Journal of Clinical and Translational Science

ACTS
ASSOCIATION FOR CLINICAL AND TRANSLATIONAL SCIENCE

TRANSLATIONAL RESEARCH, DESIGN AND ANALYSIS
REVIEW ARTICLE

Maintenance efficacy designs in psychiatry: Randomized discontinuation trials – enriched but not better

S. N. Ghaemi[1*] and Harry P. Selker[2,3]

[1] Mood Disorders Program, Department of Psychiatry, Tufts Medical Center, Tufts University School of Medicine, Boston, MA, USA
[2] Tufts Clinical and Translational Scientific Institute, Tufts University, Boston, MA, USA
[3] Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA

Journal of Clinical and Translational Science (2017), 1, pp. 198–204 doi:10.1017/cts.2017.2

Introduction. Although classical randomized clinical trials (RCTs) are the gold standard for proof of drug efficacy, randomized discontinuation trials (RDTs), sometimes called "enriched" trials, are used increasingly, especially in psychiatric maintenance studies.

Methods. A narrative review of two decades of experience with RDTs.

Results. RDTs in psychiatric maintenance trials tend to use a dependent variable as a predictor: treatment response. Treatment responders are assessed for treatment response. This tautology in the logic of RDTs renders them invalid, since the predictor and the outcome are the same variable. Although RDTs can be designed to avoid this tautologous state of affairs, like using independent predictors of outcomes, such is not the case with psychiatric maintenance studies.

Further, purported benefits of RDTs regarding feasibility were found to be questionable. Specifically, RDTs do not enhance statistical power in many settings, and, because of high dropout rates, produce results of questionable validity. Any claimed benefits come with notably reduced generalizability.

Conclusions. RDTs appear to be scientifically invalid as used in psychiatric maintenance designs. Their purported feasibility benefits are not seen in actual trials for psychotropic drugs. There is warrant for changes in federal policy regarding marketing indications for maintenance efficacy using the RDT design.

Received 9 December 2016; Accepted 26 January 2017; First published online 10 July 2017

Key words: Enriched, randomized discontinuation, maintenance, clinical trials, psychotropic drugs.

Introduction

Although classical randomized clinical trials (RCTs) are the gold standard for proof of drug efficacy, the Food and Drug Administration (FDA) has increasingly allowed a different design called randomized discontinuation trials (RDTs) for indication purposes [1]. These RDTs, called "enriched," are used routinely in psychiatric maintenance trials for FDA registration. We examined the benefits and limitations of RDTs, and conclude that their extensive use in psychiatric maintenance studies may compromise scientific validity. The core critique here is with the *concept* of RDTs, that is, their internal validity, not just their external validity or misuse or misinterpretation. We call for a more consistent requirement of classical RCTs for FDA indication.

Rationale for the RDT Design

The enriched RDT design involves the following scenario: to prove long-term maintenance efficacy of a psychiatric drug, patients who enter a double-blind RCT are initially selected, before the study begins, to receive the relevant medication nonrandomly for an acute phase of the illness (e.g., an acute depressive or manic episode). Typically, the medication has already been proven effective in RCTs in the acute phases of illness; the question is whether it is effective for maintenance treatment. If a patient responds (e.g., for the acute mood episode), he/she enters the RCT that tests whether he/she will stay well by remaining on that medication, as opposed to having it stopped (whether receiving placebo or active control). Thereby, for patients who respond to a drug in an initial nonrandomized phase of acute treatment, enriched RCTs test whether they continue to respond.

From the FDA's perspective [2], enrichment is seen as having 3 potential uses: (1) "Practical" enrichment, which produces a more homogeneous

* Address for correspondence: S. N. Ghaemi, Mood Disorders Program, Department of Psychiatry, Tufts Medical Center, Tufts University School of Medicine, 800 Washington St., Boston, MA 02111, USA.
(Email: nghaemi@tuftsmedicalcenter.org)

sample, thus reducing statistical "noise." (2) "Predictive" enrichment, which produces a more treatment-responsive sample, thereby increasing effect size. (3) "Prognostic" enrichment, which produces a sample more likely to have the desired outcome (identifying high-risk subjects who are more likely to have the outcome to be measured). All 3 strategies should enhance statistical power, allowing for more efficient, ethical, and cost-effective clinical trials.

This commentary applies mainly to the second type of *predictive* enrichment, which is the most common use of this design strategy in the psychiatric setting. Our view is that, although the other 2 types of enrichment are likely valid, certain types of predictive enrichment are prone to produce scientifically questionable results, and when used as the basis of FDA indication they could pose public health risks.

Independent Versus Dependent Predictors

For FDA indication purposes, RDTs appear to be most extensively used in psychiatric illnesses, but published RDT studies are also found in oncology, neurology, and immunology, among other fields. For instance, estrogen receptor-positive tumors are more responsive to drug treatments that affect that receptor, such as tamoxifen. It thus made sense to design studies in which patients were initially preselected as estrogen receptor positive, and then randomized to receive tamoxifen or placebo [3]. Similarly, a population of patients with high renin status would be expected to be more responsive to angiotensin converting enzyme inhibitor antihypertensive agents versus placebo [2].

Such predictors are *independent* of the treatment being studied. For instance, an individual may be estrogen receptor positive or negative; this fact has nothing to do with whether that person receives tamoxifen or not in a randomized trial. In FDA simulation analyses of oncology designs, RDTs have been seen as most valid and efficient when assessing these kinds of independent predictors, viz. molecular targets that are sensitive to chemotherapeutic agents in a subset of tumors [4].

Another kind of independently predictive RDT is when the predictor is different than the outcome. A classic example is the important negative Cardiovascular Arrhythmia Suppression Trial of drugs proven to acutely suppress paroxysmal ventricular contractions [5]. A subsample of subjects who initially responded to such agents with >70% paroxysmal ventricular contraction reduction were randomized to continue the drug or switch to placebo, with mortality as the primary outcome—a different outcome than the preselected predictor. Surprisingly to some, the antiarrhythmic agents increased mortality, a negative outcome.

Such independent predictive RDTs can be either positive or negative, and are informative either way. Unfortunately, even in oncology and cardiology, and certainly in psychiatry, predictive biological markers are often unknown, and hence enrichment in RDTs based on independent predictors often is not possible.

When independent biological predictors are not known, it has been proposed that the response to treatment by the test drug can be the predictor for enrichment to (further) assess drug treatment response [6]. Thereby, a predictor that is known to be *dependent* on the outcome is used: treatment response is used to identify treatment response. The only way this approach could be nontautological would be if the specific type of treatment response was different between predictor and outcome, that is, *acute* treatment response (treatment of a *current* episode) as a predictor of *maintenance* treatment response (prevention of future *new* episodes). Posed clinically, this design answers the following question: if a patient initially responds to a medication, will that patient continue to respond to that medication in long-term treatment?

Dependent Predictor and Outcome: Are They the Same?

The use of the dependent predictor of acute treatment response to measure the outcome of maintenance treatment response raises a question of validity: are predictor and outcome the same? If so, this variety of the RDT design would be statistically invalid.

In psychiatric conditions for which RDTs are used, primarily mood illnesses, the underlying illnesses are remitting and relapsing, not chronic. Mood episodes come and go; they tend not to come and stay unchangingly or worsen unremittingly. Chronic constant depression lasting even 1 year is rare, occurring in <10% of unselected mood illness populations [7]. In some of the early RDT proposals in oncology, the context of a progressive chronic disease was taken for granted. RDTs in those studies involved cancers that always worsened, gradually and inexorably. Spontaneous remission was rare.

This difference may be a key feature to understanding why RDTs may be misused in providing FDA indications in psychiatry.

Most psychiatric studies of maintenance treatment have involved drugs that have been repeatedly proven effective in acute disease states versus placebo, and then they are tested in RDT samples of patients in whom the acute response is confirmed. These studies are invariably positive in the putative maintenance outcome. For instance, an FDA analysis found that all *maintenance* RDTs of antidepressants in major depressive disorder (MDD) over the past 25 years have shown efficacy (14/14 studies) [8], whereas only about half (38/74 studies) of classical nonenriched RCTs of antidepressants for *acute* depression in MDD showed efficacy [9].

One reason for this outcome may be that the apparent maintenance outcome is merely a reassessment of the acute-phase outcome. In other words, the same outcome is being measured twice, once before randomization (when acute nonrandomized responders are selected) and again after randomization (when those taken off the acutely effective drug relapse back into the acute phase). The failure to measure a different outcome, prevention of a new episode in the maintenance phase, is suggested by the rapid relapse of most patients in psychiatric RDTs. The natural history of the acute phase of depressive episodes is that they tend to last 3–6 months in bipolar illness and 6–12 months in unipolar depression [7]. The majority of relapses in RDTs of those conditions occur in the first 6 months after the study begins, that is, after treatment of the acute phase to initial response [10]. Typically, the nonrandomized treatment of the enriched acute phase (before the RDT of the purported maintenance phase begins) occurs for about 2 months (range up to 4 months). Thus, most relapses occur in RDTs just a few months into an acute mood episode, which is still within the natural history of the acute phase of illness. New episodes occurring 6–12 months later are infrequently observed in RDTs [10], and drug efficacy is driven mostly by early relapses in the first 6 months of follow-up, that is, not in the maintenance phase.

In other words, psychiatric RDTs are tautologous in what they measure: they preselect patients for acute response, and then they measure acute response again (but they label it maintenance response).

Hence, it may not be surprising to note that there are few, if any, studies of a drug that is effective in an acute disease state, which is then given for maintenance treatment in a RDT design, in which the result has been negative. We identified only 1 potential case in psychiatric studies (a study of electroconvulsive treatment) [11], out of at least 30 such studies in unipolar and bipolar illness as well as schizophrenia [8, 10, 12, 13]. As a valid scientific design should be falsifiable [14], the infrequency of negative studies raise questions about the RDT design's validity.

The Reverse Enriched Design

One can see how the RDT design can mistakenly suggest drug efficacy rather than spontaneous recovery if RDTs in unipolar depressive illness are examined in what might be called a "reverse" enriched design, where study populations are enriched with placebo responders. Patients treated acutely who respond to *placebo* stay on placebo, or are switched to antidepressant, for the maintenance phase. In 7 RDTs with such data available, there is *more* maintenance depressive relapse with antidepressant (42%) treatment versus placebo (25%) [15]. Should we conclude that placebo is more effective than antidepressants in the maintenance treatment for unipolar depression? If one reverses the terms, in the standard RDT analysis of antidepressant maintenance efficacy in those preselected to respond to antidepressants, antidepressants are more effective than placebo in the maintenance phase [16]. Presumably, the reverse enriched RDT design preselects for patients who will do well on placebo, perhaps because they have a more spontaneously recovery. However, this may not mean that the placebo is inherently more effective than the drug. Similarly, the standard enriched RDT design preselects for patients who will do well on a drug that is effective acutely, at least for some time in the continuation phase of treatment. However, this equally may not mean that the drug is inherently more effective than the placebo. Which enriched RDT analysis is valid, the one enriched for placebo, or the one enriched for antidepressant? One cannot have it both ways.

Rejoinders

Rejoinders to the above analysis include the following: (A) acute response will not predict maintenance response, as acute "responders" include some patients who would have gotten better anyway, that is, are not responding to the pharmacological properties of the agent: so-called "placebo responders." (B) The maintenance period for mood illnesses is shorter than 6–12 months, and in fact begins at 2 months or longer; recent RDTs are often cited to make this claim.

These rejoinders can be subject to the following criticisms:

(a) So-called "placebo responders" still only represent a proportion of response to drugs, which are proven to be more effective than placebo in acute treatment. For instance, if olanzapine produces 60% acute response for acute mania and placebo produces 40% acute response (these are typical effect sizes), then 40% of the absolute "olanzapine" response involves placebo responders. However, 20% of the absolute "olanzapine" response is actually response to that drug pharmacologically; in relative size, this is 1/3 (20%/60%) of the acute olanzapine responders. Therefore, a RDT enriched for olanzapine response will be biased by a factor of 1/3 in favor of olanzapine. This bias can be sufficient to produce a positive response in the RDT withdrawal paradigm. In other words, even allowing for "placebo responders," the RDT design will still be biased in favor of an acutely responsive drug.
(b) The claim that the maintenance phase begins in mood illnesses from 1 to 2 months onward is very difficult to justify on the basis of 100 years of natural history outcome literature in manic-depressive illness. First, it is worth reminding readers that mood illnesses are now divided into 2 groups, unipolar and bipolar; both groups used to be combined in the diagnosis of "manic-depressive illness." There is widespread consensus that the maintenance phase in unipolar depression *begins* at 12 months or longer [17]. This fact in itself would invalidate most of the antidepressant RDT literature in "maintenance" prevention of MDD, as relapses in those studies mostly occur before 6 months of follow-up and as most of those studies are not longer than 1 year in maximum duration of follow-up (only 2 studies are longer). Regarding bipolar illness, the claim that 2 months is the beginning of the maintenance phase is based on current RDTs, which are exactly the kind of studies that are being questioned. This kind of argument obviously begs the question. Further, studies in the last 4 decades are influenced by the effects of treatment; most patients in such "outcome" studies are being treated with antidepressants extensively, as well as other psychotropic agents. These drugs in themselves can alter the course of bipolar illness, for better or for worse. For worse is relevant as 2 randomized trials have shown that antidepressants can cause or worsen a rapid-cycling course in bipolar illness; this would produce more rapid relapse, and thus speed up the apparent onset of new episodes. This antidepressant effect is completely ignored in claims regarding the onset of maintenance phase in modern treatment studies. Earlier studies in the premodern psychotropic era have the advantage of basically representing untreated natural history for the long term. Many such studies were conducted from the late 19th century into the 1960s, such as the classic studies of Kraepelin's group. They never found that relapses in manic-depressive illness, including those with manic episodes who we would now diagnose with bipolar illness, happened 2–3 months after an acute episode. Rather, the mean duration until the next episode was much longer, at least in the 6–12 months range if not more. Here is a summary from the 1990 first edition of the classic textbook, *Manic-Depressive Illness*, by Goodwin and Jamison, which reviewed the pretreatment era literature, and studies from the 1970s and 1980s that were not RDTs and used less medications than now: "Cycle length is defined as the time from the onset of one episode to the onset of the next…." The authors then cite a 1979 study: "The average cycle length between the first and the second episode is 36 months, then diminishes to about 24 months, then to 12 months." Then they cite a 1985 study: "…the mean cycle lengths for the first seven episodes were 53, 28, 25, 20, 12, 15, and 9 months…." Four other studies are cited where a mean cycle length is never obtained earlier than 6 months. Kraepelin's work from 1921, in which patients definitively received no medication, produced cycle lengths of 24 months at the shortest. Hence, it is highly questionable to claim, as many mood illness clinical trialists now claim, that an entirely new episode in the maintenance phase of the natural history of bipolar illness can be observed as early as 2–3 months after remission from the acute phase.

Statistical Power

If RDTs produce a more homogeneous sample, fewer subjects would be exposed to the usual ethical risks of RCTs, especially in long-term outcomes that include the use of placebo [18]. In one of the original simulation models of the RDT design, it was estimated that samples could be utilized that were 20%–50% smaller than in standard parallel-design RCTs [19], with similar recent simulation results [20].

However, the assumption of increased power becomes dubious if the validity of the RDT design itself is questionable, as described above. The issue of power could then be posed another way: would RDTs increase power so much that they might inflate effect sizes such that a null effect size could no longer be detected? This would threaten the validity of the design in general. The only psychiatric drug in which this inflation effect can be estimated, because of the presence of both enriched and nonenriched studies, is lithium. In nonenriched maintenance studies, lithium efficacy for bipolar disorder ranged from odds ratios of 1.9 (95% confidence intervals (CI) 1.2–2.8) (newer studies, n = 638) to 3.2 (95% CIs 0.6–15.5) (older trials, n = 130). For 3 enriched trials (n = 252), a huge pooled odds ratio of 22.0 (95% CIs 7.0–68.7) was found, corresponding to 7–10-fold higher effect estimates than observed in the nonenriched designs [21]. The question would be as follows: assuming that the true effect size was the null, would an enriched design automatically produce an odds ratio of 7–10?

The Case of Lamotrigine

There is 1 apparent exception to the above critique of the validity of RDTs in the psychiatric literature. The anticonvulsant lamotrigine has been shown to be *ineffective* in multiple acute studies of depressive and manic episodes, yet it was effective in RDTs of mood episodes in bipolar disorder [22]. In our analysis, such maintenance RDT efficacy in an acutely ineffective drug is no different than simply randomizing such patients, without previous enrichment, in a classic RCT. The RDT in that case is more expensive and unnecessarily complicated because it initially excludes a large subsample of acute "nonresponders" who might have "responded" (for prevention of new episodes) later in the maintenance phase.

In sum, if there is positive response in acute use, RDTs appear mostly, if not always, positive, raising questions about their validity. If there is no acute response, then RDTs have no advantage over a traditional nonenriched RCT design.

Phase 2 Versus Phase 3 Validity

It is worthwhile distinguishing the assumptions that underlie RDTs in oncology as opposed to in psychiatry. In oncology, the view is that tumors that do not respond acutely to a chemotherapeutic agent are unlikely to respond in longer-term treatment because those subjects would "represent a subgroup with advanced disease largely resistant to further treatment" [23]. Similar considerations apply to chronic conditions such as cystic fibrosis and degenerative joint disease. In psychiatry, study designs used for mood illnesses would not necessarily pick out subgroups with "advanced disease," or unresponsive to further treatment. In fact, mood-stabilizing agents such as lithium and lamotrigine are well-known to have less benefit in acute treatment of mood episodes, despite showing marked benefit in longer-term prevention of future mood episodes [22].

Further, in oncology, RDT designs are mainly advocated for phase 2 studies, and not phase 3 trials [24]. RDTs are seen as better alternatives for finding a suggestion of treatment benefit than options such open-label studies or historical controls. RDTs are not seen as definitive demonstrations of efficacy, but a kind of advanced pilot testing for definitive classic RCTs. A similar use was made in cardiology with a small, 38-subject, placebo-controlled RDT of nifedipine for vasospastic angina [25].

In psychiatry, in contrast, with the exception of schizophrenia studies, phase 3 protocols of maintenance efficacy in the last 2 decades have consisted mostly of nothing but RDTs. Most of these studies have involved agents for depression or bipolar disorder, that is, mood illnesses. FDA indications have ensued for a number of agents (among antipsychotics: olanzapine, aripiprazole, and quetiapine for bipolar disorder; among anticonvulsants: lamotrigine for bipolar disorder; among antidepressants: venlafaxine, paroxetine, atomoxetine, fluoxetine, among others).

Practical Aspects: Feasibility, Cost, and Withdrawal Effects

Early uses of RDTs as approved by the FDA in the past few decades might have been supported on the practical basis that so few psychotropic agents were available for certain indications, such as maintenance treatment of bipolar disorder, and thus more efficient clinical trial designs such as RDTs were justified. However, now that multiple drugs are FDA approved for all phases of mood illnesses, this justification seems less relevant for those indications. Further, as noted, RDTs, if valid, would seem most defensible in phase 2 studies, rather than definitive phase 3 studies, as they are currently accepted by FDA for psychiatric indications.

Because of increased power or inflated effect sizes, it is also judged that RDTs are more cost-efficient to conduct. Especially in psychiatry, it is claimed that patients are difficult to study and large samples are both infeasible and expensive, and therefore the enriched design allows for demonstration of efficacy in a feasible and affordable manner [26]. Psychiatric patients may be more difficult to study than patients with other medical illnesses, but recent enriched maintenance designs have begun to demonstrate the ability to enroll large samples. For instance, there were 2503 subjects in 3 studies of the antipsychotic quetiapine for bipolar disorder [27–29]. If those studies had been nonenriched, even larger samples could have been studied, as overall 5852 participants were recruited for open acute treatment, but then 56% who were acute phase nonresponders were excluded from the maintenance RDT phase. These sample sizes, studied for up to 1 year, compare favorably with the size of typical RCTs of other types of medications (Table 1). The notion that the pharmaceutical industry cannot bear the cost of large studies is somewhat undermined by the fact that some psychiatric medications such as antipsychotic agents are similar to many statin agents in worldwide profitability [30]. Further, standard RCTs may save costs by not needing to recruit and then exclude large numbers of patients in the acute enrichment phase of RDTs.

A final practical limitation to RDTs, at least in the psychiatric setting, is that they commonly produce withdrawal effects and high dropout rates, with over 90% of patients dropping out at 1–2 years of follow-up (sometimes 100% of patients drop out: see Table 1 for comparison of enriched antipsychotic studies with a nonenriched classical RCT of an antihypertensive drug) [31]. Such extremely high dropout rates make maintenance study results difficult to interpret. It is notable that one of the few contemporary psychiatric nonenriched RCT maintenance studies, a 2-year comparison of lithium versus valproate in prevention of bipolar disorder (the BALANCE

Table 1. *Comparison of randomized clinical trials (RCTs) versus randomized discontinuation trials in cardiology and psychiatry: trial considerations*

Studies	Sample size	Enrichment	Mean follow-up (months)	Maximum study duration (years)	Dropouts at 2 years (%)	Number of sites	Concomitant medications allowed*
CHARM—Candersartan (2003)	7599	No	37.7	3.5	43	628	Yes
QTP as adjunct to lithium/VPA US study (2009)	628	Yes	6.2	2	96	127	No
QTP as adjunct to VPA Europe study (2008)	706	Yes	6.3	2	96	177	No
QTP monotherapy (2011)	1172	Yes	3.0	1	100	193	No

QTP, quetiapine; VPA, valproic acid.

*For treatment of the primary illness, that is hypertension or mania/depression, as opposed to peripherally relevant medications, like hypnotics for sleep.

Table 2. *Relationship between acute and maintenance response*

	Acute response	Acute nonresponse
Maintenance response	A	B
Maintenance nonresponse	C	D

trial) [32], had a low dropout rate (23%), similar to or better than antihypertensive trials.

Clinical Assumptions: Acute Versus Maintenance Response

The main clinical claim for RDT designs is that they test whether a drug that is effective for acute use should be continued or not, and for how long. It is unclear whether the RDT design answers this question any better than a classic RCT.

Clinicians sometimes seem to assume that acute treatment response necessarily implies maintenance response. However, counter examples are common in medicine: indomethacin is an acute treatment for gout but not maintenance, whereas allopurinol is effective for maintenance treatment but not effective (indeed counter-productive) for acute use; sumatriptan is an acute treatment for migraine but not maintenance; penicillin can treat acute pneumonia but is poor at maintenance prophylaxis; and steroids are much more effective for acute episodes of autoimmune illnesses than for prophylaxis. Lithium and lamotrigine are much more effective in prevention of mood episodes of bipolar illness than in treatment of acute episodes [7]. In a large National Institute of Mental Health-sponsored study, antidepressants are about twice more effective in treating acute depressive episodes than in preventing them [33].

Hence, an equivalence between acute and maintenance efficacy needs to be proven; it cannot be assumed. There appear to be 4 logical possibilities, as shown in Table 2.

In the case of acute response to a drug, RDTs have greater statistical power in 1 scenario (A), where the drug is effective in both acute and maintenance phases of treatment. In scenario B, where an acutely effective drug is in fact ineffective for maintenance treatment, RDTs that are designed based on preselecting acute drug responders are less valid than classic RCTs for the reasons given in this review (e.g., withdrawal effects, highly infrequent or absent negative studies, high dropout rates).

In the case of acute nonresponse to a drug (scenarios B and D), classical RCTs have greater statistical power than RDTs (initial treated sample that is nonresponsive is excluded in RDTs but included in RCTs, there are more dropouts in RDTs, and more withdrawal relapses in RDTs).

In sum, RCTs have more statistical power than RDTs in 2 of 4 scenarios; RDTs have more statistical power than RCTs in 1 scenario; and RCTs are of equal or greater validity than RDTs in all scenarios.

RDTs: Going Backwards on Generalizability

This analysis of the limitations of RDTs, both in validity and efficiency, relates to an increasing appreciation that "efficacy trials" on highly selected participants are less helpful, because of their lack of generalizability and applicability to usual care, than "effectiveness trials." In addition to aiming to make trials closer to the effectiveness model, 1 possible approach is integrated efficacy-to-effectiveness clinical trials [34], whereby initially more restrictive RCT efficacy designs are later broadened, as part of the same study, to assess outcomes in more generalizable samples.

Whatever approach is taken, there is a consensus in the clinical trial community that traditional randomized efficacy trials do not provide enough clinically useful information for usual medical practice. RDTs are a step in the opposite direction, with even less generalizable results (and also diminished validity) than usual efficacy trials. Some investigators have even proposed extending the RDT approach from maintenance studies to acute trials in psychiatry [18]. This approach, which has received support from FDA, would appear to head the psychiatric field in the wrong direction, both in terms of validity and generalizability.

Hence, a reconsideration of RDTs should be undertaken in the context of this general trend, apparently in most of medicine other than psychiatry, toward recognizing a need for less, rather than more, preselection and manipulation of samples in randomized drug trials.

FDA Policy

These considerations have important implications for FDA policy, which has, in recent years, encouraged the pharmaceutical industry and academic researchers to consider RDT designs in clinical trials [1]. One implication is that such RDT designs should be limited to independent predictors and not the dependent predictor of acute treatment efficacy with the drug being studied for maintenance efficacy. Another implication is that RDTs be limited to phase 2, but not phase 3, studies, and thus should not serve as the sole basis for FDA indication of maintenance treatments.

In relation to psychiatric research, this analysis is relevant to the conclusions of a FDA Psychopharmacology Drugs Advisory Committee meeting held on October 25, 2005 (http://www.fda.gov/ohrms/dockets/ac/05/transcripts/2005-4186T1.pdf). In that meeting, the FDA committee unanimously rejected the question raised by FDA staff as to whether long-term efficacy should be required at the time of acute efficacy indications in psychiatric conditions. FDA staff also asked the Committee to comment on whether the RDT design should be limited to trials that require at least 6 months of stabilization after acute response to a drug, before the randomized maintenance RDT begins. This also was rejected at that meeting, and the current standard FDA requirement is only a minimum 2-month stabilization period before initiation of a RDT maintenance study.

What Is an Alternative to RDTs for Maintenance Efficacy in Psychiatric Studies?

On the basis of the analysis presented in this article, our recommendations for clinical research on maintenance efficacy in psychopharmacology are 2-fold:

(1) One option would be that acute response clinical trials and maintenance prevention clinical trials should be separated clearly from each other. Acute response clinical trials can and should be conducted according to usual standards: parallel randomization would occur for a short period, usually around 6–8 weeks. Entirely different studies would then assess maintenance prevention in a different sample of subjects; those subjects would be entered into maintenance studies after having improved from acute-phase episodes with a range of different medications, *not only* the specific medication being studied in the maintenance prevention clinical trial.
(2) A second option would be to conduct an acute-phase response study with the usual randomized parallel design for a short time, such as 6–8 weeks, and then to simply continue those same subjects into a maintenance prevention phase of 12 months or longer, *without*

changing the initial acute randomization assignment. Keeping the initial acute randomization assignment would ensure that maintenance treatment was not biased in favor of acute drug responders. This is because those who responded in the acute phase to the drug would still be on the drug in the maintenance phase, without any influence of withdrawal or discontinuation of the drug. Further, those who responded in the acute phase to placebo would still be on placebo in the maintenance phase, without any influence of discontinuation of placebo. Both groups would balance each other out in the bias of "enrichment" in both arms.

In our view, RDTs are unnecessary; they do not add any scientific information that cannot be obtained more validly, and in many cases even more efficiently, with standard RCTs. However, if any role for RDTs is to be preserved, we would recommend RDTs be limited in psychiatry, as in all other medical disciplines, to phase 2 trials for pilot testing purposes, and not for phase 3 marketing indications. In the latter cases, the above 2 options of standard RCTs would be recommended.

Conclusions

This summary of 2 decades of experience with RDTs suggests appropriate uses and potential avenues for changes in neuropharmacology FDA policy. RDTs appear most valid when used with (a) an independent predictor of treatment response (such as a known biological marker) and/or (b) a chronic, progressive disease, without frequent episodic recovery. If the first criterion of an independent marker is absent, RDTs should be used primarily as phase 2 pilot studies in preparation for larger classical RCTs for definitive determination of drug efficacy. The current state of antipsychotic, anticonvulsant, and antidepressant maintenance research in psychiatric indications—primarily mood illnesses—does not meet the proposed criteria. Changes in FDA policy would help better establish true drug efficacy in these psychiatric settings, ending its sole reliance on RDTs and requiring a return to some usage of traditional nonenriched RCT designs with larger samples and longer durations of treatment than are currently required.

Such classical RCTs can be conducted in ways that are practically feasible, more efficient, and have as much, if not more, statistical power compared with RDTs. Most importantly, if our analysis is correct, such RCTs are more scientifically valid than RDTs as currently conducted in psychiatric maintenance studies. Clinical practice for long-term psychiatric treatment, as currently based on potentially invalid RDT designs, would then be based on more valid scientific assessments of long-term efficacy, and thus expose the large population of patients who receive these agents to risks and side-effects only where these medications are proven effective at a level of scientific validity that is consistent with the rest of medical practice.

Acknowledgments

The authors thank Jessica Paulus, Ph.D., Paul Vohringer, M.D., M.P.H., and David Kent, M.D., M.S., for their input on earlier drafts of this article.

Financial Support

H.P.S.'s participation was supported by National Institutes of Health National Center for Advancing Translational Sciences Clinical and Translational Science Award number UL1 TR000073.

Disclosures

In the past year, S.N.G. has received honoraria from Sunovion Pharmaceuticals and has done consulting work for Advance Medical.

References

1. **Food and Drug Administration.** *Guidance for Industry: Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products* December 2012. US Department of Health and Human Services, Washington DC.
2. **Temple R.** Enrichment strategies. *FDA/DIA Statistics Forum* 2012.
3. **Fisher B, *et al.*** A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. *New England Journal of Medicine* 1989; **320**: 479–484.
4. **Freidlin B, Simon R.** Evaluation of randomized discontinuation design. *Journal of Clinical Oncology* 2005; **23**: 5094–5098.
5. **Ruskin JN.** The cardiac arrhythmia suppression trial (CAST). *New England Journal of Medicine* 1989; **321**: 386–388.
6. **Rosner GL, Stadler W, Ratain MJ.** Randomized discontinuation design: application to cytostatic antineoplastic agents. *Journal of Clinical Oncology* 2002; **20**: 4478–4484.
7. **Goodwin F, Jamison K.** *Manic-Depressive Illness*, 2nd edition. New York: Oxford University Press, 2007.
8. **Borges S, *et al.*** Review of maintenance trials for major depressive disorder: a 25-year perspective from the US Food and Drug Administration. *Journal of Clinical Psychiatry* 2014; **75**: 205–214.
9. **Turner EH, *et al.*** Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine* 2008; **358**: 252–260.
10. **Goodwin FK, Whitham EA, Ghaemi SN.** Maintenance treatment study designs in bipolar disorder: do they demonstrate that atypical neuroleptics (antipsychotics) are mood stabilizers? *CNS Drugs* 2011; **25**: 819–827.
11. **Kellner CH, *et al.*** Continuation electroconvulsive therapy vs pharmacotherapy for relapse prevention in major depression: a multisite study from the Consortium for Research in Electroconvulsive Therapy (CORE). *Archives of General Psychiatry* 2006; **63**: 1337–1344.
12. **Briscoe B, El-Mallakh R.** The evidence base for the long-term use of antidepressants as prophylaxis against future depressive episodes (poster). *American Psychiatric Association Annual Meeting*, May 23–26, New Orleans, LA, 2010.
13. **Leucht S, *et al.*** Maintenance treatment with antipsychotic drugs for schizophrenia. *The Cochrane Database Systematic Reviews* 2012; **5**: CD008016.
14. **Popper K.** *The Logic of Scientific Discovery*. New York: Basic Books, 1959.
15. **Andrews PW, *et al.*** Blue again: perturbational effects of antidepressants suggest monoaminergic homeostasis in major depression. *Frontiers in Psychology* 2011; **2**: 1–24.
16. **Geddes JR, *et al.*** Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. *The Lancet* 2003; **361**: 653–661.
17. **Frank E, *et al.*** Conceptualization and rationale for consensus definitions of terms in major depressive disorder: remission, relapse, and recurrence. *Arch Gen Psychiatry* 1991; **48**: 851–855.
18. **Fava M, *et al.*** The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics* 2003; **72**: 115–127.
19. **Kopec JA, Abrahamowicz M, Esdaile JM.** Randomized discontinuation trials: utility and efficiency. *Journal of Clinical Epidemiology* 1993; **46**: 959–971.
20. **Karrison TG, *et al.*** Estimation of progression-free survival for all treated patients in the randomized discontinuation trial design. *American Statistics* 2012; **66**: 155–162.
21. **Deshauer D, *et al.*** Re-evaluation of randomized control trials of lithium monotherapy: a cohort effect. *Bipolar Disorders* 2005; **7**: 382–387.
22. **Ghaemi SN.** The failure to know what isn't known: negative publication bias with lamotrigine and a glimpse inside peer review. *Evidence-Based Mental Health* 2009; **12**: 65–68.

23. **Capra WB.** Comparing the power of the discontinuation design to that of the classic randomized design on time-to-event endpoints. *Controlled Clinical Trials* 2004; **25**: 168–177.
24. **Sharma MR, Stadler WM, Ratain MJ.** Randomized phase II trials: a long-term investment with promising returns. *Journal of National Cancer Institute* 2011; **103**: 1093–1100.
25. **Schick EC Jr, *et al.*** Randomized withdrawal from nifedipine: placebo-controlled study in patients with coronary artery spasm. *American Heart Journal* 1982; **104**: 690–697.
26. **Bowden CL, *et al.*** Maintenance clinical trials in bipolar disorder: design implications of the divalproex-lithium-placebo study. *Psychopharmacology Bulletin* 1997; **33**: 693–699.
27. **Vieta E, *et al.*** Efficacy and safety of quetiapine in combination with lithium or divalproex for maintenance of patients with bipolar I disorder (international trial 126). *Journal of Affective Disorders* 2008; **109**: 251–263.
28. **Suppes T, *et al.*** Maintenance treatment for patients with bipolar I disorder: results from a north American study of quetiapine in combination with lithium or divalproex (trial 127). *The American Journal of Psychiatry* 2009; **166**: 476–488.
29. **Weisler R, *et al.*** Quetiapine or lithium versus placebo for maintenance treatment of bipolar I disorder after stabilization on quetiapine. *60th Institute on Psychiatric Services Congress*, Chicago, IL, 2008.
30. **IMS Institute.** The use of medicines in the United States: review of 2011. [Cited Apr 13, 2017]. (http://www.imshealth.com/imshealth/Global/Content/IMS%20Institute/Documents/IHII_UseOfMed_report%20.pdf).
31. **Goodwin FK, Whitham EA, Ghaemi SN.** Maintenance treatment study designs in bipolar disorder: do they demonstrate that atypical neuroleptics (antipsychotics) are mood stabilizers? *CNS Drugs* 2011; **25**: 819–827.
32. **Geddes JR, *et al.*** Lithium plus valproate combination therapy versus monotherapy for relapse prevention in bipolar I disorder (BALANCE): a randomised open-label trial. *The Lancet* 2010; **375**: 385–395.
33. **Rush AJ, *et al.*** Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *The American Journal of Psychiatry* 2006; **163**: 1905–1917.
34. **Selker HP, *et al.*** A proposal for integrated efficacy-to-effectiveness (E2E) clinical trials. *Clinical Pharmacology and Therapeutics* 2013; **95**: 147–153.