Mood Stabilizers and the Evolution of Maintenance Study Designs in Bipolar I Disorder

Joseph R. Calabrese, M.D., and Daniel J. Rapport, M.D.

The designs employed in bipolar maintenance studies have evolved greatly over the last 28 years. Consequently, there has been minimal consensus set for methods used to demonstrate the ability of any new putative mood stabilizers to prevent relapse and recurrence in bipolar disorder. The methods that have evolved the most include enrollment procedures, randomization schemes, use of outcome measures, statistical analyses, and country-specific commercial and regulatory issues. This article contrasts the various methods employed in first- and second-generation placebo-controlled bipolar I maintenance studies. This article also explores the advantages and disadvantages associated with various designs. *(J Clin Psychiatry 1999;60[suppl 5]:5–13)*

During the last 25 years, there has been one placebo-controlled maintenance study conducted about bipolar I disorder.[1] A critical review of other bipolar maintenance studies conducted during this period indicates that study methodologies have evolved substantially; methods employed have differed depending upon the decade in which the study was designed and whether the results were intended for use for submission to a regulatory agency. For those pivotal data sets intended for submission to a regulatory agency, country-specific requirements have also had a impact on study design.[2–9]

Over the last 10 years, efforts to develop new drugs for bipolar disorder have primarily focused on bipolar type I presentations of the illness. These studies have used double-blind placebo-controlled trials to demonstrate the antimanic efficacy of agents in hospitalized patients experiencing the manic phase of the disorder. Once an initial acute mania study has been replicated, maintenance studies designed to evaluate the ability of a putative mood stabilizer to prevent relapse and recurrence usually follow. Most recent studies have randomly assigned subjects to parallel groups and have typically performed not only responder analyses (percentage of patients completing trial who experience 50% improvement), but also survival analyses. The survival analyses are normally carried out on intent-to-treat samples (i.e., on all patients who begin treatment regardless of whether or not they complete the trial). More recent studies have been designed to investigate the acute and prophylactic efficacy of putative mood stabilizers in the various phases (depressed, mixed, hypomanic, manic) and subtypes of the disorder (bipolar I, bipolar II, rapid cycling, and comorbid with substance abuse). Separate studies either are ongoing or have been conducted in each of these various patient populations.

STUDY ENROLLMENT

The first generation of bipolar disorder maintenance studies tended to enroll patients with "classic" (euphoric) forms of manic depression. It is likely that the evolving nature of our psychiatric nomenclature has had an impact on the design and method of maintenance studies, and particularly on enrollment practices. The Research Diagnostic Criteria (RDC),[10] the second version of the *Diagnostic and Statistical Manual of Mental Disorders*[11] published by the American Psychiatric Association, and the clinical standards most commonly employed at different times placed importance on the role of hospitalization in confirming the diagnosis of mania. The removal of this requirement from the DSM-III in 1980 probably resulted in the inclusion of less impaired patients in more recent maintenance studies of bipolar disorder.[12] In the RDC, patients with mood-incongruent psychotic symptoms were included in the category of the "mainly affective variant of schizoaffective disorder," not bipolar disorder. As a result, patients with mania enrolled in early bipolar maintenance studies tended to have a disproportionate representation of

From the Mood Disorders Program, Department of Psychiatry, Case Western Reserve University School of Medicine, Cleveland, Ohio.

Supported by grants RO1-MH 50165 and MH 50165-S1 from the National Institute of Mental Health, Bethesda, Md.

Presented at the symposium "The Use of Mood Stabilizers for Treating Psychiatric Disorders," which was held March 3, 1998, Amelia Island, Fla., and supported by an unrestricted educational grant from Eli Lilly and Company.

The authors thank Jules Angst, M.D., Charles L. Bowden, M.D., and Robert M. Post, M.D., for their review of this article.

Reprint requests to: Joseph R. Calabrese, M.D., Professor of Psychiatry, Director, Mood Disorders Program, 11400 Euclid Ave., Suite #200, Cleveland, OH 44106.

classic euphoric mania. If they were psychotic, the content of their hallucinations and delusions was mood-congruent. This patient subtype is now recognized as being one of the more lithium-responsive variants of the disorder.[13]

Recently, lithium response rates in open naturalistic studies have been noted to decrease markedly from those rates reported in the early placebo-controlled maintenance studies.[14] The early maintenance studies reported mean relapse rates of 23% for mania and 21% for depression.[15] These rates are inconsistent with recently published naturalistic data, which reveal mean overall lithium relapse rates of 77% over 5 years of prospective follow-up.[14] It is likely that these deteriorating lithium response rates are due (at least in part) to the changing study patient populations. Again, this change may be secondary to the changes in the evolving nomenclature, which now includes atypical variants such as abnormal mood states with incongruent psychotic symptoms, mixed or dysphoric mania, rapid cycling, and presentations that are comorbid with alcohol/drug abuse.

Enrollment in the early maintenance studies may have selected preferentially for the more lithium-responsive bipolar patient. Changes in the most recent version of the DSM have included all of the above atypical variants of bipolar disorder,[16] and less is known about these variants of illness since they have not been extensively studied. In particular, the degree of subtype response to placebo is unclear. New studies should include these atypical subtypes because of their prevalence. When considered as a group, these atypical variants of bipolar disorder are more prevalent than the classic forms of the illness. New studies should include these more common patterns of presentation to improve generalizability to the practice of routine clinical care.

It is now recognized, moreover, that patients with less severe illnesses are less suitable for clinical treatment trials. Including these patients decreases the power of studies by decreasing the amount of potential difference between results from the active compound and results from the placebo.[1,17,18] This issue was less relevant to first-generation maintenance studies because the selection bias at the time favored enrollment of more severely ill patients. Early studies were more likely to enroll patients with more severe forms of classic manic depression because fewer effective treatments were available. More recently, patients with severe illness have been difficult to enroll because the number of effective treatment alternatives has increased. Some investigators also consider the enrollment of the severely ill in outpatient studies to be unethical.

Patients with severe illness are now usually taking 3 to 5 different medications[19] and are less interested or less able to enroll in rigidly designed studies employing complete drug washouts followed by monotherapy treatment arms or the use of placebo. In addition, health care reform has led to a decrease in the utilization of hospitalization for patients with bipolar disorder. As a result, investigators have become reluctant or unable to enroll more seriously ill patients in bipolar disorder maintenance studies, which has resulted in a selection bias toward a less ill and therefore less representative study population.

The primary advantage associated with entry criteria that facilitate the enrollment of more severely ill patients is that the power of the study to detect clinically and statistically significant differences between the putative mood stabilizer and placebo improves markedly. The disadvantages associated with the enrollment of more severely ill subjects are that dropouts due to poor compliance, adverse events, lack of efficacy, withdrawn consent, and protocol violations all increase. Accordingly, randomization rates into the blinded phase of maintenance studies that include the broader range of patients are lower, ranging from 33% to 66%. Survival analyses and intent-to-treat analyses that carry last observations forward, however, make valuable use of the study endpoints so that the data are not wasted.

Twenty-five years ago, bipolar disorder maintenance research tended to evaluate smaller cohorts of study subjects, ranging from 5 to 40 patients per study arm.[2–9] The one notable exception to the above is the study conducted by Prien and colleagues,[8,9] which enrolled 101 patients who took lithium and 104 who took a placebo. In general, the mean enrollment size of the lithium arms in these earlier studies was 31 (range, 7–101) and of the placebo arms, 28 (range, 8–104). More recently, issues associated with a drug's remaining patent life and time to marketplace have become compelling reasons for industry-sponsored drug development efforts, which have led to large-scale multicenter studies with as many as 50 to 60 sites. These studies generate data more quickly, but interrater reliability is substantially more difficult to monitor. Remaining patent life and time-to-marketplace issues are particular concerns for antiepileptic drugs under development for bipolar disorder, because many years of the patent lives of these drugs have already been used during the development of compounds for use in the treatment of epilepsy. On the other hand, drug development is often streamlined when basic phase I (pharmacokinetic) studies have already been completed.

RANDOMIZATION SCHEMES

The first-generation bipolar I disorder placebo-controlled maintenance research studies were conducted between 1960 and 1970 and published between 1970 and 1976. These pioneering studies most commonly employed double-blind crossover designs in which enriched samples of lithium responders were crossed over to placebo (discontinuation designs). A positive response to lithium was either determined retrospectively through open unblinded assessments or prospectively through blinded assessments. With one exception,[8,9] these early studies usually

did not standardize the index episode (the polarity of the episode that brought the patient into the study) in order to enroll patients either as manic or depressed.

Crossover Study Design

In a typical crossover study, each patient's response under treatment A is compared with his or her response under treatment B so that the influence of unique patient characteristics that determine the degree of response can be subtracted out of the treatment comparison. This procedure does not remove biological variation within an individual, but if these variations are small, a crossover design can provide the same statistical accuracy as a larger parallel study, but with a smaller sample of patients. In general, crossover designs are most appropriate in the study of the treatment of a stable disease process.[20]

Potential factors that can alter clinical outcome when crossover designs are employed include 2 types of order effects. Carryover effects occur when the therapeutic effects of the first treatment persist during the administration of the second, and period effects occur when the disease increases or decreases in severity during the period of investigation. These order effects can complicate data interpretation and weaken the scientific merit of the design.

Other factors that can influence the crossover design include treatment sequencing and patient assignment. If all patients receive treatment according to the same fixed sequence (i.e., A followed by B), comparisons must be based on evidence to suggest that the effects of the second treatment (B) after the first (A) do not differ from the effects B would have if it were given first. Recent data suggest this may not be the case for bipolar disorder when lithium is used for treatment. These data have suggested that the withdrawal of lithium, and in particular its rapid withdrawal, significantly increases the risk of early recurrence to a rate that exceeds that predicted by the natural course of the untreated illness.[21–23] Therefore, it is possible that previously used crossover designs artificially decreased placebo response rates through lithium discontinuation-induced relapse.

Premature drug discontinuations have also been observed to have special relevance when they occur during crossover studies. Although premature study drug discontinuations are a serious problem for any drug trial, the impact of premature dropouts in a crossover study is exaggerated because each remaining patient then contributes a larger proportion of the study data. Dropout rates can be high in crossover studies since each patient must receive at least 2 treatments to provide complete data. In studies of illnesses such as bipolar disorder, where the prevalence of poor compliance is substantial, a high dropout rate weakens a crossover study. The initial sample size should be sufficiently large to compensate for this effect. Studies designed to have short experimental periods to evaluate acute changes in clinical status are best suited for crossover designs. The experimental medications under study should not be accompanied by carryover treatment effects, and the illness under study should exhibit a stable nonperiodic disease process. Ideally, treatments used should have few side effects severe enough to result in drug discontinuation.

Parallel Study Design

Studies that employ a parallel design are less dependent on assumptions about the disease process and frequently produce a lower dropout rate because each patient is exposed to only 1 treatment. One of the first-generation studies used prospective random assignment to parallel groups receiving lithium or placebo.[4] Although double-blind study medication was continued throughout the entire drug trial, the psychiatrist in charge of the case could also prescribe treatment other than lithium as deemed clinically appropriate. The psychiatrist in charge remained blinded and assessed improvement, while another psychiatrist blindly monitored lithium levels and adjusted doses as necessary per protocol. This particular study design most resembles those employed by second-generation research. Second-generation studies, however, have typically used open stabilization periods rather than immediately proceeding with randomization to the putative mood stabilizer and placebo.

Most recently, a study randomly assigned subjects to parallel groups following open stabilization with medications chosen completely at the discretion of the investigator.[1] Given this development, the field appears to have cycled from an extreme of randomizing a homogeneous cohort of lithium monotherapy responders to the other extreme of allowing any medications for use during the open stabilization phase of the trial. As a result of this more recent study design, the study patients may become unmanageably heterogeneous. An alternative to this method might be limiting the use of psychotropic medications during the open stabilization period of the maintenance study to only the putative mood stabilizer and lithium (if lithium is being used to gauge study reliability) during the blinded phase.

Another recent series of ongoing studies has employed "enriched" designs in which responders to a putative mood stabilizer in monotherapy are again randomly assigned to continue treatment with the new drug, placebo, or lithium in parallel. This design is considered to be enriched because it randomizes a homogeneous cohort of patients that have responded to a defined regimen of medication. This method benefits from the advantages associated with enriched designs while discarding the crossover methodology in favor of a design considered by some regulatory agencies to be more suitable for disorders accompanied by frequent periodic relapse.

Unless enrollment procedures standardize for the pattern of the index episode (cycling from mania into depres-

sion, and then euthymia versus depression into mania and then euthymia) at the time of study entry, crossover designs will carry the risk of altering outcome as an artifact of spontaneous cycling. A patient who presents toward the end of an episode of mania could conceivably have a spontaneous remission falsely attributed to a study medication, and have a relapse falsely attributed to placebo or to the next study medication.

Neither design (crossover nor parallel), however, addresses the issue of discontinuation-induced relapses. If a patient is stabilized on one medication during the open stabilization phase of a maintenance study and that medication is abruptly discontinued following randomization, as is the case with both crossover and enriched parallel designs, patients who are randomly assigned to a placebo group can experience an iatrogenic-increased risk of relapse. This problem can be managed effectively by slowing the taper and starting the time-to-relapse survival analysis when blinding and the wean to study medication begin. The latter is required because a significant number of patients, especially patients with rapid cycling, will relapse quickly and will therefore be unavailable to the survival analyses if the analyses do not begin until the wean is complete.[22]

OUTCOME MEASURES

Most early lithium prophylaxis studies were designed to compare the efficacy of lithium with that of a placebo by general indices of outcome during the study period,[4] the number of manic and depressive episodes,[7,24] or the probability of manic or depressive episodes.[8,25–27] These early studies did not rate symptom severity with mood-specific rating scales such as the Hamilton Rating Scale for Depression[28] or the Young Mania Rating Scale.[29] Symptom severity is now routinely quantified with depression and mania rating scales, which then are used as secondary outcome measures.

Depression symptom severity rating scales were originally standardized using patients with unipolar depression only. They were typically designed to make cross-sectional assessments that evaluate a patient's symptoms of depression during the 7-day period prior to the completion of the instrument. In so doing, they run the risk of missing cycle activity previously experienced by the patient. Various measures of prospective daily-life charting might be a more suitable method to count episodes or to attempt to quantify the cumulative duration of (euthymic) periods.[30]

Over the last 10 to 15 years, the criteria used to define the time to relapse and recurrence into an episode have been more strictly defined. Thinking of the time to relapse as a rigidly defined episode, however, may pose new problems since specificity is sometimes achieved at the expense of sensitivity. For example, relapse into mania requiring hospitalization, which was the most common outcome measure in first-generation research, might be too rigid. This measure might not be sensitive enough to uncover lesser degrees of difference between (less potent) active compounds and a placebo. The time from the beginning of randomization to the first prescription needed of a psychotropic medication might be a more sensitive indicator of clinical efficacy (as employed by Coppen and colleagues[4]). Many studies now require a minimum rating score on a symptom severity scale for a randomized subject to reach study endpoint. This method, though, can miss other indications of medication inferiority such as patient satisfaction, quality of life, tolerability, convenience, and so on.

STATISTICAL ANALYSES

The first generation of bipolar disorder maintenance studies used responder analyses on observed data or completer data with little or no distinction between primary and secondary outcome measures. For example, rather than arbitrarily relying on only 1 primary outcome measure as the primary determinant of efficacy, Coppen[4] relied on several measures, including a 7-point global rating scale, percentage of time spent as an inpatient, percentage of time spent as an outpatient, other treatments prescribed during the trial, and premature study discontinuations.

Early analyses usually addressed the proportion of patients who experienced a relapse or recurrence. However, those analyses are difficult to generalize in maintenance trials because they do not consider the length of time that a patient remained well before relapsing. A relapse at 1 month and one at 12 months of a 1-year trial would incorrectly reflect similar degrees of efficacy. Another liability associated with this analytical procedure is that patients withdrawing prematurely without experiencing a relapse are either ignored or analyzed incorrectly. For example, in a study using relapse into mania as its primary outcome measure, important clinical data such as relapses into hypomania or depression, premature discontinuations from the study due to intolerable side effects, or dropouts due to poor compliance would not be incorporated into the primary analysis.

A statistical method known as a survival analysis commonly employed in cancer trials uses as its measure of outcome the time to relapse or recurrence (as defined in the protocol). One of the more commonly used survival analyses is called the product-limit method. This method extends the idea underlying a life-table analysis. The life-table procedure calculates the probabilities of remaining well during a study period and the median time in remission for each treatment. It was developed by Kaplan and Meier.[31] In the traditional life-table analysis, relapses are grouped into intervals of time. When the actual times to relapse are available, the life-table method wastes data such as the cumulative time spent well. The product-limit

method makes use of the exact time to relapse to compute cumulative probabilities of remaining well. Patients dropping out of the study before its completion for reasons other than a relapse or recurrence are also analyzed until the time of dropout and then are withdrawn or censored from the analysis.

A striking example of the importance of this statistical method comes from the Shapiro reanalysis[32] of the National Institute of Mental Health (NIMH) Collaborative Study bipolar patient study.[33] In the NIMH study, responder analyses were carried out on patients randomized to lithium, imipramine, or the combination of both. Using the original responder analysis for patients presenting with an index episode of depression, no differences were observed between lithium, imipramine, or the combination (li = imi = li/imi). The final outcome changed with the reanalysis, which employed survival techniques. In patients with depressive index episodes, the combination of lithium plus imipramine was significantly superior to imipramine alone or lithium alone (li/imi > imi = li). This finding differs from the original analysis, which found the combination to be no different from the other 2 monotherapies. Survival analyses are now viewed as powerful techniques that are capable of detecting effects not otherwise observed with other analytical methods.

Over the last 10 years, survival analyses have become the standard in analyzing bipolar maintenance data. However, these analysis methods were not employed during the first generation of bipolar disorder maintenance studies. Survival analyses usually require the prospective selection of a primary outcome measure. It is now well-recognized that there are many ways of defining outcome. Patients define outcome by emphasizing convenience and safety. Managed care companies define outcome by emphasizing direct and indirect costs. Investigators and pharmaceutical companies wish to define outcome based on measures most likely to separate differences in efficacy from the active compound and placebo. Power calculations conducted for single primary outcomes result in studies that are only powered to detect differences in the primary outcome when many other measures of clinical improvement are valuable and relevant.

Recent methods have erroneously placed crucial importance on the distinction between primary and secondary outcome measures at a time when health care reform is moving toward the interpretation of data from many perspectives. These different perspectives would include morbidity, efficacy, mortality, safety and tolerability, convenience, and direct and indirect costs. Using alternative perspectives minimizes the likelihood of study failure, and will likely become more important as health care becomes increasingly managed by government agencies and insurance/managed care companies.

Placebo response rates in bipolar disorder vary substantially with the specific population of patients studied, the definition of the primary outcome measure, and the environment in which the patients are studied. For example, the placebo response rate in a cohort of lithium-nonresponsive, acutely manic hospitalized patients has been reported to be as low as 10.5%.[34] In contrast, placebo response rates for patients similar in every respect, except for being unselected for prior treatment, result in markedly higher rates of 21%,[35] 24%,[36] and 45%[37]; all of these studies have defined response as a 50% reduction in the baseline mania rating. The highest of these placebo response rates occurred in a double-blind augmentation study in which Müller-Oerlinghausen and colleagues[37] found the addition of valproate to haloperidol and/or perazine to be superior to placebo.

The mean overall placebo response rate in the early maintenance studies of patients with bipolar I disorder was low—only 21%.[38] However, this low placebo response rate was based on responder analyses carried out on observed or completer data rather than survival analyses performed on last observations carried forward on the intent-to-treat sample of study patients. Recently, Bowden and colleagues[1] noted an apparently high placebo response rate of 62% in a maintenance study of patients with bipolar I disorder, but this study performed survival analyses on the intent-to-treat sample (patients who received at least one dose of study medication). This survival analysis was further weakened by the use of the traditional life-table method rather than the product-limit method, which was required because study visits were carried out at fixed intervals. Daily self-assessments could have been obtained through the use of the prospective daily life-charting method had it been employed.[30] When actual times to an event are available, the product-limit method saves data and improves the power of a survival analysis. This is particularly relevant to the second half of a maintenance study when less frequent assessments are fixed at only monthly intervals.

When a responder analysis is carried out on an intent-to-treat sample, the percentage of patients responding to placebo is artificially inflated by premature study terminations, and in particular, those dropouts occurring from withdrawn consents, protocol violations, and so on. As these early dropouts tend to disproportionately occur in those patients on placebo, fewer patients on placebo are left to be at risk for relapse, decreasing the power of a study and representing a disadvantage associated with survival analyses that employ time to relapse or recurrence as the primary outcome measure. For example, of the 94 patients assigned to placebo in the Bowden study,[1] 70 (75%) dropped out prematurely: 21 (22%) due to relapse into mania requiring hospitalization or a mania rating score of 16, 15 (16%) due to depression, 11 (12%) due to intolerance/poor compliance, and 23 (25%) because of other reasons (including protocol violations, withdrawn consents, losses to follow-up, and intercurrent illness).

Only 38% relapsed into episodes of mania or depression meeting strict criteria, which leaves the false impression that 62% overall responded to placebo.

A more meaningful analysis might employ time to any premature termination or time to any clinically necessary treatment for an emerging episode rather than time to relapse. Time to the initiation of a new medication was successfully employed in one early bipolar maintenance study.[4] Utilizing time to treatment as the event, rather than time to relapse into a full episode of mania requiring hospitalization, also has the advantage of more closely adhering to current recommendations, as early intervention for subsyndromal relapse has become the standard of clinical practice in the treatment of bipolar disorder, and especially so for relapses into hypomania.[39]

COMMERCIAL ISSUES

Until Ballenger and Post[40] and Okuma and colleagues[41] performed the first controlled trials of anticonvulsants as treatments for bipolar disorder, it was not widely recognized that antiepileptic drugs possessed mood-stabilizing properties. Since then, the pharmaceutical industry has gradually developed an appreciation for the magnitude of the unmet need in bipolar disorder treatment and the potential commercial opportunity provided by that need. The pharmaceutical industry's pipeline of antiepileptic drugs is extensive, and more companies are realizing that all antiepileptic drugs should be considered to be putative mood stabilizers and developed as such. At the same time, the drugs' efficacy in the treatment of epilepsy is explored; this strategy has resulted in some economies of scale in drug development between epileptology and the psychopharmacology of mood stabilizers.

Although most of the development of mood stabilizers has focused on the class of antiepileptic drugs, there have also been substantial resources invested in the development of atypical antipsychotic medications and on selective serotonin reuptake inhibitors. Despite recent interest in the pharmacotherapy of bipolar disorder, it is widely recognized that the ideal mood stabilizer does not exist. Such an agent would be effective in each phase of the illness and be used as monotherapy.

The ideal mood stabilizer would be

1. effective as monotherapy for the acute treatment of depression and for the prevention of relapse and recurrence,
2. effective as monotherapy for the acute treatment of mixed states and for the prevention of relapse and recurrence,
3. effective as monotherapy for the acute treatment of mania and hypomania and for the prevention of relapse and recurrence,
4. acceptable and convenient as monotherapy for the consumer with enhanced overall patient satisfaction and quality of life, and
5. successful in reducing both psychiatric resource utilization and general medical utilization as reflected by decreased frequency of office visits with the family physician.

The pharmacologic properties of such a compound would result in

1. acute onset of effect within hours,
2. low rates of drug discontinuation secondary to side effects,
3. no side effects requiring additional treatment, and
4. safety in overdose.

Considering our current pharmacologic armamentarium and the above expectations, it is evident that much opportunity exists to reduce the pain and suffering associated with bipolar disorder.

Another unique commercial issue that affects drug development in bipolar disorder is the extent to which the pharmaceutical industry has been able to rely on expert psychiatric consultation from the community of bipolar investigators. The availability of psychiatric consultation has been particularly relevant to the development of antiepileptics for use in the treatment of bipolar disorder since nonpsychiatrists have typically been employed by the pharmaceutical industry to develop these compounds for use in treating epilepsy. In the past, there was insufficient agreement among panels of expert psychiatric consultants on issues relevant to study design, and as a result, advisory panels had difficulty forming a consensus of opinion, which tended to diminish the value associated with the use of expert consultants. More recently, the degree to which a pharmaceutical company engages in discussions with expert consultants appears to have a more clearly beneficial effect on the overall process of drug development in bipolar disorder.

REGULATORY ISSUES

When examining the various requirements imposed by regulatory agencies around the world, it is clear that significant differences in requirements and preferred methodology exist. Although all regulatory agencies require a thorough exploration of basic pharmacokinetics, safety, and efficacy, the extent to which various agencies require concurrent pivotal data (intended for submission to the Food and Drug Administration [FDA]) for acute and prophylactic treatment as well as for both manic and depressive phases of the illness varies substantially. Recent experience with the U.S. FDA would suggest that 3 factors are required for acute indications in the United States: pharmacokinetic information, 2 acute pivotal data sets, and safety in both acute and longitudinal settings. European

regulatory agencies require maintenance efficacy data and quite possibly continuation phase/intermediate duration (12 weeks) data as well, in addition to all of the information required by the U.S. FDA.

Because most of the recently studied putative mood stabilizers were first developed as antiepileptic drugs, it is interesting to contrast and compare the methods by which the same drug is developed for the 2 different types of illnesses. When undergoing development for use in the treatment of epilepsy, most drugs initially have their efficacy evaluated through double-blind, placebo-controlled, add-on/augmentation trials. Not infrequently, putative antiepileptic drugs are blindly added to regimens of 2 or 3 other concomitantly prescribed antiepileptic drugs that have resulted in unsatisfactory partial responses. The apparent rationale behind this practice is that epilepsy is a life-threatening illness with a low placebo response rate (15%–20%) and that it would be unwise and unsafe to stop currently prescribed medication in order to proceed with a placebo-controlled monotherapy drug trial. In contrast, placebo response rates in bipolar disorder vary substantially. The markedly higher placebo response rates in patients with bipolar disorder as compared with those among patients with epilepsy explains why placebo-controlled add-on maintenance trials have not been previously conducted in bipolar disorder.

CONCLUSION

A critical evaluation of the various methods employed in bipolar disorder maintenance methodology over the last 30 to 40 years leads to the conclusion that each of the various designs has its own strengths and weaknesses. The primary methodological advantage associated with early maintenance study designs was their randomization of enriched patient populations that had already demonstrated an acute response to lithium; only responders to lithium were crossed over to placebo, or less commonly, randomized to continued treatment with lithium versus placebo. These enriched designs were advantageous because they increased the homogeneity of the study population and limited randomization to responders to the putative mood stabilizer. Limiting randomization to responders decreases the observed variability in the randomized patient population. The enriched design, however, is based on the assumption that there will be a sufficient number of monotherapy responders to the putative mood stabilizer to randomize; if not, the study is unable to go forward.

In general, the use of enriched designs appears to diminish the risk of study failure. Enriched study designs allow for a correlation of prophylactic and acute efficacy within the same patient, although the extent to which acute efficacy predicts long-term prophylactic efficacy is still unclear. These early crossover discontinuation designs were also advantageous because they were capable of generating controlled information about new compounds quickly, in contrast to designs employing random assignment to parallel groups, which typically require longer periods of time to complete. The primary methodological disadvantage associated with the use of crossover designs in the early maintenance studies appears to have been the risk of false positives. These designs appear more appropriate for use in psychiatric illnesses accompanied by chronic, persistent, nonperiodic disease courses such as schizophrenia, obsessive-compulsive disorder, generalized anxiety, and so on. The risk of false positives was particularly inherent in studies that employed relatively short experimental periods. For example, some of the early maintenance studies employed blinded study crossover periods lasting no more than 5[42] or 6[5,6] months.

The primary disadvantage associated with maintenance study designs employing random assignment to parallel groups is that they are more likely to be disturbed by spontaneous remissions or erratic, short-lived fluctuations in mood states. The duration of the maintenance study phases in random prospective designs is usually substantially longer and needs to increase in proportion to the slowness of cycling and recurrence rates in the study population. Longer studies allow for more accurate assessment of the impact of the putative mood stabilizer on disease course by including such secondary outcome measures as direct medical costs, quality of life, convenience, and compliance, but they are usually substantially more expensive. Symptom severity rating scales were not typically employed in early maintenance studies, but are now routinely included in study designs in order to establish the minimum severity of the index episode. These scales can also be used in secondary analyses. These instruments are advantageous because they can detect minor changes in illness severity, but are limited because their cross-sectional assessment is normally limited to the 7-day period preceding the completion of the instrument.

The study designs employed in both early and more recent maintenance studies are burdened by the liabilities associated with discontinuation-induced relapses. In both designs, patients are not randomized at the time of study entry. Instead, both designs randomize enriched samples of study patients that are eventually crossed over to placebo. As a result, there is a risk that discontinuation-induced relapses will skew study results. This problem can in part be managed by implementing a slow wean of the medication(s) used during the open stabilization phase of the study, then beginning the survival calculation at the beginning of the wean so that the data from early relapses are not wasted.

Time to treatment with an alternative psychotropic medication for a mood episode is probably the oldest, most commonly used, and most sensitive indicator of clinical outcome. More recent maintenance studies are returning to the use of this outcome measure. However, this

indicator is burdened by a risk that different investigators in different countries might have disparate thresholds for therapeutic intervention.

A more recent problem has emerged over the last decade as a result of the growing number of putative mood stabilizers available for clinical use in bipolar disorder. It has become increasingly difficult to enroll the more severely ill bipolar patient in controlled maintenance studies. This change in enrollment patterns has become a major impediment to the successful completion of any maintenance study. The enrollment of a less severely ill patient population makes it more difficult to detect differences between putative mood stabilizers and placebo, since placebo response rates in these less impaired patients are higher.[1,17,18] Use of randomized add-on designs such as those employed in recent trials for new drug approvals in epilepsy would help mitigate this problem.

In face of the risks associated with long-term maintenance studies, it has become increasingly clear that major mood disorders should be viewed longitudinally rather than solely in terms of the acute episode. Achieving adequate methodological rigor without sacrificing the study's overall feasibility has become an important scientific focus. There is no perfect design for the evaluation of the prophylactic efficacy of a putative mood stabilizer. In order to help break the virtual absence of NIMH-funded long-term studies in bipolar disorder, there will be a need for increased flexibility and compromise.

Drug names: haloperidol (Haldol and others), imipramine (Tofranil and others).

REFERENCES

1. Bowden CL, Swann A, Calabrese JR, et al. Maintenance clinical trials in bipolar disorder: design implications of the divalproex-lithium placebo study. Psychopharmacol Bull 1997;33:693–699
2. Baastrup PC, Poulsen JC, Schou M, et al. Prophylactic lithium: double blind discontinuation in manic-depressive and recurrent-depressive disorders. Lancet 1970;2(7668):326–330
3. Melia PI. Prophylactic lithium: a double blind trial in recurrent affective disorders. Br J Psychiatry 1970;116:621–624
4. Coppen A, Noguera R, Bailey J, et al. Prophylactic lithium in affective disorders: controlled trial. Lancet 1971;2(7719):275–279
5. Cundall RL, Brooks PW, Murray LG. A controlled evaluation of lithium prophylaxis in affective disorders. Psychol Med 1972;2:308–311
6. Hullin RP, McDonald R, Allsopp MNE. Further report on prophylactic lithium in recurrent affective disorders. Br J Psychiatry 1975;126:281–284
7. Stallone F, Shelley E, Mendlewicz J, et al. The use of lithium in affective disorders, III: a double blind study of prophylaxis in bipolar illness. Am J Psychiatry 1973;130:1006–1010
8. Prien RF, Caffey EM, Klett CJ. Prophylactic efficacy of lithium carbonate in manic depressive illness: report of the Veterans Administration and National Institute of Mental Health collaborative study group. Arch Gen Psychiatry 1973;28:337–341
9. Prien RF, Klett CJ, Caffey EM. Lithium carbonate and imipramine in prevention of affective episodes: a comparison in recurrent affective illness. Arch Gen Psychiatry 1973;29:420–425
10. Feighner JP, Robins E, Guze SB, et al. Diagnostic criteria for use in psychiatric research. Arch Gen Psychiatry 1972;26:57–63
11. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Second Edition. Washington, DC: American Psychiatric Association; 1970
12. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Third Edition. Washington, DC: American Psychiatric Association; 1980
13. Swann AC, Bowden CL, Morris D, et al. Depression during mania: effect on response to lithium or divalproex. Arch Gen Psychiatry 1997;54:37–42
14. Maj M, Pirozzi R, Magliano L, et al. Long-term outcome of lithium prophylaxis in bipolar disorder: a 5 year prospective study of 402 patients at a lithium clinic. Am J Psychiatry 1998;155:30–35
15. Goodwin FK, Jamison KR. Maintenance medical treatment. In: Manic Depressive Illness. Oxford, England: Oxford University Press; 1990:665–724
16. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. Washington, DC: American Psychiatric Association; 1994
17. Fisher S, Lipman RS, Uhlenhuth EH, et al. Drug effects and initial severity of symptomatology. Psychopharmacologia 1965;7:57–60
18. Uhlenhuth EH, Matuzas W, Warner TD, et al. Methodological issues in psychopharmacological research: growing placebo response rate: the problem in recent therapeutic trials. Psychopharmacol Bull 1997;33:31–39
19. Post RM, Frye MA, Leverich GS, et al. The role of complex combination therapy in the treatment of refractory bipolar illness. CNS Spectrums 1998;3:66–86
20. Louis TA, Lavori PW, Bailar JC, et al. Crossover and self-controlled designs in critical research. N Engl J Med 1984;310:24–31
21. Suppes T, Baldessarini RJ, Faedda GL, et al. Risk of recurrence following discontinuation of lithium treatment in bipolar disorder. Arch Gen Psychiatry 1991;48:1082–1088
22. Faedda GL, Tondo L, Baldessarini RJ, et al. Outcome after rapid vs gradual discontinuation of lithium treatment in bipolar mood disorders. Arch Gen Psychiatry 1993;50:448–455
23. Mander AJ. Is there a lithium withdrawal syndrome? Br J Psychiatry 1986;149:498–501
24. Fieve RR, Kumbaraci T, Dunner DL. Lithium prophylaxis of depression in bipolar I, bipolar II, and unipolar patients. Am J Psychiatry 1976;133: 925–929
25. Fyro B, Petterson U. A double-blind study of the prophylactic effect of lithium in manic depressive disease. Acta Psychiatr Scand 1977;262:17–22
26. Mendlewicz J. Lithium discontinuation in bipolar illness: a double blind prospective controlled study. In: Corsini GV, ed. Current Trends in Lithium and Rubidium Therapy. Lancaster, England: MTP Press; 1984:135–141
27. Markar HR, Mander AJ. Efficacy of lithium prophylaxis in clinical practice. Br J Psychiatry 1989;155:496–500
28. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56–62
29. Young RC, Biggs JT, Ziegler VE, et al. A rating scale for mania: reliability, validity, and sensitivity. Br J Psychiatry 1978;133:429–435
30. Denicoff KD, Blake KD, Smith-Jackson EE, et al. Morbidity in treated bipolar disorder: a one year prospective study using daily life chart ratings. Depression 1994;2:95–104
31. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–481
32. Shapiro DR, Quitkin FM, Fleiss JL. Response to maintenance therapy in bipolar illness: effect of index episode. Arch Gen Psychiatry 1989;46: 401–405
33. Prien RF, Kupfer DJ, Mansky PA, et al. Drug therapy in the prevention of recurrences in unipolar and bipolar affective disorders. Arch Gen Psychiatry 1984;41:1096–1104
34. Pope HG, McElroy SL, Keck P, et al. Valproate in the treatment of acute mania: a placebo controlled study. Arch Gen Psychiatry 1991;48:62–68
35. Bowden CL, Brugger AM, Swann AC, et al. Efficacy of divalproex and placebo in the treatment of mania. JAMA 1994;271:918–924
36. Sanger TM, Tohen M, Tollefson GD, et al. Olanzapine versus placebo in the treatment of acute mania. Presented at the 36th annual meeting of the American College of Neuropsychopharmacology; December 8–12, 1997; Kamuela, Hawaii
37. Müller-Oerlinghausen B, Retzow A for the European Valproate Acute Mania Study Group. Valproate as an adjunct to neuroleptic medication in the treatment of acute episodes of mania. Presented at the 1998 annual meeting of the International Academy of Biomedical and Drug Research, Workshop on Novel Therapeutic Strategies in the Schizophrenic Spectrum and Mood Disorders. Venice, Italy
38. Goodwin FK, Jamison KR: Manic Depressive Illness. New York: NY: Oxford University Press; 1990:603–629
39. Keller MB, Lavori PW, Kane JM, et al. Subsyndromal symptoms in bipolar

disorder: a comparison of standard and low serum levels of lithium. Arch Gen Psychiatry 1992;49:371–376
40. Ballenger JC, Post RM. Therapeutic effects of carbamazepine in affective illness: a preliminary report. Commun Psychopharmacol 1978;2:159–175
41. Okuma T, Inanaga K, Otsuki S, et al. Comparison of the antimanic efficacy of carbamazepine and chlorpromazine: a double-blind controlled study. Psychopharmacology (Berl) 1979;66:211–217
42. Baastrup PC. The use of lithium in manic depressive psychosis. Compr Psychiatry 1964;5:396–408

DISCLOSURE OF OFF-LABEL USAGE

The authors of this article have determined that, to the best of their clinical estimation, no investigational or off-label information about pharmaceutical agents has been presented that is outside Food and Drug Administration–approved labeling.