Are Two Antidepressant Mechanisms Better Than One? Issues in Clinical Trial Design and Analysis

Andrew C. Leon, Ph.D.

The introduction of newer antidepressants that affect both serotonergic and noradrenergic neurotransmission has prompted the question of whether two antidepressant mechanisms of action are better than one for the treatment of depression. Existing data do not provide a definitive answer. Whether future studies provide the answer will depend largely on the quality of trial designs. Some design aspects are dictated by the nature of such a study (e.g., superiority vs. equivalence trial), and others are somewhat more discretionary (e.g., assessment tools, statistical procedures). Issues in the design and analysis of clinical trials comparing dual- and single-action antidepressants are discussed.

(J Clin Psychiatry 2004;65[suppl 4]:31–36)

The selective serotonin reuptake inhibitors (SSRIs) have helped millions of depressed individuals via enhancement of serotonin neurotransmission. Newer antidepressants combine a serotonergic mechanism of action with a noradrenergic component and, consequently, have come to be known as dual-action antidepressants. Some research[1–3] has suggested that dual-action agents are superior to those designed to affect a single neurotransmitter system, and other research[4,5] indicates little or no difference. Because there have been few prospective trials comparing SSRIs and dual-action antidepressants, the limited evidence comes from meta-analyses of archival clinical trial data. Well-designed randomized, controlled clinical trials (RCT) are required to determine whether dual-action agents provide any advantage relative to SSRIs.

One of the primary goals of the RCT design is to minimize the bias in the estimate of the treatment effect. Does the treatment work and, if so, what is the magnitude of the effect? Several critical components of RCT design can be configured to support this goal. First, type I error typically is maintained at .05. Second, the trial must be designed with sufficient statistical power. Third, the trial must be designed in a way that is both feasible (e.g., does not require thousands of subjects) and applicable (i.e., includes the patients that will be targeted for treatment with the investigational agent). Decisions that affect how these standards are represented include selection of the participants, comparison group(s), assessments, data analytic techniques, and, with these choices in mind, sample size requirements.

Two aspects of clinical trial design that minimize bias in the estimate of the treatment effect are randomization and double-blinding. There is little debate about the indispensable nature of randomization and double-blinding for evaluating efficacy, safety, and tolerability. There is certainly less consensus about many of the other issues in RCT design that are discussed here. In fact, for some of these issues related to RCT design, there may be no clear answer.

DESIGN

Superiority, Equivalence, or Non-Inferiority Clinical Trial

In designing a study to determine if two antidepressant mechanisms are better than one, an initial consideration is whether the trial should be configured to evaluate superiority, equivalence, or non-inferiority of one treatment compared with another. The field of psychopharmacology characteristically relies on superiority trials, which are designed to detect a treatment effect that is considered clinically meaningful or an effect size that has been described in the literature. In contrast, equivalence and non-inferiority trials are guided by a margin of equivalence, which is defined as a difference in outcome that is deemed clinically trivial (e.g., a 1-point difference in the Hamilton Rating Scale for Depression [HAM-D]).[6] The margin of

From the Department of Psychiatry, Weill Medical College, Cornell University, New York, NY.

Presented at the symposium, "Pharmacologic Treatments of Major Depression: Are Two Mechanisms Really Better Than One?" which was held on February 10, 2003, in New York, N.Y., and supported by an unrestricted educational grant from Forest Laboratories, Inc.

Dr. Leon has received honoraria from Forest Laboratories, Inc.

Corresponding author and reprints: Andrew C. Leon, Ph.D., Department of Psychiatry, Weill Medical College of Cornell University, 525 E. 68th St., New York, NY 10021 (e-mail: acleon@med.cornell.edu).

Figure 1. An Example of a 2 × 2 Factorial Design Comparing 2 Mechanisms of Action

	SNRI–	SNRI+
SSRI–	Placebo	Reboxetine
SSRI+	Escitalopram	Venlafaxine

Standard ANOVA-type contrasts:

Main effect of SSRI	Main effect of SNRI	Interaction: Add SNRI
H_{01}: μSSRI+ = μSSRI–	H_{02}: μSNRI+ = μSNRI–	H_{03}: μReboxetine – μPlacebo = μVenlafaxine – μEscitalopram

Abbreviations: ANOVA = analysis of variance, SNRI = serotonin-norepinephrine reuptake inhibitor, SSRI = selective serotonin reuptake inhibitor, μ = population mean.

equivalence, which must be clearly defined in the RCT protocol, necessarily is substantially smaller than the clinically meaningful difference that might be used in the design of a superiority trial. Furthermore, equivalence trials are designed with 2-sided equivalence margins, whereas a 1-sided margin is specified for non-inferiority trials. In order to implement an equivalence or non-inferiority trial, the magnitude of medication effect must be stable and well-established in the literature, with consistent results seen from one trial to the next. This consistency clearly has not yet been achieved in the development of antidepressants. Furthermore, the trial being considered asks the question "Are two mechanisms better than one?" rather than "Is one mechanism worse than two?" or "Is one mechanism the same as two?" Hence, a superiority trial is appropriate.

Parallel vs. Crossover vs. Factorial Design

Another primary design question is whether a parallel group, crossover, or factorial design is most appropriate. The parallel group design, in which subjects are randomized to 1 of 2 or 3 treatment cells, is the approach typically used in psychopharmacology trials. In contrast, the primary strength of the crossover design is that the subjects serve as their own controls,[7] although in most psychopharmacology applications the carryover effect renders this design unworkable. Consider an example of the carryover effect in a trial in which subjects are randomized to a sequence of treatments, either AB or BA, where A is the investigational agent and B is the comparator. If a subject responds to the first treatment received in the sequence, it becomes unclear how response rates or reductions in the HAM-D can be evaluated in the second treatment period, which is one reason why a crossover design is incompatible with the example under consideration here.

On the surface it seems that a factorial design has the potential to address the antidepressant mechanism research question. The 2-by-2 factorial design in Figure 1 represents a structure for examining one mechanism versus another. Each cell either includes an SSRI mechanism of action or does not, and each cell either includes a serotonin-norepinephrine reuptake inhibitor (SNRI) mechanism of action or does not. The standard 2-factor analysis of variance (ANOVA) compares the main effects and interaction of SSRI and SNRI mechanisms of action based on the following null hypotheses:

SSRI main effect: Subjects who receive an SSRI do not respond differently than those who do not receive an SSRI.

SNRI main effect: Subjects who receive an SNRI do not respond differently than those who do not receive an SNRI.

SSRI-by-SNRI interaction: The effect of adding an SNRI to placebo does not differ from that of adding an SNRI to an SSRI.

None of these contrasts, however, correspond to the research question: What is the effect of adding the second mechanism of action on efficacy, safety, and tolerability? Thus, instead of factorial ANOVA-type contrasts, each single-mechanism agent must be compared with the dual-mechanism agent in a parallel design (e.g., escitalopram vs. venlafaxine, reboxetine vs. venlafaxine).

Placebo

Should a placebo control be included? Among other features, placebo helps to calibrate a clinical trial. It provides valuable evidence about the implementation of the trial, particularly the placebo response rate. In what kind of clinical setting(s) was the trial conducted? Was it a setting with 15% or 45% placebo response?

Placebo control also provides a context to test assay sensitivity, which represents the degree to which a trial is designed and implemented such that differences between effective and ineffective agents would be detected. By way of illustration, assume that the null hypothesis is not rejected such that a single-mechanism agent does not look different than a dual-mechanism agent. Failure to reject the null hypothesis can mean one of two things: both treatments are effective or neither is effective. The RCT data cannot disentangle those possibilities. In contrast, in a 4-cell trial with 3 active agents and placebo, assay sensitivity is demonstrated if there are differences between at least 1 active agent and placebo. Care must be taken, however, with the interpretation of results when active agents fail to separate from placebo: One may not conclude that the active agent and the comparator are equivalent. One

cost associated with the placebo-controlled design is that recruitment can be more difficult due to ill patients who do not wish to risk receiving placebo.

Sample Selection

How sample selection is defined for a given study has implications for trial design and any conclusions that may be drawn from the results. In designing a phase 4 trial to compare single and dual mechanisms of action, it may be tempting to import the inclusion and exclusion criteria typically used in phase 3 trials. This is because a trial with homogeneous subjects has less within-group variability than a trial with more diverse subjects, and, therefore, sample size requirements are reduced or, with a fixed sample size, statistical power is increased. One feature of a phase 4 trial, however, is the opportunity for more inclusive/less exclusive subject selection. For instance, such a trial could include subjects with psychiatric or other medical comorbidity, or lower baseline severity (e.g., HAM-D of 17). The benefits include faster recruitment and, at the end of the trial, the ability to generalize results to a wider range of patients. The cost includes increased within-cell variability and the consequent increase in sample size requirements.

Measurement

Biostatistician Joseph L. Fleiss opened his text *Design and Analysis of Clinical Experiments*[8] by writing, "The most elegant design of a clinical study will not overcome the damage caused by unreliable or imprecise measurement" (p. 1). He then asserted that high quality data are at least as important as randomization or blinding. To adhere to these principles, careful consideration must be given to the choice of assessments, the number of primary efficacy measures, and the frequency of assessments, each of which influences the sample size requirement.

Choice of assessments. Assessments should be selected, in part, based on the psychometric properties of the scales. Although the HAM-D may be the most widely used severity assessment in depression studies, such popularity is not necessarily indicative of superior test quality. Indeed, it has been suggested that the lack of variation and innovation in trial designs submitted to the U.S. Food and Drug Administration (FDA) during the past 2 decades may reflect an inclination to let previously successful trials dictate the design of newer studies.[9] The recently developed GRID HAM-D expands on the HAM-D by rating both intensity and frequency of each HAM-D item[10] and should prove to be a more reliable scale. There also are other instruments available for consideration, such as the Montgomery-Asberg Depression Rating Scale (MADRS),[11] which was designed to be sensitive to change in a clinical trial, or the Inventory of Depressive Symptomatology (IDS), a relative newcomer that already has extensive psychometric evaluation supporting its use.[12]

One benefit of weighing psychometric properties when selecting efficacy measures is that more reliable assessments reduce sample size requirements[13] because as the reliability of a scale increases, the within-group variability decreases (i.e., there is less measurement error). With less within-group variability, the between-group effect size is larger. Accordingly, the sample size requirement is reduced simply by carefully selecting and implementing a more reliable efficacy measure. Consequently, the trial costs less, takes less time to complete, and thus becomes more feasible.

Frequency of assessments. The typical assessment frequency in depression treatment studies is every 2 weeks, albeit somewhat more frequently in the first weeks of a trial. If one agent is hypothesized to have a faster onset of action than another, the chance of detecting the effect is greater with more frequent assessments,[14] particularly at the beginning of the study. The results of some trials[15–17] suggest that such a strategy may be appropriate in a study of whether two antidepressant mechanisms are better than one.

Constructs to assess efficacy. The conventional approach to assessment in antidepressant trials is to measure the severity of depressive symptoms. In a phase 4 study comparing single and dual mechanisms of action, alternate strategies may be considered. More than one efficacy measure could be used to compare two mechanisms of action, particularly if the different assessments specifically target symptoms that correspond to reductions expected with each mechanism. For example, a symptom severity scale might be most sensitive to the effects of an SSRI whereas a scale measuring functional impairment might be more sensitive to an SNRI. Accordingly, the study could be designed to test whether a dual-mechanism agent is superior on both dimensions by employing corresponding efficacy measures. Alternatively, the primary efficacy measures could include assessments of adverse events and symptom severity.

Type I error adjustments. If more than one primary efficacy measure is specified in the protocol, the FDA requires that a type I error adjustment be incorporated into the study and that the adjustment strategy also be specified in the protocol. The need for adjustment applies only if the protocol states that superiority on any one of the primary efficacy measures provides evidence of efficacy. In contrast, if the protocol states that efficacy requires superiority on every primary efficacy measure, no alpha adjustment is required because the threshold is elevated by design.

The adjustment is required because the risk of type I error increases with multiplicity (i.e., multiple testing). Specifically, the experimentwise (EW) probability of type I error increases with the number of statistical tests (k) and is calculated: $\alpha_{E\omega} = 1 - (1 - \alpha)^k$. For instance, if the alpha level is set at .05 for each test, then the EW type I error rate is .098 for 2 tests, .143 for 3 tests, and .185 for 4 tests.

Table 1. Sample Size Requirements for Various Response Rates

Response Rate		Sample Size	
		1 Efficacy Measure	2 Efficacy Measures
Group 1	Group 2	(α = .05)	(α = .025)
0.30	0.40	376	451
0.30	0.50	103	123
0.40	0.50	408	489
0.40	0.60	107	128
0.50	0.60	408	489
Median increase[a]		…	18%

[a]Represents percentage increase for 2 efficacy measures versus 1.

Hence, multiplicity increases the probability of falsely concluding that an ineffective agent is efficacious unless the alpha is adjusted.

The most common method of controlling for type I error is the Bonferroni adjustment, which partitions the alpha level evenly among the multiple primary efficacy measures that are specified in the protocol. For example, if there are 2 primary efficacy measures and $\alpha_{E\omega}$ of .05 is sought, the Bonferroni-adjusted alpha level would be .05/2 = .025 for the test of each efficacy measure. The Bonferroni adjustment controls type I error tightly when the null hypothesis is true. For example, with this adjustment the EW type I error rate for 2 statistical tests is maintained at .05: $\alpha_{E\omega} = 1 - (1 - .025)^2 = .05$. Another appealing feature of the Bonferroni adjustment is that it can be applied to numerous statistical procedures (e.g., tests of binary, survival, or continuous data) and combinations of those procedures.

Resistance to the Bonferroni adjustment typically stems from the reduction in statistical power for each test, when the null hypothesis is false. Yet, that is only a problem if the sample size determination is based on an unadjusted alpha. If, instead, the statistical power analyses incorporate the Bonferroni-adjusted alpha level, statistical power can be maintained. Table 1 provides examples of the sample size requirements for statistical power of 0.80 with 2-tailed chi-square tests for a range of response rates that are likely in antidepressant trials, separately for alpha levels of .05 and .025. Based on these values, the multiplicity-adjusted sample size is approximately 18% higher when a second efficacy measure is included. Accordingly, given the FDA requirement that RCT protocols state the primary efficacy measure(s) and, if necessary, the alpha-adjustment strategy, it is prudent also to specify a multiplicity-adjusted sample size. In that way, a Bonferroni adjustment does not come at the expense of statistical power when the null hypothesis is false.

ANALYSIS

Data Analytic Procedures

The choice of a statistical procedure is determined by the form of the efficacy measure and the number of assessment times per subject. Moreover, in psychopharmacology clinical trials, in which missing data are ubiquitous, the statistical procedure must readily accommodate the problem of missing data. The goal of minimizing bias in the estimate of the treatment effect was discussed earlier. Missing data are a vulnerable source of bias, particularly with differential dropout across treatments. For instance, in a study comparing single versus dual mechanisms of action, an elevated likelihood of intolerable side effects in one treatment could result in earlier dropout from that cell. For that reason, the data analytic procedure must not completely exclude subjects with missing data.

Survival analysis is one approach that can include subjects with differential follow-up times by comparing treatments on the time until response. For example, the Kaplan-Meier[18] estimate represents the cumulative percentage of subjects in each group who have *not* responded each week; its complement, therefore, is an estimate of the proportion that *has* responded. A subject continues to contribute to that estimate as long as he or she remains enrolled in the trial. A subject's data are deemed *censored* at the time of dropout, and an assumption of survival analysis is that censoring is independent of outcome. One advantage of survival analysis is that there is no reason to distinguish between last observation carried forward and completer analyses. However, survival analysis was conceived to examine the survival time until a terminal event, an event that cannot change, such as death. Thus, an implicit assumption of survival analysis is that once a subject is classified a responder during the course of the trial, the subject will not revert to partial or nonresponse status. Despite the appeal of survival analysis, this assumption is one drawback to its use in psychopharmacology trials.

Mixed-effects models have gained popularity in the last decade,[19–22] particularly as software has become accessible for mixed models.[23–26] The models include both fixed effects, such as treatment group, and random effects, such as a subject-specific intercept or slope. One particularly appealing feature of mixed-effects models is that subjects who have missing data are not completely excluded; instead, all available data are analyzed. A mixed-effects model can include a varying number of observations per subject (i.e., weekly assessments) and can account for within-subject change over time. In an RCT, a treatment by time interaction might be hypothesized, in that the illness severity is reduced more quickly in one group (i.e., the severity slopes for each group diverge over the course of a trial).

There are several forms of mixed-effects models. For example, mixed-effects linear regression models can examine weekly severity ratings such as the HAM-D, MADRS, or IDS.[24] A mixed-effects logistic regression model can be used for weekly ratings of a binary variable, such as responder versus nonresponder status.[23] This can be expanded in a mixed-effects ordinal logistic regression

Table 2. Mixed-Effects Logistic Regression: Examples of Sample Size Requirements (per group) for Various Numbers of Postbaseline Assessment Times[a]

Response Rate		Number of Postbaseline Assessment Times				
Comparator	Investigational Agent	1 (χ^2)[b]	2	4	6	8
0.25	0.35	348	230	181	165	157
	0.40	165	107	84	76	73
	0.45	98	62	49	45	42
	0.50	65	41	32	29	28
	0.55	47	29	23	21	20
	0.60	36	22	17	16	15
0.30	0.40	376	250	196	178	170
	0.45	175	114	90	82	78
	0.50	103	66	52	47	45
	0.55	68	43	34	31	29
	0.60	48	30	24	21	20
0.35	0.45	395	263	207	188	179
	0.50	182	119	94	85	81
	0.55	106	68	53	48	46
	0.60	69	44	34	31	30
0.40	0.50	407	272	214	194	184
	0.55	186	121	96	87	83
	0.60	107	68	54	49	47
0.45	0.55	411	274	216	196	186
	0.60	186	121	96	87	83
0.50	0.60	407	272	214	194	184
Median decrease[c]		…	35%	21%	9%	5%

[a]Assumptions: intraclass correlation coefficient = .40, 2-tailed α = .05, power = 0.80.
[b]For 1 observation per subject, the chi-square test is typically used; provided for comparison.
[c]Represents percentage decrease from contiguous columns (e.g., 2 vs. 4 assessments).

in which the dependent variable has ordinal categories such as full responder versus partial responder versus nonresponder.[23] Use of such a strategy prevents the HAM-D from being arbitrarily dichotomized. Typically, responder status is defined as a 17-item HAM-D < 8. A problem with dichotomization is that it implicitly designates that a score of 8 is more similar to a 30 than it is to a 6. By including partial response, an intermediate position is adopted.

Sample Size Requirements

Mixed-effects models can provide more statistical power with additional assessment times (up to a limit), or they can reduce sample size requirements for a given effect size and level of statistical power. In fact, the sample size requirement is a function of both the between-group differences and the stability of the assessments over the course of the trial as quantified by the intraclass correlation coefficient (ICC).

Examples of the sample size requirement per group for mixed-effects logistic regression analysis are presented for various numbers of postbaseline observations per subject and response rates likely to be seen in antidepressant RCTs (Table 2). For comparative purposes, the corresponding sample size requirements are also presented for the more conventional chi-square test with the continuity correction, in which there is only one observation per subject (e.g., endpoint). The tabled estimates are based on algorithms presented by Diggle et al.[27] and Fleiss,[28] respectively. The median sample size reduction with each additional observation is also presented. The examples assume a 2-tailed alpha level of .05, statistical power of 0.80, and an ICC of 0.40. (Tables for additional ICCs are presented in Leon, in press.[29]) For example, consider a trial designed to detect response rates of 30% versus 50% for single versus dual mechanism agents, respectively. If the trial were designed to compare endpoint response rates (i.e., 1 observation per subject) using a chi-square test, 103 subjects would be required per group. If instead a mixed-effects logistic regression model examined 2 postbaseline assessment points, the sample size requirement would be reduced to 66 subjects per group and for 4 postbaseline assessments, 52 subjects per group.

It should be noted that the ICC in the previous example was chosen somewhat arbitrarily because ICCs for RCTs of depression treatments are not well-established. Several factors influence the ICC including duration of the trial, inclusion and exclusion criteria, frequency of assessment, reliability of the efficacy measure, effectiveness of the agents, and the onset of action. With a lower ICC, that is, as within-subject observations are less highly correlated, sample size requirements are reduced in turn. Therefore, if the ICC were actually 0.20, these would be fairly conservative estimates in that they provide greater than 0.80 power; or stated differently, the sample size requirements for statistical power of 0.80 would be smaller.

Another issue worth considering in designing a trial for comparing mechanisms of action is the desired level of statistical power. In studies funded by the National Institute of Mental Health, power of 0.80 is typical, which leaves a 20% chance of failing to identify an effective drug. Industry-sponsored trials, particularly those unlikely to be repeated, may consider increasing sample size requirements for statistical power of 0.90. To detect group differences in response rates ranging from approximately 10% to 30%, the sample size increases by about 30% for power of 0.90 relative to that for power of 0.80. The corresponding cost is worth serious consideration.

SUMMARY

Although prescription of dual-action antidepressants is on the rise, the literature remains unclear on the issue of whether these newer agents offer incremental benefit compared with SSRIs. Well-designed trials are required to determine if an antidepressant with dual mechanisms of action is superior to a single mechanism agent. Some features of the design are clear: randomized, double-blind, parallel group, and superiority trial. Other issues are less clear. The choice of assessment deserves careful consideration. The HAM-D may have been the most commonly used assessment tool in depression studies over the past

2 decades, but other efficacy measures have greater psychometric support for their use. Furthermore, if more than one primary efficacy measure is specified in the protocol, the proposed sample size estimates should accommodate the required alpha adjustment. If any compound is expected to exhibit faster onset of action, assessments should be administered more frequently than biweekly. Finally, mixed-effects models are well-suited for many psychopharmacology trials because subjects who have missing data can be readily included, and they can provide greater statistical power or allow reduced sample sizes.

Drug names: escitalopram (Lexapro), venlafaxine (Effexor).

REFERENCES

1. Thase ME, Entsuah AR, Rudolph RL. Remission rates during treatment with venlafaxine or selective serotonin reuptake inhibitors. Br J Psychiatry 2001;178:234–241
2. Stahl SM, Entsuah R, Rudolph RL. Comparative efficacy between venlafaxine and SSRIs: a pooled analysis of patients with depression. Biol Psychiatry 2002;52:1166–1174
3. Smith D, Dempster C, Glanville J, et al. Efficacy and tolerability of venlafaxine compared with selective serotonin reuptake inhibitors and other antidepressants: a meta-analysis. Br J Psychiatry 2002;180:396–404
4. Olver JS, Burrows GD, Norman TR. Third-generation antidepressants: do they offer advantages over the SSRIs? CNS Drugs 2001;15:941–954
5. Freemantle N, Anderson IM, Young P. Predictive value of pharmacological activity for the relative efficacy of antidepressant drugs: meta-regression analysis. Br J Psychiatry 2000;177:292–302
6. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56–62
7. Senn S. Cross-Over Trials in Clinical Research. Chichester, England: John Wiley & Sons Ltd; 1993
8. Fleiss JL. Design and Analysis of Clinical Experiments. New York: John Wiley & Sons Ltd; 1986
9. Klein DF, Thase ME, Endicott J, et al. Improving clinical trials: American Society of Clinical Psychopharmacology recommendations. Arch Gen Psychiatry 2002;59:272–278
10. Engelhardt N, Bech P, Evans K, et al. A proposal for a standardized HAM-D scoring system: a collaboration among the pharmaceutical industry, academia, and government. Presented at the National Institute of Mental Health (NIMH) New Clinical Drug Evaluation Unit; May 28–31, 2001; Phoenix, Ariz
11. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979;134:382–389
12. Rush AJ, Gullion CM, Basco MR. The Inventory of Depressive Symptomatology (IDS): psychometric properties. Psychol Med 1996;26:477–486
13. Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. Arch Gen Psychiatry 1995;52:867–871
14. Leon AC. Measuring onset of antidepressant action in clinical trials: an overview of definitions and methodology. J Clin Psychiatry 2001;62(suppl 4):12–16; discussion 37–40
15. Benkert O, Szegedi A, Kohnen R. Mirtazapine compared with paroxetine in major depression. J Clin Psychiatry 2000;61:656–663
16. Gorman JM, Korotzer A, Su G. Efficacy comparison of escitalopram and citalopram in the treatment of major depressive disorder: pooled analysis of placebo-controlled trials. CNS Spectr 2002;7(suppl 1):40–44
17. Schatzberg AF, Kremer C, Rodrigues HE, et al. Double-blind, randomized comparison of mirtazapine and paroxetine in elderly depressed patients. Am J Geriatr Psychiatry 2002;10:541–550
18. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–481
19. Bryk AS, Raudenbush SW. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, Calif: Sage; 1992
20. Goldstein H. Multilevel mixed linear model analysis using iterative generalized test squares. Biometrika 1986;73:43–56
21. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–974
22. Gibbons RD, Hedeker D, Elkin I, et al. Some conceptual and statistical issues in analysis of longitudinal psychiatric data: application to the NIMH treatment of Depression Collaborative Research Program dataset. Arch Gen Psychiatry 1993;50:739–750
23. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. Comput Methods Programs Biomed 1996;49:157–176
24. Hedeker D, Gibbons RD. MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. Comput Methods Programs Biomed 1996;49:229–252
25. Littell RC, Milliken G, Stroup W. SAS System for Mixed Models. Cary, NC: SAS Institute; 1996
26. Pinheiro JC, Bates DM. Mixed-Effect Models in S and S-PLUS: Statistics and Computing. New York, NY: Springer-Verlag; 2000
27. Diggle PJ, Heagerty P, Liang KY, et al. Analysis of Longitudinal Data. 2nd ed. Oxford, England: Oxford University Press; 2002
28. Fleiss JL. Statistical Methods and Rates and Proportions. New York, NY: John Wiley & Sons Ltd; 1981
29. Leon AC. Sample size requirements for comparisons of two groups on repeated observations of a binary outcome. Eval Health Prof. In press