Evaluation of psychiatric interventions in an observational study: issues in design and analysis

Andrew C. Leon, PhD

Characteristics of randomized controlled clinical trials (RCTs) and observational studies of psychiatric intervention effectiveness are contrasted. Randomization drives treatment assignment in an RCT, whereas clinician and patient selection determine treatment in an observational study. Strengths and weaknesses of randomized and observational designs are considered. The propensity adjustment, a statistical approach that allows for intervention evaluation in a nonrandomized observational study, is described here. The plausibility of propensity adjustment assumptions must be carefully evaluated. This data analytic strategy is illustrated with the longitudinal observational data from the National Institute of Mental Health Collaborative Depression Study. Evaluations presented here examine acute and maintenance antidepressant effectiveness and demonstrate effectiveness of the higher categorical doses.

 Dialogues Clin Neurosci. 2011;13:191-198.

Keywords: *observational study; propensity score; randomized controlled clinical trial; treatment effectiveness*

Background

There are two general approaches to research design for scientific study of intervention effectiveness: the randomized controlled clinical trial (RCT) and the observational study. (The term naturalistic study is not used here, in an effort to distinguish the observational study from a study of the natural history of an untreated illness, which is often called a naturalistic study.) Typically the RCT is preferred for a variety of reasons, most importantly randomized treatment assignment. However, there are some advantages of the observational study if it is appropriately analyzed. Here the distinction between the two designs, and advantages and disadvantages of each for various settings, are considered.

Randomized clinical trials

The primary goal in designing an RCT is to minimize the bias in the estimate of the treatment effect.[1,2] Three fundamental features of RCT design play a pivotal role in minimizing bias: randomized treatment assignment, double-blinded assessments, and a credible comparison group. There are other essential goals in clinical trial design as well. For instance, a well-designed trial will

Author affiliations: Weill Cornell Medical College, New York, New York, USA

Address for correspondence: Andrew C. Leon, Weill Cornell Medical College, Department of Psychiatry, Box 140, 525 East 68th Street, New York, NY 10065, USA (e-mail: acleon@med.cornell.edu)

have a sample size that provides adequate statistical power to detect a clinically meaningful effect. An antidepressant trial, for example, must be designed to detect an effect of about .40 standard deviation units (ie, Cohen's *d*=.40) and that requires about 100 participants per treatment group for 80% statistical power for a *t*-test with a two-tailed alpha-level of .05. However, there is a tradeoff between power and feasibility. Recruitment of 100 per group is only feasible if compatible with both the budget and the study site patient flow.

Furthermore, a trial must be designed such that the false-positive rate (ie, Type I error) is acceptable; the convention is 0.05. This is because false-positive treatments, of course, do not reduce suffering in patients. Finally, the design must be applicable. That is, the recruited sample should yield results that generalize to the target patient population; ie, that for which the indication is sought.

A well-designed and well-implemented RCT is the gold standard for treatment evaluation. This is because the groups tend to be well-balanced at baseline, and therefore subsequent group differences can be attributed to treatment effects, providing strong internal validity. However, there are limitations of results from RCTs for psychiatric disorders. For example, trials for mood disorder interventions typically involve short-term treatment (4 to 12 weeks) despite the chronic nature of the disorders. Furthermore, the samples tend to be highly selected and that restricts the generalizability (ie, external validity) of the results. For instance, it has been shown that RCTs for major depression evaluate treatment in rarefied samples, excluding as many as 85% of those who are screened.[3] The exclusions are, for example, based on illness severity (too severe or not severe enough) and safety risk (eg, suicidality, concomitant medication, or psychosis). Therefore the results do not inform the treatment of patients who have such features. Finally, the attrition poses a serious threat to the internal validity of a clinical trial. The attrition rates in antidepressant trials range from 30% to 40% and they are higher for antipsychotics, ranging from 50% to 60%.[4] Self-selection of this type (ie, attrition) and of this magnitude severely compromises randomization. Features of design and analysis that reduce the impact of attrition on RCTs of psychotropics have been discussed in detail elsewhere.[1,2] These include strict adherence to the principle of intention to treat, in which *all* randomized subjects are included in the primary analyses,[5,6] use of mixed-effects models that include all available data from participants, even those who terminate the study prematurely,[7,8] and analyses of the sensitivity of results to the assumptions of the analytic model.

Observational studies

In an observational study investigators observe, but do not manipulate, the treatment that is received by participants. Randomized treatment assignment is not used, and this is the most fundamental difference between an observational study and an RCT. In addition, placebo controls and double-blinding of treating clinicians and patients are not used in observational studies, though blinded assessments could be administered. However, RCTs and observational intervention study designs share goals: minimizing bias, having sufficient statistical power, controlling Type 1 error, and providing a feasible design and widely generalizable results. The respective emphasis of each goal varies across the designs.

An observational study's strength is typically applicability, whereas it is more vulnerable to bias. A participant in an observational study receives treatment based on clinician and/or patient selection. That selection is very likely based on illness severity at time of treatment assignment. For example, those with more severe depression could much more likely receive an antidepressant than those less depressed or asymptomatic. (An example of this is provided below using data from the NIMH Collaborative Depression Study.) Furthermore, at the time a treatment decision is made it is quite possible that illness severity will be related to outcome. In other words, treatment assignment could be influenced by a confounding variable or variables. As a consequence, participants who are treated and those untreated are rarely equivalent when treatment commences.

The estimate of the treatment effect in observational studies could very well be biased without proper statistical adjustment. That is, the effect will not reflect the results that would be seen if evaluated in several well-conducted trials of the intervention. If only one variable was responsible for treatment assignment, and that variable was both known and collected, stratified analyses could control the confounding effect. For instance, consider the case where those with health insurance are much more likely to receive an antidepressant intervention (eg, pharmacotherapy, psychotherapy, or implantation device) than the uninsured. Separate analyses for the insured and uninsured (ie, stratified analyses) would

remove the influence of that confounding variable. If the treatment effect was not dissimilar for the insured and uninsured, the results could be aggregated or pooled. However, it is unlikely that the treatment delivery mechanism is explained by just one variable. The focus of this presentation is on a method to reduce bias in the observational estimate of the treatment effect in the presence of multiple confounding variables.

Propensity score adjustment

The propensity score adjustment is used to estimate causal treatment effects with nonequivalent comparison groups and is readily applied to observational studies.[9] The adjustment is a balancing strategy. Its goal is to decrease bias in the estimate of the treatment effect by reducing treatment group imbalance. The propensity score, $e(x)$, represents the conditional probability of receiving treatment, given, for example, pretreatment clinical and demographic characteristics, such that $0 \leq e(x) \leq 1$. The propensity score for each observation is derived from a logistic regression model as described below. Subjects with low propensity scores have characteristics of someone unlikely to get treatment, whereas those with high propensity scores have characteristics of someone more likely to get treatment.

Before the introduction of the concept of the propensity score, Cochran[10] showed that analyses that are stratified into quintiles of a confounding variable will remove >90% of the associated bias. Building on these findings, the propensity adjustment can be implemented through stratification. The rationale for this is that within a propensity score quintile (containing 20% of the observations), the variability in propensity scores will be greatly reduced, albeit not necessarily a constant (as in the health insurance example, above). Based on the concept of restriction of range, propensity score stratification will attenuate the association between the confounding variables and treatment. In addition to stratification, matching, inverse probability weighting, and covariate adjustment are other strategies to implement the propensity adjustment. Covariate adjustment, though most commonly used, can be problematic.[9]

Two stages of propensity analyses

The propensity adjustment is implemented in two stages: (i) propensity score estimation, and (ii) treatment effectiveness analyses. The first stage is the *propensity model*. The propensity score is estimated based on parameter estimates from a logistic regression model for cross-sectional data or mixed-effects logistic regression model for longitudinal data. A preponderance of the applications and evaluations of the propensity methods have involved cross-sectional data. However, evaluations of the performance of propensity quintile stratification with longitudinal observational data have supported its use for bias reduction[11-15] and both examples presented below involve longitudinal data. The dependent variable in the propensity model is treatment (eg, novel vs standard) and the independent variables include demographic and clinical characteristics hypothesized to be associated with treatment.

The propensity adjustment assumes that treatment assignment is strongly ignorable conditional on the propensity score.[9] The plausibility of this assumption can be examined by evaluating the between-treatment group balance on pretreatment variables (ie, the variables included in the propensity score) after implementing the propensity adjustment. For instance, this could be done by comparing treatment groups on baseline variables separately within each propensity quintile. Presumably, the between-group effect size (eg, for baseline illness severity) will be considerably smaller when the propensity adjustment has been implemented, indicating greater baseline between-treatment group balance. With baseline balance, post-baseline groups differences on illness severity can more safely be attributed to the intervention.

The second stage of implementing the propensity adjustment involves *treatment effectiveness analyses*. As implemented in the examples below, the observations are stratified into quintiles of the propensity score. Unlike unadjusted analyses, stratification involves separate analyses for each propensity quintile. Effectiveness analyses might be conducted with a *t*-test of severity ratings or chi-square test of response rates for cross-sectional data. For longitudinal data, in contrast, mixed-effects linear regression, mixed-effects logistic regression, or mixed-effects grouped time survival models, could be used. The choice among these analytic approaches depends on the form of the dependent variable. In each case, treatment is the primary independent variable. The quintile-specific results can be pooled using the Mantel-Haenszel procedure to provide one unified estimate of the treatment effect. However, pooling can

only be used if the assumption of no treatment by quintile interaction has been evaluated and supported empirically. As stated earlier matching, inverse probability weighting, and covariate adjustment provide alternatives to stratification. These alternatives are particularly useful if the sample size precludes quintile stratification, which, of course, involves only 20% of the observations in each quintile-specific analysis.

Observational studies of antidepressant effectiveness

Two examples of observational evaluation of antidepressants are presented below. Each includes two stages of analyses: a propensity model and a treatment effectiveness model. The former examines the magnitude and direction of variables hypothesized to be associated with receiving various ordered categorical antidepressant doses. The latter examines the antidepressant effect relative to a comparator, no antidepressant in these examples. Each example comes from the National Institute of Mental Health Collaborative Depression Study (CDS). The CDS is a longitudinal, observational study that recruited 955 subjects from 1978 through 1981 who sought treatment for one of the major mood disorders (major depressive disorder, mania, or schizoaffective disorder) from one of five academic medical centers in the United States (Boston, Massachusetts; Chicago, Illinois; Iowa City, Iowa; New York, New York; and St Louis, Missouri). All subjects were English-speaking, Caucasian, and at least 17 years of age. Each subject provided informed written consent.[16] Each example below included up to 20 years of follow-up data. These data capture the repeated antidepressant exposure a patient receives during the chronic course of depression: episodes, recovery periods, and recurrences.

Evaluation of acute antidepressant effects

The effectiveness of acute somatic antidepressant treatments as administered in the community was examined.[17] At intake into the CDS, each participant included in the analyses met criteria for major depressive disorder had no history of mania, hypomania, or schizoaffective disorder and had no underlying minor or intermittent depression of at least 2 years' duration. The analyses included 285 participants who recovered from their intake episode and then had at least one recurrent affective episode over the course of the follow-up period. This was done to accommodate the variables included in the propensity model (as described below). The 285 participants had 3141 different antidepressant exposure intervals over the course of time. Each of these intervals constituted a unit of analysis, each with its own propensity score—based strictly on variables assessed prior to the start of the interval. Hence both treatment and propensity for treatment were time-varying, as might be seen in clinical practice.

Classification of antidepressant exposure

Participants were classified based on the ordinal categorical antidepressant dose they received during each week of follow-up. Four ordered categorical antidepressant doses ranged from no treatment to, for example, ≥300 mg imipramine or >30 mg fluoxetine. (Categorical doses for 14 antidepressants are described in detail elsewhere[17,18]). A change from one antidepressant to another did not initiate a new exposure interval, but instead extended the current interval duration, unless the categorical dose was modified. Use of concomitant medications had no bearing on weekly exposure classification. The unit of analysis in both examples presented here is "antidepressant exposure interval," which is defined as a period of consecutive weeks during which the categorical antidepressant dose classification remained unchanged. This is in contrast to most studies where the unit of analysis is the participant per se.

Propensity model

A mixed-effects ordinal logistic regression model examined the propensity for treatment intensity. Treatment intensity was the ordinal-dependent variable, with four ordered categorical antidepressant doses as described earlier.[18,17] Demographic and clinical variables hypothesized to be associated with treatment intensity were included as independent variables in the propensity model. The results indicate that those who had more severe depressive symptoms, more prior episodes, and more intensive somatic therapy in the past were significantly more likely to receive higher antidepressant doses. This suggests that the prior course of illness was more difficult for those who subsequently received higher doses. Nevertheless, treatment comparisons could be made by stratifying effectiveness analyses on the propensity score because the propensity adjustment

removed or greatly reduced the magnitude of the association between each propensity variable and antidepressant dose.

Treatment effectiveness model

The effectiveness outcome involved survival intervals: time from commencing a categorical antidepressant dose until recovery (in weeks), as defined by Research Diagnostic Criteria (RDC).[19] Each survival interval ended in one of three ways: (i) recovery from depressive episode; (ii) change in categorical antidepressant dose; (iii) end of follow-up. The latter two were classified as censored in the survival analyses, and censoring was assumed to be unrelated to outcome. Each subject could contribute multiple survival intervals to the analyses, based on the number of distinct periods during which an antidepressant dose remained constant over the course of the 20-year follow-up.

Treatment effectiveness analyses were initially conducted separately for each propensity score quintile. The effectiveness of each of dose relative to no treatment was tested using mixed-effects grouped-time survival models.[20] The quintile-specific results were then pooled because the propensity quintile by treatment interaction was nonsignificant (-2LL=5.817, df=12, *P*=0.925). (An interaction would have indicated that the treatment effect varied across quintiles, in which case combining results would be inappropriate.) The pooled results indicate that, after controlling for propensity for treatment intensity, those who received higher doses of antidepressant treatment were significantly more likely to recover from a mood episode than those who received no treatment (hazard ratio (HR): 1.86; 95% CI: 1.27-2.72). In contrast, neither moderate doses (HR:1.13; 95% CI:0.79-1.63) nor lower doses (HR: 0.86; 95% CI:0.55-1.23) were associated with recovery.

This observational study broadened the generalizability of antidepressant RCT results. Unlike participants enrolled in RCTs, the CDS sample included those taking concomitant medication, those with substance or alcohol abuse, those with a history of serious suicide attempts, and those with comorbid medical illnesses. In summary, although more severely ill subjects were more likely to commence antidepressant treatment with higher doses, the propensity-adjusted analyses provided an opportunity to demonstrate that those receiving higher doses were more likely to recover.

Evaluation of maintenance antidepressant effects

Two hundred ninety-six CDS subjects were included in this evaluation of antidepressants for the prevention of recurrence of depressive episodes.[15] Among them they had 1782 maintenance antidepressant exposure intervals over 20 years of follow-up.

Propensity for treatment

The propensity model was implemented with a mixed-effects ordinal logistic regression model as described above. The results indicate that those with more prior episodes, those with more severe symptoms, those with primary major depression at CDS intake, and those from the New York, Iowa, and Chicago study sites were significantly more likely to get more intensive maintenance treatment, whereas younger subjects were significantly less likely to get intensive treatment. These differences among participants receiving various doses were accounted for, once again, in effectiveness analyses that were stratified by propensity score quintile.

Using the stratification process, the association in the ordinal logistic regression analysis between each of the variables in the propensity score and antidepressant dose was substantially attenuated. For example, the association of study site with categorical dose was reduced as follows (where Boston was the standard (ie, OR=1.0): New York (OR=2.89; 95% CI: 1.45-5.74; *P*=0.002 in unadjusted model vs OR=1.20; 95% CI: 0.72-1.98; *P*=0.490 in propensity adjusted model); St Louis (OR=1.30; 95% CI: 0.79-2.13; *P*=0.302 vs OR=.93; 95% CI: 0.62-1.40; *P*=0.717); Iowa (OR=2.61; 95% CI: 1.61-4.24; *P*<0.001 vs OR=1.35; 95% CI: 0.91-1.99; *P*=0.138); Chicago (OR=2.49; 95% CI: 1.41-4.41; *P*=0.002 vs OR=1.16; 95% CI: 0.76-1.77; *P*=0.484). Similarly, the association of age with categorical dose was reduced as follows (where ages 30 to 39 years was the standard): <30 years (OR=0.51; 95% CI: 0.37-0.71; *P*<0.001 in unadjusted model vs OR=0.99; 95% CI: 0.73-1.34; *P*=0.949 in propensity adjusted model); ages 40 to 49 (OR=1.11; 95% CI: 0.86-1.42; *P*=0.435 vs OR=1.01; 95% CI: 0.80-1.29; *P*=0.913); ages 50 to 59 (OR=1.31; 95% CI: 0.90-1.90; *P*=0.156 vs OR=1.13; 95% CI: 0.83-1.54; *P*=0.450); ages 60+ (OR=1.34; 95% CI: 0.87-2.07; *P*=0.188 vs OR=1.01; 95% CI: 0.74-1.36; *P*=0.971).

Treatment effectiveness analyses

The effectiveness analyses were conducted with a mixed-effects grouped-time survival model to examine the time until recurrence, which was defined as the number of consecutive weeks during which the categorical antidepressant dose remained unchanged during a "well" period (as defined by RDC[19]). The quintile-specific treatment effectiveness results were pooled because, once again, the treatment by propensity interaction was not statistically significant (-2LL=6:146; df=12; *P*=0.909). The pooled results indicate that participants treated with higher antidepressant doses were about half as likely to experience a recurrence than those who received no somatic treatment (odds ratio (OR): 0.50; 95% CI: 0.30–0.84; Z=-2:60; *P*=0.009). In contrast, moderate doses were associated with marginal protection (OR: 0.65; 95% CI: 0.41–1.01; Z=-1:92; *P*=0.055) and lower doses were not associated with significant protection from recurrence (OR: 0.98; 95% CI: 0.65–1.48; Z=-0.09; *P*=0.929).

This observational evaluation of maintenance antidepressant treatment provides empirical evidence of the effectiveness of higher categorical doses. As in the acute treatment analyses, the more severely ill subjects were more likely to commence higher doses. Nevertheless, the propensity adjustment allowed for evaluation of maintenance antidepressant interventions in a nonrandomized study with a more broadly generalizable study sample than typically seen in RCTs of antidepressants.

Discussion

The observational study has been described and considered here as an alternative to the randomized controlled clinical trial. Although observational studies do not use randomized treatment assignment, evaluation of interventions is possible with an appropriate statistical adjustment, but only if the plausibility of statistical assumptions is carefully evaluated. This data analytic strategy was illustrated with evaluations of acute and maintenance antidepressant effectiveness using the NIMH CDS data for longitudinal, observational analyses of ordered categorical antidepressant doses. Propensity score adjusted analyses demonstrated that participants receiving higher doses during an episode were significantly more likely to recover, even though subjects who received higher doses tended to be more severely ill. Similarly, participants who received a higher dose of maintenance antidepressant therapy were significantly less likely to have a recurrence.

The propensity adjustment provides an opportunity to examine treatment effects in lieu of randomization. However, there are critical assumptions of this approach. First, it is useful to examine the degree to which between-treatment group balance has been achieved with the propensity adjustment. Second, it is essential that the treatment by propensity quintile interaction is tested before pooling quintile specific results. This is because an interaction would signify that the treatment effect varied across quintiles and those quintile-specific results must be reported separately. Third, Rubin highlighted the importance of selection of variables for a propensity score prior to seeing the outcome data.[21] This parallels the practice of designating a primary outcome variable and a primary data analytic procedure in an RCT protocol, prior to collecting data. Finally, D'Agostino and D'Agostino provide an overview of the propensity adjustment and emphasize make that it is not a panacea, particularly with the assumption of no unmeasured confounding variables.[22] A mis-specified propensity model will not reduce as much bias as a model that includes all confounding variables. Simulation studies have shown this with cross-sectional[23] and longitudinal data.[24] It is therefore important to conduct sensitivity analyses.[25]

Randomization, in and of itself, does not insulate an RCT from threats to internal validity. Two common features of antidepressant RCT implementation introduce an observational aspect to group assignment. First, attrition, which is highly prevalent in trials of psychiatric interventions, introduces bias and reduces statistical power, feasibility, and generalizability. There are well-accepted strategies for reducing the impact of attrition. Adherence to the principle of intention to treat, in which *all* randomized subjects are included in the primary analyses, is critically important.[5,6] (Note that *modified* intention to treat seldom includes all randomized subjects and therefore does not reduce as much bias as a *true* intention to treat.) Mixed-effects models can include all data from participants, even those who terminate the study prematurely.[7,8] Analyses of the sensitivity of results to the assumptions of the analytic model are useful components of a data analysis plan. These can include use of pattern-mixture models[26,27] and the assessment and application of predictors of attrition such as the two-item Intent to Attend questionnaire.[28]

A second observational component of an RCT is the flexible-dose study, in which those who fail to respond to a low dose are then offered a greater dose of the intervention. Such a design is inappropriate for dose finding because "self-selection" determines dose. Fortunately, the use of flexible dose RCTs is more limited today than two or three decades ago. The problem of flexible dosing can be obviated by conducting a fixed-dose study that allows for a brief period of titration.[29]
In summary if conditions allow, a RCT is preferable for intervention evaluation. However, there are clinical contexts and patient types that do not lend themselves to randomized treatment assignment (eg, suicidal patients). In such a case, an observational study can inform treatment choice if an appropriate adjustment, such as the propensity score adjustment, is implemented. Regardless of the design, the generalizability of the results is restricted to the type of participants included in the study. ❑

Acknowledgements: Dr Leon has received research support from the National Institute of Mental Health (MH060447, MH068638 and MH092606). In the past 12 months he has served on independent Data and Safety Monitoring Boards for AstraZeneca, Pfizer and Sunovion and has been a consultant to FDA, NIMH, MedAvante and Roche. He has equity in MedAvante.

La evaluación de las intervenciones psiquiátricas en un estudio observacional: aspectos del diseño y análisis

Se contrastan las características de los ensayos clínicos controlados randomizados (ECCR) con los estudios observacionales acerca de la eficacia de las intervenciones psiquiátricas. La randomización guía la asignación del tratamiento en los ECCR, mientras que la selección de los clínicos y los pacientes determina el tratamiento en un estudio observacional. Se consideran las fortalezas y debilidades de los diseños randomizados y observacionales. También se describe el ajuste de tendencias, una aproximación estadística que permite la evaluación de la intervención en un estudio observacional no randomizado. Debe evaluarse cuidadosamente lo plausible que pueda ser la proposición del ajuste de tendencias. Esta estrategia analítica de datos está ilustrada con los datos observacionales longitudinales del Estudio Colaborativo de Depresión del Instituto Nacional de Salud Mental de EE.UU. Las evaluaciones que se presentan aquí examinan la eficacia del tratamiento antidepresivo agudo y de mantenimiento, y demuestran la eficacia de las categorías de dosis más altas.

Évaluation des traitements psychiatriques dans une étude observationnelle : problèmes de conception et d'analyse

Les caractéristiques des études cliniques contrôlées randomisées (ECR) et des études observationnelles concernant l'efficacité des traitements psychiatriques sont contrastées. Dans une ECR, l'attribution du traitement dépend de la randomisation, tandis que ce sont le médecin et la sélection du patient qui déterminent le traitement dans une étude observationnelle. Cet article compare les forces et faiblesses des schémas d'études randomisés et observationnels. Nous décrivons ici en particulier la méthode d'ajustement sur la propension une approche statistique qui permet l'évaluation du traitement dans une étude observationnelle non randomisée. La crédibilité des hypothèses utilisées pour l'ajustement sur la propension doit être soigneusement vérifiée. Cette stratégie d'analyse des données est illustrée par des données observationnelles longitudinales de la NIMH Collaborative Depression Study. *L'évaluation présentée ici examine l'efficacité des antidépresseurs aux phases aiguë et d'entretien et démontre l'efficacité des plus hautes posologies.*

REFERENCES

1. Leon AC, Davis LL. Enhancing clinical trial design of interventions for posttraumatic stress disorder. *J Trauma Stress*. 2009;22:603-611.
2. Leon AC, Mallinckrodt CH, Chuang-Stein C, Archibald DG, Archer GE, Chartier K. Attrition in randomized controlled clinical trials: methodological issues in psychopharmacology. *Biol Psychiatry*. 2006;59:1001-1005.
3. Keitner GI, Posternak MA, Ryan CE. How many subjects with major depressive disorder meet eligibility requirements of an antidepressant efficacy trial? *J Clin Psychiatry*. 2003;64:1091-1093.
4. Khan A, Warner HA, Brown WA. Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: an analysis of the Food and Drug Administration database. *Arch Gen Psychiatry*. 2000;57:311-317.
5. Hill AB. *Principles of Medical Statistics*. 7th ed. New York, NY: Oxford University Press; 1961.
6. Lachin JM. Statistical considerations in the intent-to-treat principle. *Control Clin Trials*. 2000;21:167-189.
7. Laird NM. Missing data in longitudinal studies. *Stat Med*. 1988;7:305-315.

8. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38:963-974.
9. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
10. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24:295-313.
11. Leon AC, Hedeker D. A mixed-effects quintile-stratified propensity adjustment for effectiveness analyses of ordered categorical doses. *Stat Med*. 2005;24:647-658.
12. Leon AC, Hedeker D. A comparison of mixed-effects quantile stratification propensity adjustment strategies for longitudinal treatment effectiveness analyses of continuous outcomes. *Stat Med*. 2006;26:2650-2665.
13. Leon AC, Hedeker D. Quantile stratification based on a misspecified propensity score in longitudinal treatment effectiveness analyses of ordinal doses. *Comput Stat Data Anal*. 2007;51:6114-6122.
14. Leon AC, Hedeker D. A comparison of mixed-effects quantile stratification propensity adjustment strategies for longitudinal treatment effectiveness analyses of continuous outcomes. *Stat Med*. 2007;26:2650-2665.
15. Leon AC, Hedeker D, Teres JJ. Bias reduction in effectiveness analyses of longitudinal ordinal doses with a mixed-effects propensity adjustment. *Stat Med*. 2007;26:110-123.
16. Katz MM KG. Introduction: overview of the clinical studies program. *Am J Psychiatry*. 1979;136:49-51.
17. Leon AC, Solomon DA, Mueller TI, et al. A 20-year longitudinal observational study of somatic antidepressant treatment effectiveness. *Am J Psychiatry*. 2003;160:727-733.
18. Keller M. Undertreatment of major depression. *Psychopharmacol Bull*. 1988;24:75-80.
19. Spitzer RL, Endicott J, Robins E. Research diagnostic criteria: rationale and reliability. *Arch Gen Psychiatry*. 1978;35:773-782.
20. Hedeker D, Siddiqui O, Hu FB. Random-effects regression analysis of correlated grouped-time survival data. *Stat Methods Med Res*. 2000;9:161-179.
21. Rubin D. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26:20-36.
22. D'Agostino RB J, D'Agostino RB, Sr. Estimating treatment effects using observational data. *JAMA*. 2007;297:314-316.
23. Drake C. Effects of misspecification of the propensity score on estimators of the treatment effect. *Biometrics*. 1993;49:1231-1236.
24. Leon AC HD. Quantile stratification based on a misspecified propensity score in longitudinal treatment effectiveness analyses of ordinal doses. *Comp Stat Data Anal*. 2007;51:6114-6122.
25. Rosenbaum P. *Observational Studies*. New York, NY: Springer-Verlag; 2005.
26. Little R. Pattern mixture models for multivariate incomplete data. *J Am Stat Assoc*. 1993;88:125-134.
27. Little R. Modeling the drop-out mechanism for multivariate incomplete *data. J Am Stat Assoc*. 1995;90:1112-1121.
28. Leon AC, Demirtas H, Hedeker D. Bias reduction with an adjustment for participants' intent to dropout of a randomized controlled clinical trial. *Clin Trials*. 2007;4:540-547.
29. Van Putten T, Marder SR. Variable dose studies provide misleading therapeutic windows. *J Clin Psychopharmacol*. 1986;6:249-250.