspective, the PRECIS tool, to produce results most relevant to actual clinical practice.

STATISTICAL *vs* CLINICAL SIGNIFICANCE

A common problem in all these large PCTs is the continued exclusive reliance on statistical significance, and absence of consideration of effect sizes in both study design and in reporting results. Two problems result: first, statistically significant results may not be clinically significant (particularly with large sample sizes) but interpreted as such; second, results that are not statistically significant are often misinterpreted as indicating clinical equivalence, when in fact the study simply failed to answer its research questions.

Over the last 20 years, it has become clear that an exclusive focus on '*p*-values' is a major problem (Cohen, 1995; Kline, 2005; Nickerson, 2000; Shrout, 1997; Thompson, 1999; Wilkinson and The_Task_Force_on_Statistical_Inference, 1999). Arguments have even been presented for banning the '*p*-value' entirely (Cohen, 1995; Hunter, 1997; Kline *et al*, 2005; Nickerson, 2000; Shrout, 1997; Thompson, 1999; Wilkinson and The_Task_Force_on_Statistical_Inference, 1999). However, currently the emphasis has been on presenting clinically interpretable effect sizes along with every '*p*-value' to be used to judge not only the statistical significance, but also the clinical significance in the T1 *vs* T2 comparison (Altman *et al*, 2001; Grissom and Kim, 2005; Kraemer and Kupfer, 2006; Wilkinson and The_Task_Force_on_Statistical_Inference, 1999).

Traditionally, PCTs are designed as non-inferiority studies, so that the probability of a 'statistically significant' result is $<5\%$ when the true effect is zero, and is $>80\%$ whenever the true effect size exceeds a clinical situation-dependent critical value. The problem lies in establishing such a critical value, which is often done arbitrarily, with results that in turn undermine efforts to demonstrate significant results (Maxwell, 2004). Which effect size is most easily interpretable in terms of clinical significance, and how large an effect size is needed for clinical significance in any specific situation (the critical effect size at which power calculations are done) are questions for which there is not yet general consensus. In general, however, PCTs, especially those comparing active treatments, should be powered to detect small to moderate effects, an indicator that at least some patients within the population may be better benefited by one of the treatments.

There remain major questions about the value of PCTs designed to establish equivalence of two treatments. Some (Kraemer and Kupfer, 2006; Lavori, 2000) argue that PCTs should be powered to detect the smallest reasonable clinically meaningful effect size (ie, non-inferiority studies) rather than aim to demonstrate clinical equivalence (Bristol, 1999; Ebbutt and Frith, 1998; Jones *et al*, 1996), which requires much larger sample sizes. In essence, if a newer drug can't beat an older drug at a minimal clinically meaningful level, what practical value does it add to our therapeutic armamentarium? Often the response to this question is that the newer drug may not be more effective in reducing symptoms in the entire population but only in a specific subpopulation, or may have a better side effect profile, be more convenient to use, or cost far less. However, in that case, the PCT should test such a more complex non-inferiority hypothesis and not focus merely on equivalence for symptom reduction. However, this contentious issue is resolved, the PCT must be designed to achieve its stated goal, be it the usual type of study needed to establish non-inferiority, or the much larger studies necessary to establish equivalence or clinical superiority.

Table 2 presents the approximate sample size per treatment group for comparing two treatments, using a two-tailed 5% test and requiring 80% power to detect various determinations of the critical effect size. Even in large PCTs, it should be rare to see sample sizes larger than 400 patients per treatment group (the exception being in prevention studies done in the general population where the vast majority will not have the event even in absence of intervention). Likewise, it should be rare to see sample sizes smaller than 30 per treatment group, as such sample sizes taken from the heterogeneous populations in PCTs will typically result in failed trials.

In reporting the results of PCTs, it would preferable to report the 95% confidence intervals for a clinically interpretable effect size in addition to the '*p*-values'. None of the T-PCTs did so uniformly. Then T1 and T2 are shown to be 'statistically significantly different' when the null value of the effect size is not contained in the confidence interval. On the other hand, T1 and T2 would be shown to be 'clinically equivalent' if the 95% confidence interval for an effect size such as number needed to treat does not include any effect size exceeding the critical value. It is possible (although rare) to have a 'statistically significant' result when T1 and T2 are 'clinically equivalent', because the confidence interval contains neither the null effect size nor any effect sizes exceeding the critical value. T1 will be shown 'clinically superior' to T2 if the confidence interval contains no effect size below the critical effect size. Finally, a failed PCT is one in which the confidence interval contains both the null effect size (a 'non–statistically significant' result) as well as effect sizes exceeding the critical value, thus demonstrating neither superiority nor equivalence nor non-inferiority. Such wide confidence intervals are usually the result of inadequate sample sizes. Then, the state of knowledge after the study remains just as it had been before the study.

It is possible for a well-designed trial to fail, either because of misleading information in the literature on which rationale and justification for the trial was based, or simply because of bad luck. However, failed trails are most often the result of poor design, poor measurement, and inadequate sample size.

Table 2 Approximate Sample Size per Group to Compare two Treatments using a 5% Two-Tailed test, to have at Least 80% Power to Detect any Effect Size Exceeding The Indicated Critical Effect Size

Effect size (d): standardized mean difference (normally distributed outcomes)	Risk difference (success/failure outcome)	Number needed to treat (NNT)	Sample size per treatment group
0.1	0.056	17.7	1560
0.2	0.112	8.8	390
0.3	0.185	5.9	174
0.4	0.223	4.5	100
0.5	0.276	3.6	63
0.6	0.329	3.0	45
0.7	0.379	2.6	33
0.8	0.428	2.3	25
0.9	0.475	2.1	21
1.0	0.521	1.9	17

Recommendation no. 2

In proposing and designing a PCT, careful thought should be given to what effect sizes are of clinical significance, ie, the critical value of the effect size. Then, if the goal of the PCT is to demonstrate non-inferiority, it should be designed with a 5% two-tailed significance level (as is typical), and to have more than 80% power to detect any effect size exceeding that critical value. Reporting the results using a 95% confidence interval on the effect size, as well as the *p*-value, will clearly show that (a) T1 and T2 are significantly different from each other, and/or (b) T1 and T2 are clinically equivalent, or (c) the study was underpowered to demonstrate either superiority or equivalence. Use of a confidence interval would also indicate the most likely effect size and thus provide guidance as to whether or not the effect is of clinical significance.

THE POPULATION OF INTEREST

A key distinction between a PCT and an explanatory (efficacy) trial involves the choice of the population to be sampled (Hoagwood *et al*, 1995). In general, explanatory trials exclude many (sometimes most) of the patients that

clinicians treat in everyday practice (Humphreys *et al*, 2005). Thus, the results of explanatory trials do not necessarily apply to the patients clinicians treat. In contrast, PCTs must exclude only those patients whom they are ethically obliged to exclude (eg, those who refuse informed consent, those who would be harmed by any of the treatments, and those that previous research indicate could not be helped by the treatments in the PCT.) Beyond that, PCTs should enroll patients in the settings in which they are found and should use the treating clinicians as the providers (Glasgow *et al*, 2005; March *et al*, 2005; Tunis *et al*, 2003).

The results of any trial apply only to those in the population represented by the sample. It may well be that in the population represented in an efficacy study, T1 may be superior to T2, whereas in the types of complex clinical populations represented in an effectiveness study, T1 and T2 may be clinically equivalent or T2 may be superior to T1. Thus, although the results of such efficacy studies affect regulatory decision making, they should probably not affect clinical decision making (although they often do), as they usually deal with narrowly defined populations using investigational protocols optimized for research signal detection rather than for clinical practice. Because the type of sample recruited into a PCT is heterogeneous and those sampled into an explanatory trial are often selected as those most likely to respond better to T1 than to the control condition, to be free of potentially confounding comorbidities, and to be less likely to drop out, larger treatment differences are more commonly seen in efficacy studies, one reason why PCTs sample sizes must be larger.

All the T-PCTs tended toward the pragmatic end of the sampling frame. If anything, criticisms tend to focus on whether the inclusion criteria were wide enough. Most of the patients recruited into these trials were recruited in specialty mental health settings. Yet many patients with mental health needs are identified and treated in primary care settings. Some of the T-PCTs imposed severity criteria for entry. Yet many patients with mental health needs may not present with high severity scores. In future studies, consideration might be accorded to the full spectrum of patients with the indication that might be addressed by the treatments in the study, preferably recruited from practice, not research settings. If later, it is found that the treatments do not 'work' for certain patients (eg, if clinical setting, initial severity moderate the effects of treatment), the criteria for entry for future PCTs would require exclusion of such patients, but the exclusion would then be evidence-based, rather than based on accessibility or clinical intuition.

All of these T-PCTs involved multiple treatments. Traditionally, the approach has been to include only those willing to accept randomization to all the treatments considered in the PCT (as was true in MTA). When there is randomization to multiple treatments, excluding those who refuse any one treatment not only raises the chance of introducing a sampling bias relative to the clinical population of interest, but also makes recruitment much more difficult (and thus the study becomes longer and more costly).

To deal with this problem, STAR*D introduced the concept of 'equipoise randomization' (Lavori *et al*, 2001). Each eligible patient was asked which of the treatments she/he was willing to accept, and was included in the PCT as long as there were choices. Then, each pair of treatments was compared on that subsample, willing to consider that choice. Equipoise randomization has the advantage of providing clues as to which treatments are or are not acceptable to which types of patients, and why certain treatments are not acceptable to certain types of patients. The analyses of the results in such studies are necessarily more complicated, but the quality of the recommendations such a PCT might generate may be more clinically useful.

Recommendation no. 3

Pragmatic PCTs should continue to focus on sampling the population clinicians are required to treat (effectiveness) in the settings in which they are typically seen excluding only those patients unwilling to participate and those already known to be harmed or not benefited by one or more of the treatments being studied. Where multiple treatments are being studied, equipoise randomization should be considered.

A CLINICALLY RELEVANT OUTCOME

One of the persistent major problems in many PCTs is that of multiple outcomes. It has long been known that testing multiple outcomes separately without adjustment of *p*-values proliferates false positive results (type I error). But with statistical adjustments (eg, Bonferroni or false detection rate (Benjamini and Hochberg, 1995)), proliferation of false negatives (type II error) is likely unless the sample size is increased commensurate with the number of outcome measures and their intercorrelations.

Nevertheless, as is the case for each of the T-PCTs, multiple primary outcomes are common. When there are statistical adjustments for multiple testing to protect type I errors as well as an increased sample size to preserve power, the results on the different outcome measures may conflict, leaving the clinician in doubt as to whether T1 is better than T2, T2 is better than T1, or they are equivalent, another version of a failed study.

Given the effect size for each of multiple outcomes separately, it is not possible to assess which of T1 or T2 is clinically preferable or how strongly, for there are two crucial pieces of information missing: (1) the correlation between harms and benefits in each treatment (does harm happen to the same patients who experience benefit or to different patients?), and (2) the balance in clinical importance of the two outcomes (if benefit and harm happen to the same patient, which, if either, is predominant?).

As CATIE indicated in its justification for the primary outcome (Lieberman *et al*, 2005), the purpose of a PCT is not to assess the differential effects of T1 and T2 on different outcome measures, but to assess those effects on patients, each of whom experience all the different outcome measures. CATIE suggested that it is the benefit–harm balance within each individual that needs to be considered in making such a decision. They attempted to do this by using time to failure of a treatment, with failure defined as the time point when, from the point of view of the clinician and patient (both blinded as to which treatment was being used), the clinical harm outweighs the clinical benefit. Then the outcome of the PCT would be based on this single integrated clinical outcome measure, and information on specific

benefits and harms would be presented descriptively as they are now, not to change the conclusion of the study, but to illustrate and amplify the interpretation of that conclusion. The problem here was that many patients got a partial response (ie, were not treatment failures), and they and/or their families with their study clinician's assent decided to move to the next phase.

Time to failure of a treatment as an integrative outcome measure also addresses other important issues in PCTs. In PCTs where the primary outcome is determined only at the end of the study period, patients for whom it is clear early that harm outweighs benefit are encouraged to stay on the assigned treatment to avoid sample attrition and missing data. The analysis of any trial results requires analysis 'by intention to treat', ie, every patient randomized is included in the analysis in the treatment assigned, and non-random dropouts and missing data are problematic.

The ethics of keeping patients on a treatment that is clearly failing are questionable. In many cases, patients indicate such failure by simply dropping out of the PCT. Statistical analyses to deal with missing data and dropouts are based on the assumption that such dropouts are 'missing at random', which is clearly not true when the dropout is becauseof failure of the treatment. Thus, the value of attempting to keep patients who are clearly failing on an assigned treatment is questionable from ethical and pragmatic, as well as scientific reasons.

With time to treatment failure, as soon as it is clear that the treatment is failing, the primary outcome is determined. The patient is removed from the study, to be given standard clinical treatment. Both ethical and scientific needs are satisfied.

Admittedly, this strategy will not cover many situations faced with multiple measures in future PCTs. However, there are other existing approaches that integrate the measures of benefit and harm, and many more that could be developed, were there a commitment to use such integrated outcome measures.

Recommendation no. 4

In addition to not burdening patients and clinicians with multiple unnecessary and often confusing outcome measures, the development of methods should be encouraged to integrate clinical consideration of benefits and harms to individual patients (thus 'personalizing' the outcome measure), so that the comparison of T1 *vs* T2 would result in only one test comparing the two treatments, one effect size and its confidence interval, but one test that would take into appropriate consideration the effect on individual patients of all harms and benefits of clinical importance.

HETEROGENEITY AND PERSONALIZATION

By definition, PCTs sample a patient population heterogeneous in numerous dimensions. Most of this heterogeneity is of little concern in designing a PCT. If the research question is 'what is the effect size of T1 *vs* T2 in the population sampled?', simple randomization to T1 and T2 is all that is required to obtain the answer. Yet many researchers and reviewers strongly urge either 'controlling for' certain baseline variables (by manipulations such as stratifying the sample), or 'adjusting for' such baseline variables (by inclusion in some mathematical model). However, if one 'controls for' or 'adjusts for' a baseline variable unrelated to treatment response, or 'controls for' or 'adjusts for' a baseline variable incorrectly (eg, using a model that ill fits the clinical situation), there is not only a loss of power but the results may be biased and generalization to the population of interest may be impaired. All of the T-PCTs suffered such problems.

Recommendation no. 5

Stratification of the sample should be done only when there is empirical justification for a moderator hypothesis, in which case, the sample size would need to be large enough to detect a clinically significant moderator effect. Such a sample size is usually much larger than that needed without stratification.

In a multisite study, site is always a special case. There is always rationale and justification to hypothesize site as a possible moderator of treatment response, for different sites recruit from different catchment areas and the patient populations may be different; entry, treatment, and assessment are carried out by different research staffs. All the T-PCTs were multisite PCTs.

There are two important reasons for a multisite study: (a) to generate a large enough sample size for adequate power to test the hypothesis and (b) to examine the generalizability of the conclusions over multiple sites (site moderating treatment). In designing a multisite PCT, each site should replicate the same design, which means that either the design must be a simple one (few cells in the design) or the sites should be carefully selected to be able to generate sufficient numbers of patients to replicate a complex design. Then, the analysis of the results of a multisite study must always include consideration of site and site by treatment interactions in order to test the generalizability of the results over sites.

Despite training to implement a common protocol and to provide checks on the fidelity to those protocols excessive relative to that expected for a PCT, site was an important factor in the majority of these T-PCTs where site was evaluated. For example, the MTA had six sites, four cells (treatments) in the design and considered site and site by treatment interactions in their analyses. In their primary outcome, no significant site-moderation effects were detected, although large site differences were detected. In contrast, CATIE had 57 sites, 16 cells in an unbalanced design. Many sites had fewer patients than there were cells in the design, and every site had one or more empty cells. Consequently, CATIE could not fully consider site or site by treatment interactions that were likely present. In the TADS, site differences also emerged (March and Vitiello, 2009).

Multisite PCTs are generally more costly than are single-site PCTs. The larger the number of sites the more costly is the study, for separate research staffs must be supported at the different sites. In addition, processes such as randomization, data management, data analysis, etc. need to be centralized, which means further support for an executive committee to oversee the study as a whole, as well as a central data unit, both supported separately from the individual sites. In all of these studies, the clinical burden and research burden were well in excess of what is typical in clinical practice, which in turn dramatically inflated costs.

For example, quality control measures for the treatments in the MTA and TADS were very much like classical explanatory studies and accounted for as much as 15% of trial expense. Reducing such burden, which is practical (Eisenstein *et al*, 2005), should be a key goal of pragmatic trials.

Recommendation no. 6

To optimize considerations of generalizabilitiy, the number of sites should be large, but to maximize the power and to minimize cost, it is preferable to have as few sites as possible, each contributing adequate numbers of patients to the PCT. The number of patients per site must be adequate to have minimally two patients per cell of the design. Then analysis of results should include consideration of site and site by treatment effects.

There are, of course, numerous possible moderators of treatment response beyond site. MTA had an '*a priori*' hypothesis that comorbid anxiety disorder moderated the effect of treatment, a hypothesis that was supported (March *et al*, 2000; The_MTA_Cooperative_Group, 1999). However, '*post hoc*' exploration (ie, not driven by the '*a priori*' hypothesis, the study was specifically designed to test) suggested that the strongest moderator of treatment response was the presence of mental health problems in the parents of the ADHD patients and initial severity of symptoms (Owens *et al*, 2003), each of which is a possible risk factor for comorbid anxiety disorder. In '*post hoc*' analysis, previous medication was suggested as a moderator of treatment in CATIE (Essock *et al*, 2006). TADS also '*post hoc*' investigated predictors and moderators of acute outcome (Curry *et al*, 2006).

In general, age, gender, ethnicity, educational level, severity, and/or comorbidity at baseline, are all common possibilities as moderators. As Garber and Tunis (2009) point out, genotypes, biomarkers, and parameters of brain structure and function may also moderate treatments. Clearly, the number of possible moderators is huge, but the number of actual moderators for any treatment choice is probably small. Moreover, the factors that moderate the treatment effects of T1 *vs* T2 are not necessarily the same as those that moderate the treatment effects of T1 *vs* T3, or T2 *vs* T3. Finally, the factors that moderate the treatment effect of T1 *vs* T2 may be different for men and women, for young and old, in general medical practice *vs* a specialty mental health clinic.

Designing a PCT to demonstrate that a specific baseline variable is a moderator for a certain pair of treatments in a certain population is relatively easy. Generating the rationale and justification to propose that specific baseline variable as a moderator is very difficult. Such rationale and justification can only come from '*post hoc*' exploratory studies, using the data accumulated during a PCT comparing T1 *vs* T2 in a population. In fact, one can argue, as has been done for biomarkers (Davis *et al*, 2008), that the identification (in contrast to validation) of moderators is intrinsically exploratory, whereas studies focused on validating moderators, biological or otherwise, can only be designed and powered using the information from such exploratory studies.

Recommendation no. 7

To move toward stratified/personalized medicine, '*post hoc*' exploration of the data resulting from any PCT to detect possible moderators should be encouraged to generate moderator hypotheses to be tested in future PCTs with results reported as hypotheses yet to be tested, not as conclusions. Identification of moderators is an important advantage of pragmatic trials, as they include greater sample variability and larger sample sizes necessary for this purpose. This means that in designing PCTs, potential moderators should be considered '*a priori*' and efforts made to assess them that are consistent with limited patient and investigator burden, eg, single or a few questions rather than long questionnaires.

TREATMENTS IN COMBINATION OR SEQUENCE

One of the recurring themes motivating the T-PCTs is that of combining or sequencing treatments already shown to be effective, but not effective enough. The MTA and TADS investigated whether combined two such treatments worked better than either alone. STAR*D began with all patients on one such treatment, then investigated the effectiveness of switching, augmenting, or combining treatments for those that did not respond. CATIE's phase 2 proposed switching treatments for those who failed the selected treatment in phase 1. One of the STEP-BD randomized studies assessed the effectiveness of three different adjunctive treatments in patients with bipolar depression already resistant to a mood stabilizer plus an antidepressant (Nierenberg *et al*, 2006).

The effect sizes for individual treatments may be small because of unrecognized moderators of treatment, or because of poorly designed studies, or simply because the individual treatments available are, in fact, generally weak. The foregoing discussion has focused on the first possibilities. What follows focuses on the last.

If individual treatments available are indeed effective but weak, the question arises as to whether combinations of treatments either presented simultaneously, or in sequence, might be stronger than any individual treatment alone. Certainly, such choices are the only options until completely new treatments are developed.

MTA and TADS evaluated the effect of combined treatments *vs* each individual treatment. For two treatments T1 and T2, there would be three treatment groups to be compared: T1 + T2, T1, T2 (plus possibly a usual care control group). Patients were randomized to one of these groups. If comparing T1 *vs* T2 would require 100 per treatment group to detect any clinically significant effect, a total of 200 patients, one now requires 300 (or 400 if a control group were included).

However, even with only two treatments, there are numerous other possibilities, eg:

- Randomly assign to T1 and T2. After a period of time, re-randomize each patient either to stay on the originally assigned treatment or to be switched to the other treatment. In absence of dropouts, this would be the same as initially randomizing to one of four treatment strategies: T1 in both time periods, T1 followed by T2, T2 followed by T1, T2 in both time periods. With dropouts, there are essentially three trials corresponding to the three points of randomization, each on the population eligible and willing to be randomized.

- Randomly assign to T1 and T2. After a period of time, leave those who respond on the originally assigned treatment, those who do not are switched to the other treatment (two treatment strategies: switch strategy).
- Randomly assign to T1 and T2. After a period of time, leave those who respond on the originally assigned treatment, those who do not are given the other treatment as well (two treatment strategies: augment strategy).
- Randomly assign to T1 and T2. After a period of time, leave those who respond on their originally assigned treatments, and randomly assigned those who fail in each treatment to the other treatment (switch) or add the other treatment (augment). (Three trails, each comparing two treatment strategies on different sub-populations.)

Each of these poses a different research question and might require different designs and sample sizes for adequate power. Here, we have considered only two treatments, T1 and T2. For PCTs with more than two treatments, the number of such options becomes greater, and the necessary sample sizes become even larger to address switching, augmentation, etc. in multiple stages, especially if the switches or augmentations are done conditionally, randomly or not, on the response of patients to earlier assigned treatments.

When the focus of a PCT is on testing personalized treatment strategies, it is preferable to employ experimental designs appropriate to adaptive treatment strategies sometimes called treatment algorithms, stepped care models, or dynamic treatment regimes (Murphy *et al*, 2007). An adaptive treatment strategy is a sequence of individually tailored decision rules that specify whether, how, and when to alter the intensity, type, or delivery of treatment at critical decision points in the treatment of adolescent psychiatric disorders. To develop adaptive treatment strategies that inform and improve the clinical management of mentally ill patients requires a clinical trial design that allows the investigative team to consider which treatments to choose, the order or sequencing of treatments, the timing of changes in treatment, and the ability of clinicians to make use of measures of benefit, harm, and acceptability (adherence) collected during treatment to make further treatment decisions. The recently developed statistical design methodology, Sequential Multiple Assignment Randomized Trials (SMARTs), created by Susan Murphy and colleagues (Collins *et al*, 2004; Collins *et al*, 2007; Murphy *et al*, 2007) and also by Lavori and colleagues (Dawson and Lavori, 2004, 2008; Dawson *et al*, 2007; Lavori and Dawson, 2004) give promise of an efficient and ecologically valid format through which such issues can be addressed.

On the other hand, large and costly multitreatment, multistage studies are very unlikely to ever be independently validated or confirmed. If such a PCT fails, or there is some sampling, measurement, or analytical error that produces a false-positive result, deleterious effects on clinical decision making and future research may persist for a very long time. When the primary aim is CER, multiple-staged PCTs, adequately powered with smaller sample sizes, each building on its predecessors, may be better both scientifically and economically, than one large PCT with multiple stages.

Recommendation no. 8

As, even with only two treatments, there are multiple different ways of articulating the relevant research hypothesis, with each articulation requiring different designs and sample sizes, it is crucial that the precise articulation with greatest clinical import be addressed. A sequence of PCTs each building upon the results of earlier PCTs, a SMART design, or a simple design comparing treatment strategies (rather than treatments) might be a better investment than one large-scale PCT with multiple treatment, stages, and combinations of treatments, having a greater chance of failure.

BUILDING A SUSTAINABLE INFRASTRUCTURE FOR PCTs

Many of the issues here discussed have long been known among PCT specialists. However, most PCTs are reviewed, designed, executed, and reported by non-specialists, often with very little training and experience in the complexities of PCT design and analysis. It is not unusual that, whereas those who propose a PCT are required to provide the rationale and justification for the hypotheses to be tested, and to present a design and analytical plan valid and powerful to test that hypothesis, the reviewers of a PCT propose modifications to the research question without rationale and justification, and without acknowledging the necessary modifications to the design and analytical plan necessary to test the modified hypotheses. A common example: review committees often require stratification or adjustment for covariates not known to influence response to treatment. This not only changes the research question, but also imposes a requirement for larger sample sizes than available or affordable under the proposed budget. However, where is the training necessary for successful PCT design and execution available either to proposers or to reviewers?

There are many available options. Options for training in PCT methods already include NIMH-supported workshops either at the New Clinical Drug Evaluation Unit (NCDEU) or in Washington DC; web-based presentation on specific PCT issues; study groups at professional meetings (such as the one from which this discussion grew at ACNP) critically examining completed studies to identify what is successful and unsuccessful in such studies. Clinical trials coordination, including project management site management, data management and statistical services, safety reporting, web site development and maintenance, human patients protection, network services, and dissemination activities, is critical to pragmatic trials, and those proposing new PCTs might avail themselves of training by affiliating with organizations with experience in and success in conducting PCTs. Finally, researchers might avail themselves of software tools such as that developed by the Pragmatic RAndomized Controlled Trials in Health Care (PRACTIHC) group that supports the writing of protocols for pragmatic randomized controlled trials (Treweek *et al*, 2006).

As outlined in the National Institutes of Health (NIH) Roadmap, and strategic plan, PCT networks are envisioned as a means to reduce costs associated with launching

multisite studies, while increasing patient and physician participation in clinical research. To enhance the field's capacity to conduct public health relevant CER and, in so doing, to enhance the evidence base to the benefit of mentally ill patients and their families, it also will be necessary to develop the PCT infrastructure. The traditional approach used in these PCTs of assembling an '*ad hoc*' network for each PCT will not suffice. Inefficiencies in network construction and in clinical trial data collection cause delays, increase costs, and reduce clinician participation in medical research (McCourt *et al*, 2007). Established networks of sites that could be used as the clinical research platform for multiple different PCTs conducted by different researchers would enhance the ability to conduct clinical studies better, faster, and cheaper. Using electronic medical records as a platform for CER in such a network of sites would significantly reduce errors, inefficiencies, and costs by eliminating double and some times triple data entry, data justification, and much of the need for site monitoring (Afrin *et al*, 2003; Gersing and Krishnan, 2003). Doing this may also improve clinical record keeping at those sites, and would also enhance clinical dissemination of research findings. Paying such sites only for the cost of research above the cost of clinical care (as in the T-PCTs) produces additional savings.

Thus, unlike the T-PCTs, which together cost over 100 million dollars (Wang *et al*, 2009), PCTs implemented in an established network of sites would cost far less, and afford many more researchers the opportunity to design and conduct PCTs.

Moreover, as called for in the NIH Roadmap, (Zerhouni, 2003; Zerhouni, 2006) cooperation between academic medical centers and community partners, both serving as sites in such networks, has become a mutual necessity. In addition to the Clinical Science Translation Award consortia funded by the Roadmap (Berglund and Tarantal, 2009; Zerhouni, 2007), the NIMH-funded networks such as the Depression Trials Network and Bipolar Trials Network, and the recently established National Network of Depression Centers, the Schizophrenia Trials Network, along with networks like Kaiser (Glasgow *et al*, 2005) and MindLinc (Gersing and Krishnan, 2003), make it possible to integrate the participation of community clinics with a prestigious academic health center (AHC) so as to effectively and efficiently conduct data mining and retrospective/prospective studies, as well as to conduct PCTs.

Recommendation no. 9

To reduce costs associated with launching multicenter studies while increasing patient and physician participation in clinical research, PCTs should rapidly transition away from trial-specific networks to a comprehensive infrastructure that includes both clinical practice sites and AHCs.

DATA SHARING

The product of any research study is not the set of conclusions drawn by the investigators reported in publication; it is the set of data collected in the course of that research study, the evidence to support those conclusions. It is important that the datasets resulting from each adequately powered PCT be explored both to check the internal validity of the conclusions of the PCT (to prevent unrecognized false positives in the research literature), and to generate hypotheses in more breadth and depth for future PCTs and the information necessary to design well-powered PCTs to test those hypotheses. In that way, each PCT answers the research question it is designed to answer and provides the platform for future PCTs. Thus the value of each PCT can be leveraged to advance scientific knowledge.

Consequently, the exploration of each PCT dataset should be available not only to the participants of the study, but to other researchers as well. Funders, such as NIH, are often reluctant to approve for funding validation studies because they are not novel enough, and exploratory studies because they are considered too novel (ie, 'fishing expeditions'). As a result, PCTs are often based on findings in the research literature that are false positives, but never identified as such because of absence of validation of any kind, and are often weakly designed because of absence of information that would be readily available in exploring past studies. If it were required that researchers share the data that underlie each publication of a PCT with other qualified researchers immediately upon publication, and to make all the data from a PCT available, say, 3 years after completion of the PCT, both such problems would be mitigated without the investment of major further funding. The knowledge that data will be shared also often improves data checking and documentation as well.

Recommendation no. 10

The NIMH should continue (as it has for the T-PCTs) to make publicly funded clinical trial data sets available and should expand this practice to other NIMH-funded studies in order to provide the best possible basis for design new PCTs.

SUMMARY

The completion of any well-designed, well-executed and carefully reported research study has two effects. The better recognized effect is that of extending the knowledge in that particular area of research. Here, we have focused on the second effect. That is with each well-done study, researchers learn a little more as to how to do such studies better in the future, better both in terms of scientific quality, but also in terms of cost efficiency. What is learned may warn against certain tactics, multiplicity of research hypotheses that cannot be optimally addressed with a single research design, or ambiguous conclusions that can mislead subsequent clinical decision making. However, what may also be learned are better methods of research design, such as the moderator/mediator testing of the MTA, the equipoise randomization of the STAR*D, or even an integrated outcome measure of CATIE.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

Afrin LB, Oates JC, Boyd CK, Daniels MS (2003). Leveraging of open EMR architecture for clinical trial accrual. *AMIA Annu Symp Proc* 16–20.

Altman DG, Schulz KF, Hoher D, Egger M, Davidoff F, Elbourne D *et al* (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* **134**: 663–694.

Armitage P (1998). Attitudes in clinical trials. *Stat Med* **17**: 2675–2683.

Avorn J (2009). Debate about funding comparative-effectiveness research. *N Engl J Med* **360**: 1927–1929.

Baigent C (1997). The need for large-scale randomized evidence. *Br J Clin Pharmacol* **43**: 349–353.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57**: 289–300.

Berglund L, Tarantal A (2009). Strategies for innovation and interdisciplinary translational research: removal of barriers through the CTSA mechanism. *J Investig Med* **57**: 474–476.

Bristol DR (1999). Clinical equivalence. *J Biopharm Stat* **9**: 549–561.

Cohen J (1995). The earth is round ($p<0.05$). *Am Psychol* **49**: 997–1003.

Collins LM, Murphy SA, Bierman KL (2004). A conceptual framework for adaptive preventive interventions. *Prev Sci* **5**: 185–196.

Collins LM, Murphy SA, Strecher V (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *Am J Prev Med* **32**(5 Suppl): S112–S118.

Curry J, Rohde P, Simons A, Silva S, Vitiello B, Kratochvil C *et al* (2006). Predictors and moderators of acute outcome in the Treatment for Adolescents with Depression Study (TADS). *J Am Acad Child Adolesc Psychiatry* **45**: 1427–1439.

Davis M, Hanson S, Altevogt B (2008). *Neuroscience Biomarkers and Biosignatures: Converging Technologies, Emerging Partnerships: Workshop Summary*. National Academies Press: Washington, DC.

Dawson R, Lavori PW (2004). Placebo-free designs for evaluating new mental health treatments: the use of adaptive treatment strategies. *Stat Med* **23**: 3249–3262.

Dawson R, Lavori PW (2008). Sequential causal inference: application to randomized trials of adaptive treatment strategies. *Stat Med* **27**: 1626–1645.

Dawson R, Lavori PW, Luby JL, Ryan ND, Geller B (2007). Adaptive strategies for treating childhood mania. *Biol Psychiatry* **61**: 758–764.

Demaria AN (2009). Comparative effectiveness research. *J Am Coll Cardiol* **53**: 973–975.

Ebbutt AF, Frith L (1998). Practical issues in equivalence trials. *Stat Med* **17**: 1691–1701.

Eisenstein EL, Lemons II PW, Tardiff BE, Schulman KA, Jolly MK, Califf RM (2005). Reducing the costs of phase III cardiovascular clinical trials. *Am Heart J* **149**: 482–488.

Essock SM, Covell NH, Davids SM, Stroup TS, McEvoy JP, Rosenheck RA *et al* (2006). Effectiveness of switching antipsychotic medications. *Am J Psychiatry* **163**: 2032–2033.

Garber AM, Tunis SR (2009). Does comparative-effectiveness research threaten personalized medicine? *N Engl J Med* **360**: 1925–1927.

Gersing K, Krishnan R (2003). Clinical computing: Clinical Management Research Information System (CRIS). *Psychiatr Serv* **54**: 1199–1200.

Glasgow RE, Magid DJ, Beck A, Ritzwoller D, Estabrooks PA (2005). Practical clinical trials for translating research to practice: design and measurement recommendations. *Med Care* **43**: 551–557.

Grissom RJ, Kim JJ (2005). *Effect Sizes for Research*. Lawrence Erlbaum Associates: Mahwah, New Jersey.

Hazell PL (2009). 8-year follow-up of the MTA sample. *J Am Acad Child Adolesc Psychiatry* **48**: 461–462.

Hoagwood K, Hibbs E, Brent D, Jensen P (1995). Introduction to the special section: efficacy and effectiveness in studies of child and adolescent psychotherapy. *J Consult Clin Psychol* **63**: 683–687.

Humphreys K, Weingardt KR, Horst D, Joshi AA, Finney JW (2005). Prevalence and predictors of research participant eligibility criteria in alcohol treatment outcome studies, 1970–98. *Soc Study Addict* **100**: 1249–1257.

Hunter JE (1997). Needed: a ban on the significance test. *Psychol Sci* **8**: 3–7.

Johnston SC, Hauser SL (2009). Comparative effectiveness research in the neurosciences. *Ann Neurol* **65**: A6–A8.

Jones B, Jarvis P, Lewis JA, Ebbutt AF (1996). Trials to assess equivalence: the importance of rigorous methods [see comments] [published erratum appears in BMJ 1996 Aug 31;313(7056):550]. *BMJ* **313**: 36–39.

Keefe RS, Bilder RM, Harvey PD, Davis SM, Palmer BW, Gold JM *et al* (2006). Baseline neurocognitive deficits in the CATIE schizophrenia trial. *Neuropsychopharmacology* **31**: 2033–2046.

Kline JA, Johnson CL, Pollack Jr CV, Diercks DB, Hollander JE, Newgard CD *et al* (2005). Pretest probability assessment derived from attribute matching. *BMC Med Inform Decis Making* **5**: 26.

Kline RB (2005). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. American Psychological Association: Washington, D.C.

Kraemer HC, Glick ID, Klein DF (2009). Clinical trials design lessons from the CATIE study. *Am J Psychiatry* **166**: 1222–1228.

Kraemer HC, Kupfer DJ (2006). Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry* **59**: 990–996.

Lauer MS (2009). Comparative effectiveness research: the view from the NHLBI. *J Am Coll Cardiol* **53**: 1084–1086.

Lavori PW (2000). Placebo control groups in randomized treatment trials: a statistician's perspective. *Biol Psychiatry* **47**: 717–723.

Lavori PW, Dawson R (2004). Dynamic treatment regimes: practical design considerations. *Clin trials (London, England)* **1**: 9–20.

Lavori PW, Rush AJ, Wisniewski SR, Alpert J, Fava M, Kupfer DJ *et al* (2001). Strengthening clinical effectiveness trials: equipoise-stratified randomization. *Biol Psychiatry* **50**: 792–801.

Lee SS, Mudaliar A (2009). Medicine. Racing forward: the Genomics and Personalized Medicine Act. *Science* **323**: 342.

Lieberman JA (2007). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia: efficacy, safety and cost outcomes of CATIE and other trials. *J Clin Psychiatry* **68**: e04.

Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO *et al* (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *N Engl J Med* **353**: 1209–1223.

March JS, Silva SG, Compton S, Shapiro M, Califf R, Krishnan R (2005). The case for practical clinical trials in psychiatry. *Am J Psychiatry* **162**: 836–846.

March JS, Swanson JM, Arnold L, Hoza B, Conners C, Hinshaw SP *et al* (2000). Anxiety as a predictor and outcome variable in the Multimodal Treatment Study of Children with ADHD (MTA). *J Abnorm Child Psychol*, 2000 **28**: 527–541.

March JS, Vitiello B (2009). Clinical messages from the Treatment for Adolescents With Depression Study (TADS). *Am J Psychiatry* **166**: 1118–1123.

Maxwell S (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol Methods* **9**: 147–163.

McCourt B, Harrington R, Fox K, Booher K, Hammond W, Walden A *et al* (2007). Data standards: at the interesection of sites,

clinical research networks and standards development initiatives. *Drug Inf J* **41**: 303–404.
Molina BS, Hinshaw SP, Swanson JM, Arnold LE, Vitiello B, Jensen PS *et al* (2009). The MTA at 8 years: prospective follow-up of children treated for combined-type ADHD in a multisite study. *J Am Acad Child Adolesc Psychiatry* **48**: 484–500.
MTA (1999). A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. Multimodal Treatment Study of Children with ADHD [in process citation]. *Arch Gen Psychiatry* **56**: 1073–1086.
Murphy SA, Collins LM, Rush AJ (2007). Customizing treatment to the patient: adaptive treatment strategies. *Drug Alcohol Depend* **88**(Suppl 2): S1–S3.
Nierenberg AA, Ostacher MJ, Calabrese JR, Ketter TA, Marangell LB, Miklowitz DJ *et al* (2006). Treatment-resisteant bipolar depression: a STEP-BD equipoise randomized effectiveness trial of antidepressant augmentation with lamotrigine, inositol, or risperidone. *Am J Psychiatry* **163**: 210–216.
Nickerson RS (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods* **5**: 241–301.
Olsen L, Aisner D, McGinnis JM (2007). *The Learning Health Care System: Workshop Summary*. National Academies Press: Washington, DC.
Owens EB, Hinshaw SP, Kraemer HC, Arnold LE, Abikoff HB, Cantwell DP *et al* (2003). Which treatment for whom for ADHD? Moderators of treatment response in the MTA. *J Consult Clin Psychol* **71**: 540–552.
Rush AJ (2007). STAR*D: what have we learned? *Am J Psychiatry* **164**: 201–204.
Sachs GS, Thase ME, Otto MW, Bauer M, Miklowitz D, Wisniewski SR *et al* (2003). Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (STEP-BD). *Biol Psychiatry* **53**: 1028–1042.
Shrout PE (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychol Sci* **8**: 1–2.
TADS (2003). Treatment for Adolescents With Depression Study (TADS): rationale, design, and methods. *J Am Acad Child Adolesc Psychiatry* **42**: 531–542.
The_MTA_Cooperative_Group (1999). Moderators and mediators of treatment response for children with attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry* **56**: 1088–1096.
Thompson B (1999). Journal editorial policies regarding statistical significance tests: heat is to fire as p is to importance. *Educ Psychol Rev* **11**: 157–169.
Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG *et al* (2009). A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol* **62**: 464–475.
Treweek S, McCormack K, Abalos E, Campbell M, Ramsay C, Zwarenstein M (2006). The trial protocol tool: the PRACTIHC software tool that supported the writing of protocols for pragmatic randomized controlled trials. *J Clin Epidemiol* **59**: 1127–1133.
Trivedi MH, Rush AJ, Gaynes BN, Stewart JW, Wisniewski SR, Warden D *et al* (2007). Maximizing the adequacy of medication treatment in controlled trials and clinical practice: STAR(*)D measurement-based care. *Neuropsychopharmacology* **32**: 2479–2489.
Trusheim MR, Berndt ER, Douglas FL (2007). Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nat Rev Drug Discov* **6**: 287–293.
Tunis S, Helms WD, McGinnis JM, Pearson SD (2009). Rountable on expanding capacity for comparative effectiveness research in the United States. *Health Res Educ Trust* **44**: 327–342.
Tunis SR, Stryer DB, Clancy CM (2003). Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *Jama* **290**: 1624–1632.
Wang PS, Ulbricht CM, Schoenbaum M (2009). Improving mental health treatments through comparative effectiveness research. *Health Aff (Millwood)* **28**: 783–791.
Wilkinson L, The_Task_Force_on_Statistical_Inference (1999). Statistical methods in psychology journals: guidelines and explanations. *Am Psychol* **54**: 594–604.
Zerhouni E (2003). Medicine. The NIH Roadmap. *Science* **302**: 63–72.
Zerhouni EA (2006). Clinical research at a crossroads: the NIH roadmap. *J Investig Med* **54**: 171–173.
Zerhouni EA (2007). Translational research: moving discovery to practice. *Clin Pharmacol Ther* **81**: 126–128.
Zwarenstein M, Treweek S (2009). What kind of randomized trials do we need? *CMAJ* **180**: 998–1000.
Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B *et al* (2008). Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ* **337**: a2390.