Statistical Approaches to Effectiveness Measurement and Outcome-Driven Re-Randomizations in the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Studies

by Sonia M. Davis, Gary G. Koch, C.E. Davis, and Lisa M. LaVange

Abstract

The design of the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) schizophrenia and Alzheimer's disease studies pose several statistical challenges, including issues related to performing multiple comparisons, defining effectiveness outcomes, and collecting and analyzing data from a design with multiple outcome-driven re-randomizations. We discuss the CATIE strategy for addressing many hypotheses within the context of one clinical trial while controlling the overall type I error rate. We provide motivation for the use of two effectiveness outcomes: time to all-cause discontinuation and composite endpoints that combine outcomes from multiple domains, such as efficacy, safety, cost-effectiveness, and quality of life. Methods for statistical analysis of an outcome-driven re-randomization trial are compared and evaluated. We describe analysis within each phase, analysis based on the first randomization or treatment algorithms, and repeated measures modeling. Finally, strategies are described for designing an electronic data collection system for trials with repeated outcome-driven re-randomizations.

Keywords: Multiple comparisons, effectiveness, treatment discontinuation, composite endpoint, outcome-driven re-randomization, electronic data capture.

***Schizophrenia Bulletin*, 29(1):73–80, 2003.**

The Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project, sponsored by the National Institute of Mental Health (NIMH), consists of two clinical trials with the goal of studying the effectiveness of atypical antipsychotic medications for the treatment of schizophrenia and Alzheimer's disease (Schneider et al. 2001; Stroup et al., this issue). Both studies are multiphase double-blind clinical trials in which patients are randomized to one of several medications and are subsequently re-randomized to a new medication upon failure or discontinuation of the first treatment.

The schizophrenia trial has a planned sample size of 1,600 patients. In the first phase, patients are randomized to one of four atypical antipsychotic medications (olanzapine, quetiapine, risperidone, and ziprasidone) or a typical antipsychotic (perphenazine). Upon treatment failure assessed by the investigator, patients may discontinue the first treatment phase and be re-randomized to a new double-blind treatment. At the second phase, patients choose a path between an arm that randomizes them equally to either ziprasidone or one of the three other atypical antipsychotics, and an arm that randomizes them equally to either clozapine or one of the three atypicals. The investigator may subsequently choose to discontinue the second treatment phase and enter the patient in an unrandomized open-label phase. The goal is to follow each patient for 18 months. If patients refuse treatment with study medication, they may enter into a followup phase to measure key assessments for the remainder of the 18 months.

The Alzheimer's disease trial has a planned sample size of 450 patients. In the first phase, patients are randomized to one of three atypical antipsychotic medications (olanzapine, quetiapine, risperidone) or placebo. If the investigator chooses to discontinue the medication, patients may be re-randomized to a new double-blind treatment. At the second phase, patients originally randomized to an atypical antipsychotic are re-randomized to a new atypical antipsychotic or citalopram in a 3:3:3:2

Send reprint requests to Dr. S. Davis, Quintiles, Inc., 5927 South Miami Boulevard, Morrisville, NC 27560; e-mail: sonia.davis@quintiles.com.

ratio, and patients originally randomized to placebo are randomized in a 1:1:1:3 ratio, so that half of patients originally randomized to placebo receive citalopram. The investigator may subsequently choose to discontinue the second treatment phase and enter the patient in a randomized open-label phase, in which patients receive one of the active medications not previously received. The goal is to follow each patient for 36 weeks. If patients or caregivers refuse treatment with study medication, they may enter into an open choice phase for the remainder of the 36 weeks.

These study designs pose several challenges for statistical analysis: (1) consideration of how the type I error rate can be controlled for multiple comparisons while addressing many hypotheses within a complex research protocol, (2) strategies for developing overall effectiveness outcomes, (3) strategies for analyzing data from a study with multiple unequally timed treatment phases, and (4) strategies for designing a data collection system for this study design.

Multiple Comparisons

The CATIE complex study designs involve three distinct types of multiple comparison issues. First, there are many assessment outcomes, including time to treatment phase discontinuation, efficacy assessments, safety assessments, neurocognitive measures, cost-effectiveness, quality of life, other secondary scales, and genetics measures. Second, there are multiple treatment phases in which patients are re-randomized. And third, there are many treatments to be compared. In most phases, there are four to six possible treatments.

At the completion of the studies, the data bases for both protocols will be some of the largest and richest research data bases available in the areas of schizophrenia and Alzheimer's disease. Designed as research projects, they contain many outcomes worthy of careful analysis. However, with so many assessments, the question of how to control for multiple comparisons looms large.

In the regulatory setting, the number of questions that can be definitively addressed by a clinical trial is generally one per study, or one sequence of questions handled through appropriate methods of adjusting for multiple comparisons, so that the overall study-wise type I error rate is maintained at 0.05. Most commonly, one primary outcome is identified. All other evaluations are secondary, supportive, or exploratory. Another option is to identify a small number of primary outcomes and control the type I error rate across these outcomes by a method such as (1) a composite analysis, (2) a step-down testing procedure, or (3) by applying a significance-level adjustment such as Bonferroni or Bonferroni-Holm (Hochberg 1988; Bauer 1991; Westfall et al. 1999; Koch 2000; Moye 2000).

For a complex research project like the CATIE trials, this regulatory model is overly stringent. In the setting of research protocols that rigorously measure many outcomes and have substantial sample size, a reasonable strategy for addressing multiple assessment outcomes is to define a set of domains that are distinct and not confounded with each other, and then to maintain the type I error rate at 0.05 within each domain. One primary domain has been identified by each CATIE protocol: the overall assessment of effectiveness in the first phase. Other a priori domains are loosely defined by the types of assessments collected (Swartz et al., this issue). These domains include, but are not limited to, efficacy and safety, cost-effectiveness, quality of life, neurocognitive functioning, caregiver quality of life for Alzheimer's disease, and competency to give informed consent for schizophrenia. Some research questions within these secondary domains have clearly defined objectives, such as comparisons of cost-effectiveness and specific adverse event profiles known to be associated with the treatments. Other research questions are in fact exploratory and hypothesis generating.

Within the primary domain of overall effectiveness, each protocol has one clearly defined primary outcome: time to Phase 1 treatment discontinuation. The schizophrenia trial has an additional primary domain based on the overall assessment of effectiveness in Phase 2. In the Alzheimer's disease trial, time to Phase 1 discontinuation is the primary outcome for the placebo versus active treatment comparison, and the primary outcome for comparison of the active treatments in Phase 1 is actually a noninferiority test of an assessment within the efficacy domain, the Clinical Global Impression (CGI; Schneider et al. 1997) score at week 12. The protocols clearly identify both a primary outcome and a primary phase for each primary domain. The error rate for multiple treatment group comparisons within the primary outcomes is maintained at 0.05 by methods specified in the protocol via step-down tests and Bonferroni-Holm significance-level adjustments.

Other overall effectiveness outcomes may be evaluated as secondary outcomes. For all secondary outcomes, statistical tests are interpreted as descriptive rather than confirmatory tools. *P* values obtained from comparisons will be presented to identify substantial treatment differences without drawing conclusions about statistical significance. However, within each outcome, the type I error rate will be maintained by appropriate strategies for comparing multiple treatment groups and multiple phases.

The CATIE publications committee ensures that each statistical analysis reported by the CATIE team is focused on one domain and that each domain has a multiple com-

parisons strategy that addresses multiple outcomes, multiple treatments, and multiple phases, so that each domainwise error rate is maintained at 0.05. The complete list of domains need not be specified a priori, as long as each domain is determined by the publications committee to be logically independent of other domains.

Effectiveness Outcomes

The objective of effectiveness trials is somewhat shifted from that of traditional clinical trials. Rather than a protocol in which efficacy and safety can be carefully measured and evaluated in a controlled setting, the goal is instead to evaluate a treatment in a setting as close as possible to usual patient care. Outcomes of interest for effectiveness trials include cost-effectiveness and quality of life as well as efficacy, safety, and tolerability measurements (Bombardier and Maetzel 1999; Revicki and Frank 1999).

The CATIE trials are designed to mimic a real-world experience by having open inclusion and exclusion criteria, by allowing the physician to change the dose of the double-blind randomized treatment whenever warranted, and by allowing the physician to discontinue the randomized medication at any time and for any reason. Additional randomized phases allow for the opportunity to evaluate second line medications and allow the patients to be followed for the full length of the trials. The primary CATIE outcome domain is overall effectiveness in Phase 1. Within this domain, we chose to pick one single primary outcome that would evaluate components of efficacy, safety, and tolerability combined. Individual efficacy and safety outcomes, as well as cost-effectiveness and quality of life, will be evaluated as secondary domains. Several types of overall effectiveness outcomes are described here. The one that we have chosen for the primary analysis is time until all-cause treatment discontinuation.

Time to Discontinuation. Time until treatment phase discontinuation has several advantages that make it an excellent choice as a measurement of overall effectiveness. First, it is a simple assessment that can easily be understood and interpreted by a large audience, without needing to know details of specific measurement instruments or complex statistical techniques (beyond knowledge of survival analysis). This outcome can be simplified even further in a secondary analysis to the proportion of patients who discontinue the phase from each treatment group. Another trait adding to its simplicity is that by concentrating on the data from only one phase, we eliminate complexities introduced by attrition and subsequent re-randomizations (see Analyzing Multiple Outcome-Driven Re-Randomizations).

Second, by including all reasons for treatment discontinuation, this outcome encompasses lack of efficacy, intolerable side effects, or both, as well as lack of compliance, plus any other reason that led to substantial dissatisfaction with the medication, without having to identify these reasons. The reason for discontinuation is collected, so that secondary evaluations can be carried out on subsets of patients who discontinue for specific reasons. The simplicity and generality of time to discontinuation make it a very appropriate overall effectiveness outcome.

Time to discontinuation would not be an appropriate effectiveness outcome for acute illnesses that can be completely healed within a short period, or intermittent illnesses with symptoms that come and go. In these cases, patients who discontinue the drug because it is no longer required would need to be carefully separated from those for whom the drug was ineffective or intolerable. Unfortunately, this scenario is not applicable to long-term illnesses such as schizophrenia and Alzheimer's disease. Although no longer needing the medication is an acceptable reason for discontinuation in the Alzheimer's disease trial, this has so far been a very rare occurrence. For these rare cases, patients can be reclassified as completers and their time to Phase 1 discontinuation can be recoded to the planned study duration.

One might consider excluding administrative discontinuations from the outcome definition. However, defining the overall effectiveness outcome as all-cause treatment discontinuation is a more unbiased assessment than one that attempts to identify whether a discontinuation is related to treatment or not. In the effectiveness setting, many discontinuations that appear administrative—such as patient refusal, patient dropout, and loss to followup—may in fact be related to poor efficacy, safety, or compliance. Administrative discontinuations that are truly unrelated to treatment, such as a patient's family moving away from a site, are fairly uncommon during a trial and should be balanced across treatment groups by randomization. For the Alzheimer's disease trial, placement in a nursing home does not require the discontinuation of a phase, because investigators continue treatment of patients after nursing home placement whenever possible.

Composite Measures. The broad generalization of efficacy, safety, and compliance is one of the strengths of time to treatment discontinuation as an effectiveness outcome. However, the generality can also be thought of as a detriment because it is not overly descriptive of a patient's condition. It is of interest to define secondary outcomes that are more descriptive. Many outcomes describe either efficacy or safety separately, such as time to discontinuation by reason for discontinuation, primary efficacy assessments at the last observation in the treatment phase,

or individual safety assessments such as rates of adverse events during the phase. One refinement that is more descriptive than time to treatment discontinuation and yet still combines efficacy, safety, and tolerability is a composite effectiveness outcome. This type of outcome can be defined as an ordinal classification that combines discontinuation status with the outcome of a key efficacy measurement. A five-level composite outcome, as defined in table 1, generally has a sufficient number of categories to describe the discontinuation status in addition to how well the patient was doing immediately prior to discontinuation. Treatment groups can be compared with a Mantel-Haenszel mean-score chi-square test for ordinal data or proportional odds regression.

A third strategy for combining outcomes is to form a composite score by averaging rankings for several continuous responses per person, in the style of an O'Brien rank-sum test (O'Brien 1984). This can be done by converting the time to treatment phase discontinuation to a Wilcoxon or log-rank score (Koch et al. 1985), converting the Positive and Negative Syndrome Scale (PANSS; Kay et al. 1987) total score percentage change from baseline to a rank score, standardizing each set of scores to z scores having a mean of 0 and a standard deviation of 1, averaging the two z scores, and comparing this average across the treatment groups with either a Mantel-Haenszel mean-score chi-square test or a Wilcoxon rank-sum test.

The main advantage of an O'Brien-type composite score is that it uses the complete information from measurements, rather than a categorization, and therefore is more powerful than an ordinal composite categorization. In addition, it can easily be more generalized by averaging other domains of effectiveness, such as assessments of cost-effectiveness and quality of life. With several outcomes averaged together, one could contemplate assigning different weights to the outcomes, so that some outcomes, such as time to discontinuation, may contribute more to the overall analysis than other outcomes, such as quality of life. The disadvantage of this type of composite analysis is that it is difficult to know what a clinically meaningful difference between treatment groups might be. Descriptive parameters are limited to statistical p values comparing treatment groups, followed by descriptive statistics of each individual parameter. Treatment group comparisons for the individual parameters can be evaluated for statistical significance in a step-down fashion (Lehmacher et al. 1991).

Analyzing Multiple Outcome-Driven Re-Randomizations

The primary analysis of time to Phase 1 discontinuation is not affected by subsequent randomizations. However, there are many hypotheses regarding the subsequent phases and outcome responses across the phases that we wish to investigate. Analyses based on data from subsequent phases can become complex, depending on the strat-

Table 1. Composite ordinal effectiveness outcome

Composite effectiveness category	Phase 1 treatment discontinuation status	PANSS total score percent improvement from baseline to end of phase
1. Best outcome: Completed with improvement	Completed entire study without discontinuing the first treatment phase	> 20%
2. Discontinued with improvement	Discontinued for any reason other than death, adverse event, or side effect	> 20%
3. Completed without improvement	Completed entire study without discontinuing the first treatment phase	≤ 20%
4. Discontinued without improvement	Discontinued for any reason other than death, adverse event, or side effect	≤ 20%
5. Worst outcome: Discontinued for safety	Discontinued because of death, adverse event, or side effect	Any score

Note.—PANSS = Positive and Negative Syndrome Scale.

egy employed. Here we describe four main analysis strategies: (1) analyze each phase separately; (2) analyze data from any time point but focus comparisons on the first randomized treatment group; (3) analyze data from any time point, incorporating all of the preceding randomizations in a treatment algorithm approach; and (4) employ a complex repeated measures model. Each of these methods has both advantages and limitations, and the choice of which to use depends on the research questions being asked.

Analysis of Each Phase Separately. Analyzing each phase separately addresses questions of the form "Which is the best first line medication?" and "Which is the best second line medication?" Endpoints for subsequent phases can be defined exactly as for the first phase, such as time from re-randomization to discontinuation, change from baseline to last available observation in the current phase for an efficacy assessment such as the PANSS total score, or incidence of phase-emergent adverse events. Here the baseline of interest is the last measurement immediately prior to re-randomization in the current phase. Statistical models comparing the treatment groups in a given phase should adjust for the treatment received in the earlier phase(s), as well as other covariates such as the investigator site and the baseline value. The interaction between prior and current treatment assignments could be used to explore whether any sequences of treatments yielded substantially better or worse outcomes during the second phase. However, because there are many possible treatments in CATIE, such comparisons will have little power.

For the schizophrenia trial, it is of particular interest to investigate second line medications. The second phase is designed to compare either clozapine or ziprasidone against the other three atypical antipsychotics combined in subjects for whom the first line of treatment failed. To fully address the second line hypotheses, the study design specifies that a complete set of all assessments be taken at the end of each phase. These assessments ensure that the patient's state is thoroughly evaluated immediately prior to discontinuing the phase. These data provide endpoint measurements for the previous phase as well as baseline assessments for the subsequent phase. Although the Alzheimer's disease trial is also interested in citalopram as a second line medication, this comparison is of secondary interest, and the study design does not require a complete assessment prior to switching phases. Therefore, by-phase analyses for subsequent phases in the Alzheimer's trial are limited to outcomes such as time until phase discontinuation, or outcomes measured at each visit such as GCI.

Analysis Based on the First Randomization. It is sometimes of interest to evaluate study outcomes at fixed points in time by comparing the first randomized treatment and ignoring subsequent re-randomizations. This analysis addresses the question of whether there are any differences in long-term outcomes as a function of the first treatment a patient received. This strategy is particularly applicable to continuous cost-effectiveness and quality of life outcomes, as well as some efficacy assessments.

The goal of this approach is to assess whether or not any treatments are superior or inferior as initial therapy, regardless of any subsequent treatments. The patient's state at a particular time point (i.e., the end of the trial) is the main focus, and the patient's state at the time of treatment re-randomization is of less interest. This analysis strategy is not intended to identify acute treatment differences and may lead to apparently contradictory conclusions about individual treatments. For example, if many patients are discontinued from a first line treatment at an early stage and re-randomized to a more effective treatment, the short poor performance of the first treatment is likely to be masked in the final outcome by a longer improved response to the subsequent treatment. An analysis of measurements taken at the end of the study may find no substantial difference between weaker and stronger first line medications because of the effect of later medications.

Therefore, analyses of long-term outcomes for first line medications are often most helpful as subsequent analyses in addition to the analysis of the responses at Phase 1 discontinuation. If differences are found between treatments at the end of Phase 1, then it is of interest to find out if such differences are maintained or increased as time passes, or if an initial advantage is ultimately lost. For some parameters, if no long-term treatment differences are identified, then a conclusion may be that it does not matter which medication is prescribed first. However, if long-term results are contradictory to results obtained at earlier time points or at the end of the first phase, then conclusions about the relative effectiveness of the treatments can be difficult to interpret. Other complex analyses such as analysis of treatment algorithms or repeated measures modeling can be pursued to more fully describe the treatment response across phases.

The Alzheimer's disease trial is actually a hybrid between an effectiveness study and an efficacy study, because the primary comparison among active treatments is based on an efficacy assessment, the CGI at 12 weeks. The initial treatment groups will be compared for noninferiority on the CGI score at 12 weeks, regardless of subsequent randomizations. It is understood that in the trial many patients may be switched between medications as early as 2 weeks after initiating treatment. As an effectiveness trial, the CATIE Alzheimer's disease protocol was

not designed to evaluate whether or not an atypical antipsychotic has the best immediate response but instead to see whether or not the active treatments are essentially equivalent in terms of efficacy after 12 weeks, regardless of which antipsychotic is begun first.

The limited interpretations supported by the first randomization analysis make it an unattractive method for evaluating pure efficacy or safety outcomes. However, in addition to the Alzheimer's disease protocol evaluation of short-term efficacy, it is an appropriate analysis for some long-term effectiveness outcomes, particularly cost-effectiveness. Because some CATIE treatments are significantly more expensive than others, an important health policy outcome of the CATIE trials would be to determine any long-term cost differences between the treatments chosen as first line medications.

Additional analyses of the initial treatment groups may be stratified according to whether or not patients were re-randomized, and if so, to which treatment, in order to examine the heterogeneity due to different courses of treatment. But if the research question focuses primarily on the course of treatment, then evaluating treatment algorithms may be of more interest.

Analysis of Treatment Algorithms. An alternative way of viewing outcome-driven re-randomizations is to consider all treatments a patient ultimately receives as a treatment algorithm. In this framework, the question of interest is whether or not there are treatment sequences that lead to a superior or inferior outcome at a particular set time point. Similar to analyses based on the first randomization, the patient's state at a particular time point (i.e., the end of the trial) is the main focus, and the patient's state at the time of treatment re-randomization is of less interest. For the Alzheimer's disease trial, we are particularly interested in investigating the effect of treatment algorithms at set time points, and the study design therefore specifies that a full set of all assessments is taken at these time points (12, 24, and 36 months) rather than at the time of phase switch.

An analysis of treatment algorithms is essentially identical to an analysis based on the first treatment, with the modification to account for the treatment-phase combinations, so that all sequences of treatments can be compared, rather than just the treatments from Phase 1. The set of treatments a patient is ultimately randomized to in a series of phases can essentially be viewed as an algorithm of treatments that was completely specified by one randomized assignment at the beginning of the trial. This series of treatments can form the basis for statistical comparisons of the outcome measure.

Analyses of treatment algorithms have some important disadvantages. Because there are many possible treatment algorithms obtained by combinations of treatments across phases, the sample size for each algorithm is fairly low; therefore, the power for identifying treatment algorithm differences is low. And, although analyses of algorithms do account for the ordering of treatments across phases, they make no use of the temporal relationship between a patient's treatments and the outcome. Most important, in CATIE, the number of treatments that a patient receives is driven by the patient's response to the previous treatment. The fact that a patient was re-randomized is actually part of the patient's outcome, and thus the treatment algorithm is part treatment and part response. Therefore, although an analysis of treatment algorithms may be of interest for exploratory purposes, interpretation of results can be very difficult.

Repeated Measures Modeling. Longitudinal data analysis is a strategy that accounts for all treatments received by the patient and incorporates the temporal relationship between the treatments and the outcomes. In addition, rather than focusing on the outcome at one point in time, it models all of the repeated measurements that were collected for that outcome over the course of the study. Longitudinal models can be used to address a variety of research questions regarding the change in an outcome over time. One question of interest is whether or not any treatments result in a superior or inferior change in a particular outcome over time, combining information from all randomized phases. Another question may be whether or not changes in an outcome for the second phase are different from changes seen in the first phase. These types of questions are applicable to efficacy outcomes and numerous secondary assessment instruments.

The correlation of the repeated measurements within the same person are taken into account by fitting a repeated measures longitudinal model, via either mixed linear models or generalized estimating equations (Diggle et al. 1994). An appropriate model would include fixed effects for baseline value, clinical site, initial treatment, and other important covariates, plus time trends, a time-varying indicator of treatment phase, a time-varying treatment effect that accounts for re-randomizations, and time by treatment interactions. Specific research hypotheses would be addressed by the model through contrasts that compare slopes or other measures that are interpretable as the change in the outcome over time. This type of repeated measures model has substantial flexibility in the type of questions that can be addressed, through specifying complex contrasts and adding additional covariates. Results may be difficult to interpret, however, because of the complex nature of the model and possible time lags between treatment change and evidence of response. Guidelines for handling the multiple treatment comparisons described in

the section "Multiple Comparisons" will be applied. Because there will be early study dropouts, the robustness of the models to missing data can be evaluated by comparing results from several models that address missing data in different ways.

For CATIE, complex repeated measures models will be largely exploratory and hypothesis generating and will be used to provide further information about the impact of the current treatment on study outcomes beyond that available from analyses based on each phase separately or based on the first randomized treatment.

Implications of Outcome-Driven Re-Randomization on Data Base Design

The outcome-driven re-randomizations are an integral component of the CATIE study designs. They allow the investigators to address many research questions in an effectiveness trial in accordance with the real-world setting. As we have seen, the statistical analysis of data from CATIE is greatly affected by this multiphase design. However, statistical analysis is not the only affected component of the clinical trial. Multiple outcome-driven re-randomizations add substantial complications to the automated randomization system and the clinical data base design, including data from external vendors such as laboratories, electrocardiograms, and hair samples. These extra complications affect the work of study coordinators, clinical monitors, and data managers during the trial.

Re-randomization occurs whenever the investigator chooses, so subsequent re-randomizations are not set at planned time intervals. Yet, the visits within the study period are set, as in standard clinical trials. Therefore, tracking the treatment phase for the patient is a separate task from tracking the visit number for the patient. Knowing a patient's treatment phase at any given visit is important for patient tracking and patient safety during the trial as well as for statistical analysis.

Because close tracking of patients is important, we employed electronic data capture (EDC) rather than a traditional paper system. Personnel at each study site enter visit data into EDC software loaded on their computer. The data are transferred online to a central data base, and subsequently to the monitors and data managers at Quintiles, who have the ability to view all case report form (CRF) data online and issue queries within hours of data entry. Individual patient data are also viewable from a Web portal by project administrators and principal investigators. Randomization and medication dispensing are handled by the Quintiles interactive voice-recorded (IVR) telephone system. The clinical data entry and IVR data bases are linked. Patient identification information is sent from the clinical data management system to the IVR system prior to the first randomization, and medication dispensing information (date, bottle identification numbers, and treatment phase) is sent from the IVR system to the CRF data base. Because the medication dispensing information is downloaded to the data entry system, everyone with access to the data base, either directly or through the Web portal, can view the dispensing and CRF data simultaneously. This system allows study personnel to identify what phase a patient is in and permits queries for quick resolution of any inconsistencies between phase status recorded by the site on the CRF pages and the medication dispensing information from the IVR system. Other advantages of the electronic data capture system include a large number of preprogrammed instant edit checks that prevent many data entry errors, and an up-to-date clinical data base used by Quintiles statisticians to provide patient disposition and safety information in quarterly Data and Safety Monitoring Board reports to the NIMH.

Our solution to tracking both study visits and treatment phases within the clinical CRF data base itself was to track patients by visit number as in standard clinical trials. The treatment phases are tracked by a CRF page indicating phase status at each scheduled or unscheduled visit, combined with a special CRF (Alzheimer's disease) or packet of CRFs (schizophrenia) completed at each visit in which a phase switch occurs. These "end of phase" visits are recorded and tracked in the system by a special visit number for each phase. Two other data base design options for the schizophrenia trial, which calls for all outcomes to be measured at the end of each phase, were to create a separate data base for each treatment phase or to have the electronic data collection software create visit structures dynamically for all patients as they progressed through subsequent phases of the trial. However, these options were deemed overly complex and cumbersome.

Summary

This article has addressed some of the fundamental statistical issues involved with design and analysis planning for the CATIE trials. We have presented a framework for addressing multiple comparisons, given motivation for the use of time until Phase 1 treatment discontinuation as the primary overall effectiveness outcome, described the advantages and limitations of several strategies for analyzing the data collected across multiple outcome-driven re-randomized phases, and identified what research questions each strategy addresses. In addition, we have described how we designed the CATIE electronic data collection system to reliably track both visit and phase information.

References

Bauer, P. Multiple testing in clinical trials. *Statistics in Medicine,* 10:871–890, 1991.

Bombardier, C., and Maetzel, A. Pharmacoeconomic evaluation of new treatments: Efficacy versus effectiveness studies? *Annals of the Rheumatic Diseases,* 58(Suppl 1):82–85, 1999.

Diggle, P.J.; Liang, K.; and Zeger, S.L. *Analysis of Longitudinal Data.* New York, NY: Oxford University Press, 1994.

Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika,* 75:800–802, 1988.

Kay, S.R.; Fiszbein, A.; and Opler, L.A. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin,* 13(2):261–276, 1987.

Koch, G.G. Discussion for alpha calculus in clinical trials: Considerations and commentary for the new millennium. *Statistics in Medicine,* 19:781–784, 2000.

Koch, G.G.; Sen, P.K.; and Amara, I.A. Logrank scores, statistics, and tests. In: Johnson, N.L., and Kotz, S. *Encyclopedia of Statistical Sciences.* Vol. 5. New York, NY: John Wiley and Sons, 1985. pp. 136–142.

Lehmacher, W.; Wassmer, G.; and Reitmeir, P. Procedures for two-sample comparisons with multiple endpoints controlling the experiment wise error rate. *Biometrics,* 47:511–521, 1991.

Moye, L.A. Alpha calculus in clinical trials: Considerations and commentary for the new millennium. *Statistics in Medicine,* 19:767–779, 2000.

O'Brien, P.C. Procedures for comparing samples with multiple endpoints. *Biometrics,* 40:1079–1087, 1984.

Revicki, D.A., and Frank, L. Pharmacoeconomic evaluation in the real world: Effectiveness versus efficacy studies. *PharmacoEconomics,* 15(5):423–434, 1999.

Schneider, L.S.; Olin, J.T.; Doody, R.S.; Clarck, C.M.; Morris, J.C.; Reisberg, B.; Schmitt, F.A.; Frundman, M.; Thomas, R.G.; and Ferris, S.H. Validity and reliability of the Alzheimer's Disease Cooperative Study-Clinical Global Impression of Change. The Alzheimer's Disease Cooperative Study. *Alzheimer Disease and Associated Disorders,* 11(Suppl 2):S22–32, 1997.

Schneider, L.S.; Tariot, P.N.; Lyketsos, C.G.; Dagerman, K.S.; Davis, K.L.; Davis, S.; Hsiao, J.K.; Jeste, D.V.; Katz, I.R.; Olin, J.T.; Pollock, B.G.; Rabins, P.V.; Rosenheck, R.A.; Small, G.W.; Lebowitz, B.; and Lieberman, J.A. National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE): Alzheimer disease trial methodology. *American Journal of Geriatric Psychiatry,* 9(4):346–360, 2001.

Stroup, T.S.; McEvoy, J.P.; Swartz, M.S.; Byerly, M.J.; Glick, I.D.; Canive, J.M.; McGee, M.F.; Simpson, G.M.; Stevens, M.C.; and Lieberman, J.A. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: Schizophrenia trial design and protocol development. *Schizophrenia Bulletin,* 29(1):15–31, 2003.

Swartz, M.S.; Perkins, D.O.; Stroup, T.S.; McEvoy, J.P.; Nieri, J.M.; and Haak, D.C. Assessing clinical and functional outcomes in the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) schizophrenia trial. *Schizophrenia Bulletin,* 29(1):33–43, 2003.

Westfall, P.H.; Tobias, R.D.; Wolfinger, R.D.; and Hochberg, Y. *Multiple Comparisons and Multiple Tests Using the SAS System.* Cary, NC: SAS Institute, 1999.

Acknowledgments

This article was based on results from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project, which was supported with Federal funds from the National Institute of Mental Health (NIMH) under contract NO1 MH90001. The aim of this project is to examine the comparative effectiveness of antipsychotic drugs in conditions for which their use is clinically indicated, including schizophrenia and Alzheimer's disease. The project was carried out by principal investigators from the University of North Carolina, Duke University, the University of Southern California, the University of Rochester, and Yale University, in association with Quintiles, Inc., and the program staff of the Division of Interventions and Services Research of the NIMH and investigators from 84 sites in the United States. AstraZeneca Pharmaceuticals LP, Bristol-Myers Squibb Company, Forest Pharmaceuticals, Inc., Janssen Pharmaceutical Products, L.P., Eli Lilly and Company, Otsuka Pharmaceutical Co., Ltd., Pfizer Inc., and Zenith Goldline Pharmaceuticals, Inc., provided medications for the studies.

The Authors

Sonia M. Davis, DrPH, is Director of Biostatistics, Quintiles, Inc., Morrisville, NC; and Adjunct Assistant Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC. Gary G. Koch, Ph.D., is Professor, and C.E. Davis, Ph.D., is Chair, Department of Biostatistics, University of North Carolina, Chapel Hill. Lisa M. LaVange, Ph.D., is Vice President of Biostatistics and Data Management, Inspire Pharmaceuticals, Durham, NC; and Adjunct Professor, Department of Biostatistics, University of North Carolina, Chapel Hill. During the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) protocol development, Dr. LaVange was Vice President of Biostatistics, Quintiles, Inc.