Key Words: attrition, cure model, onset of action, multiplicity

Statistical Strategies for Randomized Controlled Clinical Trials to Detect Differential Onset of Action

By Andrew C. Leon

ABSTRACT *~ Designing a randomized controlled clinical trial to detect differential onset of action of treatments for mood and anxiety disorders poses several challenges. The definition of onset of action must embrace both the degree of symptom reduction and the duration of reduction over time. Perhaps the most difficult aspect of the definition is the classification of subjects with symptom reduction that is not sustained. The scheduling of assessments in such a trial is critical. Unlike standard RCTs for mood or anxiety disorders, weekly or biweekly assessments could prove to be insensitive to group differences. The data analytic procedure must focus on change over time in order to detect differential onset, yet account for problems of multiplicity and attrition. Repeated statistical tests at each successive assessment time might appear to be intuitively appealing, but they require multiplicity adjustments. The data analytic approach must also be flexible enough to incorporate available data from subjects who fail to complete some assessments. The applicability of mixed-effects models and the cure model to challenges in the study of onset of action is described.* Psychopharmacology Bulletin. *2009;42(2):39–46.*

INTRODUCTION

Eleven antidepressants were approved by the US FDA from 1985 to 2004 (US FDA, 2007). As part of the approval process, each of these was shown to be superior to placebo. In an effort to gain a competitive advantage, sponsors have sought to show that their antidepressant has a more rapid onset of action than others. A randomized controlled clinical trial (RCT) designed to demonstrate such an advantage faces many methodological challenges. For instance, the definition of onset of action must be clearly operationalized. Is it response, defined as a predetermined percent reduction in severity, or must remission be achieved? Must onset be sustained? How often should onset be measured? Does the study focus exclusively on efficacy or do safety and tolerability have a role? The investigator must carefully consider the most appropriate of comparison group; and if an active comparator is used must a placebo also be included? The selection of a data

Dr. Leon, Department of Psychiatry, Weill Cornell Medical College, New York, NY.

To whom correspondence should be addressed: Andrew C. Leon, PhD, Weill Cornell Medical College, Department of Psychiatry, Box 140, 525 East 68th Street, New York, NY 10065. Phone: (212) 746-3872; Fax: (212) 746-8754. Email: acleon@med.cornell.edu

analytic strategy for group comparisons is critical. It is the latter that is the focus of consideration here.

In choosing an appropriate statistical procedure, the operational definitions and other features of design must be well characterized. For instance, what is the target for the onset study? Is it probability of response, timing of onset of action, or the duration of effect? A different statistical approach might be used for each.

REPEATED CROSS-SECTIONAL COMPARISONS

The statistical strategies from which to choose can be grouped into cross-sectional comparisons (e.g., χ^2 tests or t-tests) or longitudinal analyses (e.g., mixed-effects models and survival analyses). Perhaps the most common approach is to conduct repeated cross-sectional comparisons and to define onset as the first week of statistically significant separation. For instance, χ^2 tests of weekly onset status or t-tests of weekly ratings were used to compare pindolol + paroxetine with placebo + paroxetine (Bordet et al. 1998). There are several reasons that repeated cross-sectional comparisons are suboptimal. First, the approach is sensitive to small effects with large sample sizes. Second, it is vulnerable to the ubiquitous problem of attrition. Finally, repeated tests inflate the probability of false positive results.

Sensitivity to Small Effects

Initially consider the sensitivity of repeated cross-sectional comparisons to small effects with large sample size. For example, 400 subjects per group would provide 80% statistical power to detect a between group effect size (Cohen's *d*) of $d = 0.20$. Yet few clinicians would consider a difference of 0.20 standard deviation units (e.g., 1 to 2 points on the HAMD or MADRS) to be a meaningful group difference. For that reason the protocol must define what would be viewed as a *clinically meaningful* effect. For example, the magnitude might be identified a priori as Cohen's $d = 0.40$, a success rate difference (*SRD*; i.e., a difference in response or remission rates) of $SRD = 0.20$, or the number needed to treat of $NNT = 5$ as meaningful (see Kraemer and Kupfer, 2006). If an effect is identified that meets the a priori criterion and it is statistically significant, the finding would more likely be viewed as noteworthy.

Multiplicity

A further problem with repeated cross-sectional comparisons is that of multiple hypothesis tests. Multiplicity inflates the probability of a false positive result (i.e., Type I error). For instance, if hypothesis tests were conducted on weekly assessments in a six week RCT,

the probability of type I error would not be the conventionally accepted, nominal level of 0.05, but instead substantially inflated: $1 - (1 - 0.05)^6 = 0.265$. This problem can be circumvented using a multiplicity adjustment. The so-called Bonferroni adjustment, the most common multiplicity adjustment, partitions the α-level of 0.05 among the k weekly tests such that $\alpha^* = \alpha/k$. In an RCT with k assessments over time, the alpha-threshold would be adjusted such that $\alpha^* = 0.0167$, 0.0125, 0.01 and 0.0083 for $k = 3$, 4, 5, 6, respectively. By invoking the adjustment, an upper limit on *experimentwise* type I error (α_{EW}) is imposed. For example, with 6 weekly tests and an adjusted alpha of $\alpha^* = 0.05/6 = 0.0083$, the $\alpha_{EW} = 1 - (1 - 0.0083)^6 \cong 0.05$. That is there would be a 5% chance of false positive result, which is the scientific standard in the field.

There are two primary concerns about using the Bonferroni adjustment. First, it does not account for correlations between the weekly outcome measures. However, it has been shown that the Bonferroni adjustment is not conservative (i.e., it controls Type I error appropriately) unless the correlation is 0.60 or greater (Pocock et al. 1987; Leon et al. 2005). In the case of highly correlated successive weekly assessments (i.e., $r \geq 0.60$), a method that incorporates the correlation in the calculations for the multiplicity adjustment should be used (e.g., James, 1992; Leon et al. 2005; Leon et al. 2007). The second reservation about employing the Bonferroni multiplicity adjustment has to do with reductions in statistical power, which could result in false negative findings. However, this is only a problem if the sample size estimates, calculated before the trial has commenced, fail to take into account the multiple testing issue. A researcher can maintain statistical power at the design stage of an RCT with multiplicity-adjusted sample size estimates; that is, those based on adjusted α^*. A word of caution: this will increase the required sample size, for example, by over 50% for six hypothesis tests. As a consequence, multiple tests increase research costs, study duration and the number of subjects exposed to risk of an experiment (Leon, 2004). Alternatives to multiple cross-sectional analyses are described below in the discussion of longitudinal analyses.

Attrition

The third problem with the strategy of using repeated cross-sectional comparisons to identify the groups with the most rapid onset of action is that of attrition. Attrition introduces bias, reduces statistical power, feasibility and generalizability of an RCT (Leon et al. 2006). There are three general data analytic approaches to attrition: 1) analyze complete cases only 2) impute data 3) analyze incomplete data. Complete case analyses fail to adhere to the randomization strategy; that is, they do not analyze

subjects as they were randomized. Instead, self-selection plays a role in treatment assignment. As a consequence, complete case analysis is exceedingly vulnerable to biased estimates of the treatment effect.

The most common approach to imputation (i.e., replacing missing values) in psychopharmacology has been last observation carried forward (LOCF). There is no statistical theory supporting its use. LOCF has also been shown to yield biased estimates, which despite common belief, could favor either active or placebo (Mallinckrodt et al. 2004). If imputation is to be used, multiple imputation is much more appropriate (Rubin, 1987; Shafer, 1997) because inherent in the strategy is incorporating the uncertainty of the imputation process. The technical details regarding multiple imputation have been described elsewhere and are beyond the scope of this presentation. Instead, two approaches to the analysis of incomplete data are considered: mixed-effects models and survival analysis. Each analyzes longitudinal data, albeit those of markedly different configurations.

LONGITUDINAL DATA ANALYSES

Mixed-effects Models

Mixed-effects models can examine illness severity over the course of an RCT (Laird and Ware, 1982; Hedeker and Gibbons, 2006). The unit of analysis is not the subject, per se, but instead it can be the weekly or biweekly assessments within a subject. Therefore, there is no need to exclude subjects with incomplete data. Consider the mixed-effects linear regression model that is used to examine a dimensional outcome such as the HAMD, MADRS, or QIDS-SR. In an RCT, of course, group assignment is based on randomization and as a result, the cells should be fairly well-balanced at baseline. Therefore, assume baseline similarity on the primary outcome. If we further assume a linear change in severity over the course of the RCT, the treatment effect can be quantified as group differences in the rate of change over time, i.e., differential slopes. The slope represents the weekly rate of change in illness severity. If attrition can be more or less accounted for by the observed measures of outcome or observed covariates that are included in the model, mixed-effects models are an advantageous approach to the problem of attrition. Attrition of this sort is referred to as "ignorable" and with ignorable attrition, mixed-effects models can be used for valid inference (Laird, 1988).

A mixed-effects linear regression model can examine the onset of action of an antidepressant by empirically determining the timing of differential onset of symptom reduction. This would initially involve an

omnibus test of the treatment by time interaction. If this test provides evidence that there is a differential rate of decline of symptom severity at some point during the trial, post hoc tests could identify the earliest separation between groups with treatment by time interaction terms that use weekly indicator variables, as opposed to comparing linear slopes as done in the omnibus test. The fact that there are differential slopes does not reveal the magnitude, direction, or stability of the treatment effect. Three of the possible scenarios involving differential slopes consistent efficacy have been contrasted (Leon, 2001): 1) Earlier onset for one cell and superiority maintained throughout the trial 2) Earlier onset for one cell, but by the final week of the trial there is no difference between cells 3) Earlier onset for one cell, but by the end of the trial the other cell has superior performance. Therefore in reporting results, an investigator must clearly indicate the magnitude and direction of the treatment effect.

There are also mixed-effects models for categorical outcomes. For instance, a mixed-effects logistic regression model can examine weekly binary response status and mixed-effects ordinal logistic regression (Hedeker and Gibbons, 1994) can examine weekly ordinal response status (e.g., full response vs. partial response vs. nonresponder as applied in Kocsis et al. 2009).

Survival Analysis

An alternative longitudinal approach to examining onset of action is survival analysis, which can target the time until onset. Survival analyses will include some data from all subjects and therefore, there is no need to impute data for dropouts. The information collected on a participant who discontinues from the trial prematurely is incorporated up until the point at which the data are censored, i.e., the point at which that subject discontinued study participation and no longer provided assessments. The Kaplan–Meier product limit estimate (Kaplan and Meier, 1959) quantifies the cumulative response rates over the course of the trial. Three assumptions of survival analysis must be considered. First, it is assumed that dropout is not related to outcome, also referred to as "non-informative censoring". This assumption if often plausible, but would be problematic if numerous subjects became so depressed during the course of the trial that they did not attend subsequent assessment sessions. Second, it is assumed that the response, once achieved, is sustained. This is undoubtedly plausible when analyses examine time until a terminal event, such as death. It is less plausible in analyses of a disorder characterized by symptoms that wax and wane. In a study of onset of action, the assumption requires that the event of interest be a state that is sustained until the end of the trial. Finally, it is implicitly assumed in survival analysis that with enough

time all participants will respond. This final assumption is clearly implausible in studies of antidepressants. Those who have not responded within six or twelve weeks are unlikely to respond without a change in intervention.

The logrank test is typically used to compare groups on response rates over time, but its applicability to the group comparison of onset of action is less clear. Consider a hypothetical example. Assume that during an RCT for depression there was no response before week 4 in either cell. Then, at week 4, 20% of those randomized to placebo responded and 50% of those randomized to active responded. The active cell clearly has higher probability of response. However, those in the active cell do not have faster onset. Yet with a sufficient sample size, these groups would differ significantly in a logrank test comparing survival curves. Hence an alternative approach is necessary for investigations of onset of action.

Cure Model

The Cure Model provides an alternative data analytic strategy. It is referred to as a mixture model in that it mixes the experience of responders and nonresponders (Boag, 1949; Farewell, 1982; Farewell, 1986; Laska and Meisner, 1992). The model separately examines the probability of response and the time to onset. The latter model is a conditional survival analysis, in that it only includes responders with sustained response.

The cure model is expressed: $H(t) = p\ S(t) + (1 - p)$, where p = probability of response; $S(t)$ = probability that time to response $>t$ among subjects who respond and $H(t)$ = probability that time to response $>t$.

"Cure" refers to surviving beyond a fixed follow-up time with no change in status. The nomenclature stems trials such as those for analgesics, in which survival time is time until relapse (e.g., time after relief from pain; Laska and Meisner, 1992). In an RCT examining onset of action of an antidepressant, the "cured" are the non-responders, hence the nomenclature is counterintuitive for trials of onset of action for acute treatments (Tamura, Faries and Feng, 2000). Two variations on the cure model are the Cramer-von Mises statistic (Tamura, Faries and Feng, 2000) and a competing risks model (Betensky and Schoenfeld, 2001). Each assumes no more than moderate censoring.

To provide an approximation on the required sample size in a cure model, consider the simulation study of Tamura et al. (2000). In using the Cramer-von Mises approach a sample size of about 100/group needed for >80% power to detect hazard ratio of 2.0, but the sample

size depends on censor rates (i.e., the proportion who either do not respond or discontinue the study prematurely). Although this sample size might seem manageable, it assumes a hazards ratio of 2.0, which is a rather substantial treatment effect.

SUMMARY

The choice among statistical strategies for an RCT that is designed to detect onset of action has been considered. The data analytic model that is selected must incorporate available data without excluding subjects with incomplete data. Both survival analysis and mixed-effects models have such capacity. Mixed-effects models can be used to empirically identify differential onset with a focus on group by time interaction. The cure model, which is a variant of survival analysis, is appealing for examining onset of action because it focuses on two critically important aspects of the response. It examines the probability of response and then, among responders, examines the time to event. Mixed-effects models and the cure model can detect differences in response rates even when initial observations of onset are simultaneous across groups. In contrast, repeated χ^2 tests or t-tests inadequately account for attrition and, if used, must account for multiplicity and make requisite sample size adjustments. Assumptions of the cure model and mixed-effects models were described above and the plausibility of those assumptions will vary with study design and attributes of attrition.

It is important to note that a data analytic procedure cannot be selected until operational definitions are in place. The definition of onset of action must address both the degree of symptom reduction and the duration of reduction, including the classification of subjects with symptom reduction that is not sustained. Each aspect of the design and analysis of an RCT for investigating onset of action must be sensitive to differential change over time.✤

ACKNOWLEDGEMENTS

This manuscript was prepared, in part, with funding from the National Institute of Mental Health (MH060447 and MH068638). Presented, in part, at the annual meeting of the American College of Neuropsychophamacology in Scottsdale, AZ, December 8, 2008.

DISCLOSURES (PAST 12 MONTHS)

Dr. Leon serves on Data and Safety Monitoring Boards for AstraZeneca, Dainippon Sumitomo Pharma America and Pfizer; serves as a consultant/advisor to FDA, NIMH, Cyberonics, MedAvante and Takeda; and has equity in MedAvante.

REFERENCES

1. Betensky RA, Schoenfeld DA. Nonparametric estimation in a cure model with random cure times. *Biometrics*. 2001;57:282–286.
2. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J Royal Statistical Society, Series B*. 1949;11:15–44.
3. Bordet R, Thomas P, Dupuis B. Effect of pindolol on onset of action of paroxetine in the treatment of major depression: intermediate analysis of a double-blind, placebo-controlled trial. *Am J Psychiatry*. 1998;155:1346–1351.
4. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*. 1986;38:1041–1046.
5. Farewell VT. Mixture models in survival analysis. Are they worth the risk? *Canadian J of Statistics*. 1986;14:257–262.
6. Hedeker D, Gibbons RD. *Longitudinal Data Analysis*. Hoboken N.J.: John Wiley and Sons, 2006.
7. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics*. 1994;50:933–944.
8. James S. The approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. *Statistics in Medicine*. 1991;10:1123–1135.
9. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958;53:457–481.
10. Kocsis JH, Leon AC, Markowitz JC, Manber R, Arnow B, Klein DN, Thase ME. Patient preference as a moderator of outcome for chronic depression treated with Nefazodone, Cognitive behavioral analysis system of psychotherapy, or their combination. *J Clinical Psychiatry*. 2009;370:354–361.
11. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*. 2006;59:990–996.
12. Laird NM. Missing data in longitudinal studies. *Statistics in Medicine*. 1988;7:305–315.
13. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics*. 1982;38:963–974.
14. Laska EM, Meisner MJ. Nonparametric estimation and testing in a cure model. *Biometrics*. 1992;48:1223–1234.
15. Leon AC. Measuring onset of antidepressant action in clinical trials: an overview of definitions and methodology. *J Clinical Psychiatry*. 2001;62(suppl 4):12–16.
16. Leon AC. Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power when using the bonferroni adjustment. *J Clinical Psychiatry*. 2004;65:1511–1514.
17. Leon AC, Heo M, Teres JJ, Morikawa T. Statistical power of multiplicity adjustment strategies for borrelated binary endpoints. *Statistics in Medicine*. 2007;26:1712–1726.
18. Leon AC, Heo M. A comparison of multiplicity adjustment strategies for correlated binary endpoints. *J Biopharmaceutical Statistics*. 2005;15:839–855.
19. Leon AC, Mallinckrodt CH, Chuang-Stein C, Archibald DG, Archer GE, Chartier K. Attrition in randomized controlled clinical trials: methodological issues in psychopharmacology. *Biological Psychiatry*. 2006;59:1001–1005.
20. Mallinckrodt CH, Watkin JG, Molenberghs G, Carroll RJ. Choice of The primary analysis in longitudinal clinical trials. *Pharmaceutical Statistics*. 2004;3:161–169.
21. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics*. 1987;43:487–498.
22. Rubin D.B. *Multiple imputation for nonresponse in surveys*. New York: John Wiley and Sons, 1987.
23. Tamura RN, Faries DE, Feng J. Comparing time to onset of response in antidepressant clinical trials using the cure model and the Cramer-von Mises test. *Statistics in Medicine*. 2000;19:2169–2184.
24. United States Food and Drug Administration. *Briefing document for Psychopharmacologic Drugs Advisory Committee*. December 13, 2006:64. www.fda.gov/ohrms/dockets/ac/06/briefing/2006-4272b1-01-FDA.pdf (Accessed April 18, 2009).