COMMENTARY

Two Clinical Trial Designs to Examine Personalized Treatments for Psychiatric Disorders

Andrew C. Leon, PhD

The National Institute of Mental Health Strategic Plan calls for the development of personalized treatment strategies for mental disorders. In an effort to achieve that goal, several investigators have conducted exploratory analyses of randomized controlled clinical trial (RCT) data to examine the association between baseline subject characteristics, the putative moderators, and the magnitude of treatment effect sizes. Exploratory analyses are used to generate hypotheses, not to confirm them. For that reason, independent replication is needed. Here, 2 general approaches to designing confirmatory RCTs are described that build on the results of exploratory analyses. These approaches address distinct questions. For example, a 2 × 2 factorial design provides an empirical test of the question, "Is there a greater treatment effect for those with the single-nucleotide polymorphism than for those without that polymorphism?" and the hypothesis test involves a moderator-by-treatment interaction. In contrast, a main effects strategy evaluates the intervention in subgroups and involves separate hypothesis-testing studies of treatment for subjects with the genotypes hypothesized to have enhanced and adverse response. These designs require widely disparate sample sizes to detect a given effect size. The former could need as many as 4-fold the number of subjects. As such, the choice of design impacts the research costs, clinical trial duration, and number of subjects exposed to risk of an experiment, as well as the generalizability of results. When resources are abundant, the 2 × 2 design is the preferable approach for identifying personalized interventions because it directly tests the differential treatment effect, but its demand on research funds is extraordinary.

J Clin Psychiatry 2011;72(5):593–597

Submitted: *August 3, 2009; accepted October 21, 2009.*
Online ahead of print: *July 13, 2010 (doi:10.4088/JCP.09com05581whi).*
Corresponding author: *Andrew C. Leon, PhD, Weill Cornell Medical College, Department of Psychiatry, Box 140, 525 East 68th St, New York, NY 10065 (acleon@med.cornell.edu).*

Over the past 5 decades, there has been considerable progress in the development of therapeutic interventions for mood and anxiety disorders. Nevertheless, given that response rates in randomized controlled clinical trials (RCTs) are typically less than 50% for these interventions, and remission rates much lower, many patients must undergo trials of multiple treatments before symptom relief is achieved. Thus, there is a need to replace clinical serendipity with a more systematic algorithmic approach to identifying the medication and/or the psychotherapy most likely to reduce the suffering of a particular patient. For this reason, the National Institute of Mental Health (NIMH) has called for an effort to develop more targeted, personalized treatments, that is, treatments that have been shown to benefit patients with particular profiles defined by demographic, clinical, or genetic characteristics.[1]

To achieve this critically important goal, investigators must identify moderators of treatment, the baseline subject characteristics that are positively or negatively associated with the treatment versus control effect size. For example, if active medication and placebo did not separate in an RCT for males, but there was a clear advantage of active medication for females, gender would be a moderator of treatment. The choice of hypothesized moderators to study could be based on preliminary evidence from chart reviews, pilot studies, or clinical experience. Alternatively, it could be derived from systematic empirical evidence from the literature. The Macarthur Group has presented a coherent approach to analyze hypothesized moderators of treatment.[2] When a moderator is selected on the basis of preliminary evidence, they advocate the use of exploratory analyses that focus on magnitude of effect of a moderator and not on the use of significance testing and *P* values.[2] The magnitude of effect can be quantified in several ways[3]; 4 examples are provided here. A standardized group difference (Cohen *d*) represents a between-group difference expressed in standard deviation units on the outcome measure. The number needed to treat represents the number of subjects that must treated with active to have 1 more responder than if the same number of subjects was treated with the control. Number needed to harm parallels number needed to treat, but, as the name suggests, quantifies adverse effects. Finally, if 1 active subject and 1 control subject were sampled from an RCT dataset, the area under the curve represents the probability that the active subject had a superior response.

Exploratory analyses typically involve multiple testing, and, therefore, in lieu of some independent confirmation or validation, the results of exploratory analyses should be viewed as tentative and not be used for treatment decision making. If there is a reasonable theoretical or empirical basis for the finding, the results can be applied to guide the design of subsequent RCTs. For instance, they can be used to refine inclusion and exclusion criteria. Alternatively, a researcher could attempt to replicate exploratory findings with secondary analyses of archival RCT data. If either the prospective or archival RCT data lend confirmation to the evidence, they could serve as a basis for clinical decisions. In contrast, when a moderator is selected on the basis of published evidence, the design could call for confirmatory analyses (not exploratory) with the necessary multiplicity adjustments and multiplicity-adjusted sample size estimates.[4]

Consider 2 examples of exploratory analyses. Panic Focused Psychodynamic Psychotherapy (PFPP) was compared

to Applied Relaxation Therapy (ART) and shown to be efficacious for panic disorder.[5] The presence of a cluster C diagnosis was identified as a putative moderator of treatment in exploratory analyses. PFPP was superior to ART for those without a cluster C diagnosis (n = 30), with a between–treatment group effect size of $d = 0.69$ (95% CI, −0.06 to 1.44).[6] This quantity (d) represents a between-group difference of 0.69 standard deviation units on the Panic Disorder Severity Scale.[7] In contrast, the advantage of PFPP was substantially greater for those with a cluster C diagnosis (n = 19): $d = 1.35$ (95% CI, 0.38 to 2.32). Nonetheless, stratum-specific sample sizes are quite small, and, as a result, estimates of the treatment effect lack precision, as seen in wide confidence intervals,[8] underscoring the recommendation that tentative results not be used to inform clinical decisions.

Another example of analyses that identified an intuitively appealing strategy for personalized treatments was recently presented for chronic depression.[9] The RCT compared nefazodone, cognitive behavioral analysis system of psychotherapy, and their combination. Subjects were asked to specify their treatment preference prior to randomization (medication, psychotherapy, combination, or no preference). Exploratory analyses showed that treatment preference appeared to moderate the effect of treatment, especially for subjects who stated a preference for either monotherapy. Among those who preferred psychotherapy, subjects who received psychotherapy had a higher remission rate (50.0%) than those who received medication (7.7%). In contrast, among those who preferred medication, participants randomly assigned to medication had a higher remission rate (45.5%) than those randomly assigned to psychotherapy (22.2%).

The results of these 2 post hoc exploratory analyses have been applied in the design of subsequent RCTs. The objective of this commentary is to contrast 2 of the possible designs for RCTs that provide confirmatory evaluation of exploratory moderator results. These designs address different research questions, test different hypotheses, and have widely disparate sample size requirements. Implications of the choice of design impact the generalizability of results, research costs, clinical trial duration, and number of subjects exposed to risk of an experiment. For simplicity, the following discussion focuses on a study with a normally distributed outcome (eg, a severity rating), a binary moderator (eg, present or absent), and equal sample sizes across treatment groups and at each level of the moderator.

HYPOTHESIS-TESTING APPROACHES TO EXAMINING MODERATORS IN RCTs

Biomarkers hold great promise for identifying personalized treatments, and, therefore, for illustration, a genotype is used as the hypothesized moderator throughout the remaining presentation. For example, 2 genetic polymorphisms that have been implicated as potential moderators of antidepressant treatment are the serotonin transporter–linked polymorphic region (5HTTLPR)[10–13] and brain-derived neurotrophic factor.[14] We will assume that exploratory analyses have identified a genotype as a moderator of treatment. Two approaches to designing the subsequent study will be considered: (1) an interaction strategy and (2) a main effects strategy. The choice between these is driven by both the hypothesis and the research resources. If financial, temporal, and human resources are plentiful, the interaction strategy is preferable for identifying personalized treatments because it provides a direct comparison of the treatment effect across subgroups.

Interaction Design

Consider first an RCT that is prospectively designed to examine a genotype by treatment interaction with hypothesis testing (as opposed to the post hoc exploratory analyses described above). The research question asks, "Is there a greater treatment effect for those with the single-nucleotide polymorphism (SNP+) than for those without that polymorphism (SNP–)?" A 2×2 factorial design is used in which subjects are randomly assigned to either the investigational intervention or the control. In an effort to assure balanced treatment assignment in each stratum, randomization is stratified by the moderator, such that separate randomization lists are used for the SNP+ subjects and SNP– subjects. (Oversampling of the less prevalent genotype could be necessary.) The null hypothesis for the interaction involves the relation among population means (μ) on severity ratings across the 4 cells of a 2×2 factorial design:

$$H_0: \mu_{\text{Active/SNP+}} - \mu_{\text{Control/SNP+}} = \mu_{\text{Active/SNP–}} - \mu_{\text{Control/SNP–}}$$

and the alternative hypothesis:

$$H_A: \mu_{\text{Active/SNP+}} - \mu_{\text{Control/SNP+}} \neq \mu_{\text{Active/SNP–}} - \mu_{\text{Control/SNP–}}$$

A significance test examines the RCT sample data for a treatment-by-SNP interaction, which compares the treatment group differences across the genotypes and could be used to guide the choice of personalized treatments. The statistical model might involve an analysis of variance (ANOVA) for cross-sectional data or a mixed-effects linear regression model for longitudinal, repeated-measures data.

Required Sample Sizes

The choice of study design is guided primarily by the scientific question; nevertheless, budgetary constraints, corresponding limitations on sample size, and ethical concerns are also factors. In fact, the sample size needed to provide sufficient statistical power to detect a clinically meaningful effect of an interaction can be as much as 4-fold that needed to detect a main effect of the same magnitude. This has been shown for ANOVA models[15] and mixed-effects linear models that examine repeated measures within subject over time.[16,17] For example, approximately 100 subjects per group would be needed to detect a main effect of $d = 40$ with 80% statistical power, whereas 400 subjects per group would be needed for an interaction of the same magnitude (ie, 0.40

SD units) in an ANOVA. [As a technical clarification: the interaction effect sizes are presented here in standard deviation units of the outcome (ie, the pooled standard deviation of the severity rating across treatment groups), not in standard error units of the interaction parameter estimate.] Indeed, if the interaction effect is expected to be smaller than that of the main effect, the sample size disparity is greater; conversely, a larger interaction effect would reduce the disparity.

Main Effects Design

Once again, it is assumed that exploratory analyses have identified a genotype as a putative moderator of treatment. If the results suggest that one level of the moderator is associated with enhanced treatment response, a subsequent RCT can be designed to examine the treatment effect with hypothesis testing. In contrast, if the moderator is associated with decreased response, those findings can be used to inform the selection or development of a novel intervention that can be tested in a subsequent RCT. In the former case, the research question could ask, "Is there a treatment effect for SNP+?" That is, in contrast to the interaction model that compares the treatment effect across subgroups, this study will focus exclusively on an enriched sample, the subgroup of patients who, based on exploratory analyses, appeared to benefit especially well from the investigational intervention. This design, also referred to as a "targeted" or "restricted entry" design,[18,19] would be particularly appealing during a time of constrained resources. In such a case, 1 inclusion criterion requires that the subjects have the characteristic that appeared to be associated with enhanced response. Thus, only SNP+ patients would be recruited and randomly assigned to either the investigational intervention or the control. The null hypothesis is

$$H_0: \mu_{\text{Active/SNP+}} = \mu_{\text{Control/SNP+}}$$

and the alternative hypothesis is

$$H_1: \mu_{\text{Active/SNP+}} \neq \mu_{\text{Control/SNP+}}$$

The significance test examines the main effect of treatment, comparing those randomly assigned to the investigational intervention with those randomly assigned to the control. The generalizability of those results is limited to the SNP+ population. The test does not compare treatment effects across subgroups (eg, SNP+ vs SNP–), but its results could guide the choice of personalized treatments for the genotype included. Although the sample size and cost are reduced substantially, the impact on recruitment of participants and study duration could prove to be impractical if this strategy focuses on a rare group of patients, perhaps an uncommon genotype, yet this limitation would apply to both main effects and interaction designs.

Test a Novel Intervention for Those Expected to Have Decreased Response

On the other hand, if the exploratory analyses suggested that a subgroup of patients (eg, SNP–) is unlikely to benefit from the investigational intervention, an alternative intervention could be selected, or a novel intervention developed, for preliminary evaluation in an RCT. The RCT would be designed to recruit only those with putative characteristics of decreased response (as identified in exploratory analyses). The research question is, "Is there a treatment effect for SNP–?" The null hypothesis is

$$H_0: \mu_{\text{Active/SNP–}} = \mu_{\text{Control/SNP–}}$$

and the alternative hypothesis is

$$H_1: \mu_{\text{Active/SNP–}} \neq \mu_{\text{Control/SNP–}}$$

Again, the significance test involves the main effect of treatment. The strategy does not compare treatment effects across subgroups, the generalizability of the results is limited to SNP– patients, and, being a preliminary evaluation, the findings would require replication in a confirmatory trial.

SUMMARY

The NIMH has initiated an effort to develop personalized treatment strategies for mental disorders. In keeping with that goal, numerous investigators have conducted exploratory analyses of RCT data focusing on baseline subject characteristics, the hypothesized moderators. These analyses examine the association between the moderators and the magnitude of the treatment effect sizes. Exploratory results generate, but do not confirm, hypotheses. Therefore, independent replication is needed from existing literature, secondary analyses of archival RCT data, or a new RCT.

Here, 2 general approaches to designing subsequent studies have been described that build on the results of exploratory analyses. These approaches address distinct questions. A 2×2 factorial design provides an empirical test of the question, "Is there a differential treatment effect for those with various levels of the moderator?" for which the hypothesis test involves a moderator-by-treatment interaction. If resources permit, this approach provides a direct test of the moderating effect. In contrast, the main effects strategy is a targeted design that involves separate hypothesis-testing studies of treatment for those hypothesized to have enhanced and adverse response (based on the exploratory results), perhaps initially testing the investigational intervention with the former group and then a novel intervention with the latter. This approach provides a less costly alternative, reducing the study duration, costs, and number of subjects that must be recruited and exposed to the risk of an experiment. The savings afford an opportunity to attempt to replicate the findings in a subsequent clinical trial and provide definitive evidence to guide clinical decisions.

Clearly, the main effects approach does not compare the treatment effect across complementary subgroups, but instead focuses on what appeared to be an enriched subgroup identified in exploratory analyses. During periods of copious resources, the interaction design is the superior approach because it can examine differential treatment response across

subgroups. However, when funding is limited, it might be more reasonable to focus on evaluating efficacy in what appears to be the more promising target population, particularly when the alternative is no funding to pursue the line of research. This is a very real possibility given the requirement that NIMH program prereview and authorize submission of any grant that has an annual budget exceeding $500,000. There is a risk with this strategy. If the exploratory result was, in fact, a false-positive result, the main effects design would focus on a subgroup that responds no differently than those excluded. Moreover, the evaluation of the investigational intervention would have been curtailed in a subgroup (based on exploratory analyses of a relatively small amount of data) that might have benefitted from the intervention had they been given access.

A main effects design could compare the investigational intervention with either a placebo or an active comparator. If an active comparator is chosen, and the necessary increase in sample size is implemented such that the smaller difference between active agents could be detected, a test of comparative effectiveness could provide pragmatic results to guide the clinician choosing between 2 interventions. Alternatively, an RCT could be designed in a way that subjects are randomly assigned to receive what is deemed either a matched or an unmatched intervention, based on the direction of the hypothesized effect associated with the putative moderator.

It is possible that a clinical trial will examine the effect of multiple moderators. Whether these are tested individually or simultaneously will impact the design. If 1 hypothesis is proposed for each of several biomarkers, multiplicity adjustments must be used to control for the risk of false-positive results seen with multiple tests, and multiplicity-adjusted sample sizes[4] must be identified to control false-negative results associated with a lower α threshold. In contrast, if the study seeks a biosignature for personalized interventions, 1 hypothesis could very well implicate combinations of moderators. One approach to testing the combinations would involve testing higher-order interactions, further increasing the required sample size.

There are several caveats regarding the approaches that have been discussed. First, it is conceivable that a main effects design could forgo the exploratory analyses, if it is driven by the theoretical basis of the mechanism of action for a novel treatment. Second, in an effort to articulate principles, this discussion has focused on normally distributed outcomes and a binary moderator with balanced sample sizes. If an alternative design is used, particularly with a binary or survival outcome, the same general concepts hold, but the specific details vary. For example, with an unequal number of subjects with each level of the moderator, the inflated sample size for the interaction design becomes even greater. Most importantly, the sample size required to detect an interaction remains vastly greater than that required for the main effect. Third, the efficiency of recruiting and screening for either approach is dependent on the prevalence of the putative moderating characteristic and its ease of ascertainment, each of which have bearing on study costs and enrollment duration. For example, if the classification of subjects requires a genetic test result, there may be considerable delay in treatment for a large number of people who ultimately will not be eligible for participation. Although this may seem to be especially problematic for the main effects approach, the rare group would also be needed in a study involving the interaction. Clearly, the main effects trial tests the treatment effect in a rarefied group, and therefore the approach limits the generalizability of results, and, if regulatory approval is involved, the indication on the product label would be restricted. Yet, with constrained resources, perhaps the focus should be on the subgroup hypothesized to have an enhanced response on the basis of exploratory results. Ultimately, however, the choice between the 2 designs should be guided primarily by the research question, but budgetary constraints cannot be ignored.

Author affiliation: Department of Psychiatry, Weill Cornell Medical College, New York, New York.
Potential conflict of interest: Dr Leon has served on data and safety monitoring boards for AstraZeneca, Dainippon Sumitomo, and Pfizer; been a consultant/advisor to Cyberonics, US Food and Drug Administration, MedAvante, National Institute of Mental Health, Schering Plough, and Takeda; and has equity in MedAvante.
Funding/support: This research was supported, in part, by grants from the National Institute of Mental Health (MH060447 and MH068638).
Previous presentation: Presented, in part, at the annual meetings of the National Institute of Mental Health, New Clinical Drug Evaluation Unit (NCDEU); May 27–30, 2008; Phoenix, Arizona; and the International Society for CNS Clinical Trials and Methodology (ISCTM); March 2009; Alexandria, Virginia.
Acknowledgment: The author gratefully acknowledges Lori L. Davis, MD, and anonymous reviewers for valuable critiques of an earlier draft of this manuscript. Moonseong Heo, PhD, and Hakan Demirtas, PhD, provided comments on technical aspects of the presentation. None of the acknowledged individuals report potential conflicts of interest related to the commentary.

REFERENCES

1. National Institute of Mental Health Strategic Plan. http://www.nimh.nih.gov/about/strategic-planning-reports/index.shtml. Accessed August 3, 2009.
2. Kraemer HC, Wilson GT, Fairburn CG, et al. Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry*. 2002;59(10):877–883.
3. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry*. 2006;59(11):990–996.
4. Leon AC. Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. *J Clin Psychiatry*. 2004;65(11):1511–1514.
5. Milrod B, Leon AC, Busch FN, et al. A randomized controlled clinical trial of psychoanalytic psychotherapy for panic disorder. *Am J Psychiatry*. 2007b;164(2):265–272.
6. Milrod BL, Leon AC, Barber JP, et al. Do comorbid personality disorders moderate panic-focused psychotherapy? an exploratory examination of the American Psychiatric Association practice guideline. *J Clin Psychiatry*. 2007a;68(6):885–891.
7. Shear MK, Brown TA, Barlow DH, et al. Multicenter collaborative Panic Disorder Severity Scale. *Am J Psychiatry*. 1997;154(11):1571–1575.
8. Kraemer HC, Mintz J, Noda A, et al. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry*. 2006;63(5):484–489.
9. Kocsis JH, Leon AC, Markowitz JC, et al. Patient preference as a moderator of outcome for chronic forms of major depressive disorder treated with nefazodone, cognitive behavioral analysis system of psychotherapy, or their combination. *J Clin Psychiatry*. 2009;70(3):354–361.

10. Yu YW-Y, Tsai S-J, Chen T-J, et al. Association study of the serotonin transporter promoter polymorphism and symptomatology and antidepressant response in major depressive disorders. *Mol Psychiatry*. 2002;7(10):1115–1119.
11. McMahon FJ, Buervenich S, Charney D, et al. Variation in the gene encoding the serotonin 2A receptor is associated with outcome of antidepressant treatment. *Am J Hum Genet*. 2006;78(5):804–814.
12. Murphy DL, Lerner A, Rudnick G, et al. Serotonin transporter: gene, genetic disorders, and pharmacogenetics. *Mol Interv*. 2004;4(2):109–123.
13. Hariri AR, Holmes A. Genetics of emotional regulation: the role of the serotonin transporter in neural function. *Trends Cogn Sci*. 2006;10(4):182–191.
14. Bath KG, Lee FS. Variant BDNF (Val66Met) impact on brain structure and function. *Cogn Affect Behav Neurosci*. 2006;6(1):79–85.
15. Fleiss JL. *The Design & Analysis of Clinical Experiments*. New York, NY: John Wiley and Sons; 1986.
16. Leon AC, Heo M. Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Comput Stat Data Anal*. 2009;53(3):603–608.
17. Heo M, Leon AC. Sample sizes required to detect two-way and three-way interactions involving slope differences in mixed-effects linear models. *J Biopharm Stat*. 2010;20(4):787–802.
18. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res*. 2004;10(20):6759–6763.
19. Schork NJ, Topol EJ. Genotype-based risk and pharmacogenetic sampling in clinical trials. *J Biopharm Stat*. 2010;20(2):315–333.