Evolving methodologies in bipolar disorder maintenance research

JOSEPH R. CALABRESE, DANIEL J. RAPPORT, MELVIN D. SHELTON and SUSAN E. KIMMEL

Background During the development of a new treatment for bipolar disorder, maintenance studies are used to evaluate the ability of the putative mood stabiliser to prevent relapse and recurrence of further episodes. Comparisons with the early bipolar disorder maintenance studies indicate that the methodologies of recent trials have evolved substantially.

Aims To review the methods used in the first- and second-generation maintenance studies, highlighting the differences of the various designs.

Method Literature review.

Results Methods that have evolved the most include patient enrolment, randomisation schemes and the use of outcome measures and statistical analyses. In addition, regulatory and commercial issues have also influenced study design.

Conclusion There is little consensus on the methodology of bipolar disorder maintenance studies. As the integration of newer therapies into routine clinical practice is dependent on the evidence from controlled studies, it is essential that future maintenance trials in bipolar disorder achieve adequate methodological rigour without sacrificing overall feasibility.

Declaration of interest Funding and consultation fees (detailed in Acknowledgements) received from various sources, including the pharmaceutical industry.

The pharmacological management of bipolar disorder involves not only the acute treatment of manic and depressive episodes, but also maintenance treatment to prevent relapse and recurrence of further episodes. The study by Bowden *et al* (1997) is the first placebo-controlled maintenance trial in bipolar I disorder to be conducted since 1973. Comparisons with earlier maintenance studies indicate that study methodologies have evolved substantially over that period. The methods employed have differed depending upon the decade in which the study was designed and whether the results were intended for submission to a regulatory agency, country-specific requirements have also had a strong influence on study design.

STUDY ENROLMENT

The first-generation maintenance studies evaluated the effects of long-term lithium therapy in patients with bipolar I disorder. These were conducted between 1960 and 1970, and published between 1970 and 1976 (Baastrup *et al*, 1970; Melia, 1970; Coppen *et al*, 1971; Stallone *et al*, 1973; Hullin *et al*, 1975; Fieve *et al*, 1976). Most of the studies tended to evaluate small cohorts of patients, typically including five to 40 patients per study arm. A notable exception, however, is the study by Prien *et al* (1973*a*,*b*), which enrolled 205 patients (101 patients received lithium and 104 patients were given placebo). Generally, the mean enrolment size of the study population in the early maintenance trials was 31 (range 7–101) for the lithium arms and 28 (range 8–104) for the placebo arms. More recently, issues associated with a drug's remaining patent life and time to market-place have become compelling reasons for industry-sponsored drug development efforts, which have led to large-scale multicentre studies with as many as 50 to 60 sites. For example, the study by Bowden *et al* (1997) enrolled 571 patients from 35 centres.

It is likely that the enrolment practices of maintenance studies have been affected by the evolving nature of our psychiatric nomenclature. At the time of the early studies, concepts such as bipolar II disorder, secondary bipolar disorders and rapid cycling were yet to be introduced into clinical practice. The Research Diagnostic Criteria (RDC) (Feighner *et al*, 1972), the second version of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM–II; American Psychiatric Association, 1970) and the clinical standards then most commonly employed by psychiatrists placed importance on the role of hospitalisation 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 the more recent maintenance studies (American Psychiatric Association, 1980). In the RDC, patients with mood-incongruent psychotic symptoms were categorised as having the "mainly affective variant of schizoaffective disorder", not bipolar disorder. Consequently, patients with mania enrolled in the early studies tended to have a disproportionate representation of classic euphoric mania. If they had psychotic symptoms, the content of their hallucinations and delusions was mood-congruent. This patient subtype is now recognised as being one of the more lithium-responsive variants of the disorder (Swann *et al*, 1997).

Comparisons between studies suggest that enrolment practices of the early maintenance trials may have selected preferentially for the inclusion of patients with the more lithium-responsive type of bipolar disorder. For example, in the early studies the mean relapse rates with lithium therapy were 23% for mania and 21% for depression over maintenance phases usually lasting 12 months (Goodwin & Jamison, 1990*a*). These rates are inconsistent with the results of the study by Maj *et al* (1998), which showed that over 5 years of prospective follow-up of 402 patients with bipolar I disorder, mean overall relapse rates with lithium were 76%. The decline in response rates is possibly owing (at least in part) to the changing patient populations of the recent studies. However, this shift may be secondary to the changes in the nomenclature, which now includes such atypical variants of bipolar disorder as abnormal mood states with incongruent psychotic symptoms, mixed or dysphoric

https://doi.org/10.1192/bjp.178.41.s157 Published online by Cambridge University Press

mania, rapid cycling and presentations that are comorbid with alcohol or drug misuse.

In addition to selecting the more lithium-responsive patients, first-generation maintenance studies also tended to enrol more severely ill patients. The main advantage of enrolling these patients is that the power of the study to detect clinically and statistically significant differences between the putative mood stabiliser and placebo improves markedly. Nevertheless, enrolment of more severely ill patients can increase the number of participants dropping out because of poor compliance, adverse events, lack of efficacy, withdrawn consent and protocol violations. Accordingly, randomisation 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 end-points so that the data are not wasted. Alternatively, the use of this methodology may introduce bias if there were differential attrition rates from different treatment conditions.

Early maintenance studies were more likely to enrol 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 enrol because the number of effective treatment alternatives has increased. Patients with severe illness are now usually taking three to five different medications (Post *et al*, 1998) and are less interested or less able to enrol in rigidly designed studies employing complete drug wash-outs followed by monotherapy treatment arms or the use of placebo. In addition, health-care reform has led to a decrease in the use of hospitalisation for patients with bipolar disorder. Consequently, investigators have become reluctant or unable to enrol more seriously ill patients in bipolar maintenance studies, considering the enrolment of the severely ill in out-patient studies to be unethical.

A recent example of study design influencing selection of a relatively mildly affected cohort of patients is the maintenance study by Bowden *et al* (1997). These investigators reported that the long placebo-based treatment phase (52 weeks) discouraged enrolment of patients experiencing frequent episodes or with serious functional impairment as a consequence of episodes, since these patients were endeavouring to avoid a repeat of such problems. In addition, a requirement for entry into the randomised phase of the study was a sustained Global Assessment Scale (GAS) (Endicott *et al*, 1976) score of more than 60. The definition of a GAS score in the range of 61 to 70 includes the statement that "most untrained persons would not consider the person sick". There is evidence that even responsive patients recovering from an acute manic episode retain substantial functional impairment for several months (Dion *et al*, 1989). Another element of the Bowden study that may have contributed toward enrolment of less severely ill patients is that eligible patients included those whose manic episode met the inclusion criteria but subsided without specific antimanic treatment during the open phase (Bowden *et al*, 1997). As the investigators note, "it is likely that such patients are atypical, in that most full manic episodes require treatment for resolution over several weeks. Additionally, such patients may have cycled spontaneously into depressive symptomatology, which added variability to the randomised sample."

Uhlenhuth *et al* (1997) presented evidence that therapeutic drug trials are much less likely to demonstrate a significant drug–placebo difference if patients with mild symptoms are enrolled. Using the example of panic disorder, they demonstrated that only patients with many panic attacks at baseline assessment benefited more from the active treatments than from placebo. The explanation for the inability of studies of mildly ill cohorts to demonstrate a true drug–placebo difference has been developed by Benjamin (1963) and Fisher *et al* (1965), who proposed that effective treatments suppress symptoms to a uniformly low level irrespective of baseline severity, whereas placebo-responding patients have a uniform but modest symptom reduction across the spectrum of initial severity. Thus, only for patients with moderate severity or greater will the difference between treatments be apparent.

RANDOMISATION SCHEMES

Crossover design

The first-generation bipolar studies most commonly employed double-blind crossover designs in which enriched sample groups of patients who responded to lithium were crossed over to placebo (discontinuation designs). The advantage of these enriched designs was that the homogeneity of the study population was increased and the randomisation of responders to the putative mood stabiliser was limited, decreasing the observed variability in the randomised 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 stabiliser to randomise; 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. Additionally, crossover discontinuation designs generate controlled information about new compounds quickly, in contrast to designs employing random assignment to parallel groups, which typically take longer 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-positive results. This was particularly inherent in studies that employed short experimental periods. For example, some of the early maintenance studies employed blinded study crossover periods lasting no more than 5 months (Baastrup, 1964) or 6 months (Candela *et al*, 1972; Hullin *et al*, 1975). In addition, all of the early studies discontinued lithium treatment without tapering in those patients assigned to placebo. The withdrawal of lithium, and in particular its rapid withdrawal, has been shown to significantly increase the risk of early recurrence to a rate that exceeds that predicted by the natural course of the untreated illness (Mander, 1986; Suppes *et al*, 1991; Faedda *et al*, 1993). Therefore, it is possible that, as a result of discontinuation-induced relapses when patients crossed over to placebo, early maintenance studies favoured the appearance of improved outcomes in the lithium-treated participants.

Unless enrolment procedures standardise for the pattern of the index episode (the polarity of the episode that brought the patient into the study) at the time of study entry, crossover designs will carry the risk of altering outcome as an artefact 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

https://doi.org/10.1192/bjp.178.41.s157 Published online by Cambridge University Press

medication, and have a relapse falsely attributed to placebo or to the next study medication. Since bipolar disorder is characterised by periodic relapses and remissions, there is the likelihood of false-positive results from spontaneous remissions. Consequently, crossover designs are more suited to studies of psychiatric illnesses accompanied by chronic, persistent, non-periodic disease courses such as schizophrenia, obsessive–compulsive disorder and generalised anxiety.

Premature drug discontinuations can also affect studies using a crossover design (Louis *et al*, 1984). Although these are a serious problem for any drug trial, the impact of premature drop-out in a crossover study is exaggerated because each remaining patient then contributes a larger proportion of the study data. Drop-out rates can be high in crossover studies since each patient must receive at least two treatments to provide complete data. In studies of illnesses such as bipolar disorder, where the prevalence of poor compliance is substantial, a high drop-out rate weakens a crossover study. Accordingly, the initial sample size should be sufficiently large to compensate for this effect. In addition, treatments used should have few side-effects severe enough to result in drug discontinuation.

Parallel design

Although a crossover design has the potential advantage of providing a direct comparison of treatments in the same patient, the parallel design is less dependent on assumptions about the disease process, and generally produces a lower drop-out rate because each patient is exposed to only one treatment. Nevertheless, the primary disadvantage of employing random assignment to parallel groups is that these studies 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 stabiliser 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.

One of the first-generation studies by Coppen *et al* (1971) used prospective random assignment to parallel groups receiving lithium or placebo. Although double-blind study medication was continued throughout the entire drug trial, the psychiatrist in charge of the case could also prescribe additional 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 stabilisation periods rather than immediately proceeding with randomisation to the putative mood stabiliser and placebo.

Most ongoing studies are now employing 'enriched' designs in which responders to a putative mood stabiliser in monotherapy are again randomly assigned to continue treatment with the new drug, placebo or lithium in parallel. The design is considered to be enriched because it randomises 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 favour of a design considered by some regulatory agencies to be more suitable for disorders accompanied by frequent periodic relapse. However, there are liabilities associated with the use of enriched designs as well, including limited generalisability to the general population of patients with the illness under study. In effect, response rates are artificially inflated owing to the enriched nature of the population of patients who are randomised.

More recently, the study by Bowden *et al* (1997) randomly assigned subjects to parallel groups following open stabilisation with medications chosen completely at the discretion of the investigator. Given this development, the field appears to have cycled from an extreme of randomising a homogeneous cohort of lithium monotherapy responders to the other extreme of allowing any medications for use during the open stabilisation phase of the trial. As a result of this more recent study design, the participation group may become unmanageably heterogeneous. An alternative to this method might be limiting the use of psychotropic medications during the open stabilisation period of the maintenance study to only the putative mood stabiliser and lithium (if lithium is being used to gauge study reliability) during the blinded phase.

OUTCOME MEASURES

Most early lithium prophylaxis studies did not measure mood severity using symptom severity rating scales, such as the Hamilton Rating Scale for Depression (Hamilton, 1960) or the Young Mania Rating Scale (Young *et al*, 1978). Typically, these trials evaluated the efficacy of lithium treatment by general indices of outcome during the study period (Coppen *et al*, 1971), the number of manic and depressive episodes (Stallone *et al*, 1973; Fieve *et al*, 1976) or the probability of manic or depressive episodes (Prien *et al*, 1973*a*; Fyro & Petterson, 1977; Mendlewicz, 1984; Markar & Mander, 1989). Symptom severity is now routinely quantified with rating scales to establish the minimum severity of the index episode. These scales are also being used as secondary analyses outcome measures. The advantage of rating scales is that minor changes in illness severity can be detected. However, it should be noted that there are limitations because their cross-sectional assessment is normally limited to the 7-day period preceding the completion of the study. In addition, symptom severity scales do not adequately assess sub-syndromal presentations of hypomania or depression. Many studies now require a minimum rating score on a symptom severity scale for a randomised subject to reach study end-point. This method, though, can miss other indications of medication inferiority such as patient satisfaction, quality of life and tolerability.

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 the risk that different investigators in different countries might have disparate thresholds for therapeutic intervention. Over the past 10–15 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 hospitalisation, which was the most common outcome measure in

https://doi.org/10.1192/bjp.178.41.s157 Published online by Cambridge University Press

first-generation research, may be too rigid. This measure may not be sensitive enough to uncover lesser degrees of difference between (less potent) active compounds and a placebo. The time from the beginning of randomisation to the first prescription of a psychotropic medication might be a more sensitive indicator of clinical efficacy (as employed by Coppen *et al*, 1971). However, there is also a problem associated with the use of this outcome measure because it is difficult to prevent non-compliant patients in some countries from receiving psychiatric medications through non-research-related sources and being incorrectly categorised as responders.

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 one primary outcome measure as the primary determinant of efficacy, Coppen *et al* (1971) relied on several measures, including a seven-point global rating scale, percentage of time spent as an in-patient, percentage of time spent as an out-patient, other treatments prescribed during the trial and premature study discontinuations.

Most of the early bipolar disorder maintenance studies were analysed by comparing the proportion of patients who experienced a relapse or recurrence. However, this approach is inadequate for a long-term trial such as a maintenance study because it does not consider the length of time that patients remain well before relapsing. As a result, relapses at 1 month and 12 months of a 1-year trial would incorrectly suggest similar degrees of efficacy. A further drawback of this method is that the numbers of patients withdrawing prematurely without experiencing a relapse is either ignored or analysed incorrectly. For example, in a study using relapse into mania as the primary outcome measure, important clinical data such as relapse into hypomania or depression, premature discontinuation from the study due to intolerable side-effects, or drop-out due to poor compliance, would not be included in the primary analysis.

The use of survival analyses has now become the standard method of examining data from bipolar disorder maintenance studies. This approach uses the time to relapse or recurrence (as defined in the protocol) as the measure of outcome. One of the more commonly used survival analyses is the product-limit method developed by Kaplan & Meier (1958). This procedure extends the idea underlying life-table analysis (Fleiss, 1986). The life-table calculates the probabilities of remaining well during a study period and the median time in remission for each treatment. Both the life-table and product-limit methods make no model assumptions and represent a summary of the raw data. However, in traditional life-table analysis, relapses are grouped into intervals of time, consequently, when the actual times to relapse are available, the life-table method wastes data such as cumulative time spent well. In contrast, the product-limit method makes use of the exact time to relapse to compute cumulative probabilities of remaining well. Cases where patients withdraw prematurely from the study for reasons other than a relapse or recurrence are also analysed until the time of drop-out and are then withdrawn or censored from the analysis.

The importance of survival analysis methods was highlighted in the re-analysis of the National Institute of Mental Health (NIMH) collaborative study (Prien *et al*, 1984) by Shapiro *et al* (1989). In the NIMH study, patients were randomised to receive lithium, imipramine or a combination of both treatments. Using responder analyses, the results showed that for patients with a manic index episode, lithium and the combination medication were superior to imipramine in preventing relapse. In contrast, all three treatments were equally effective in patients with a depressive index episode (Prien *et al*, 1984). When re-examined using survival analysis (Shapiro *et al*, 1989), similar findings to the NIMH study were reported for patients with manic episodes; both lithium and the combination were superior to imipramine. However, re-analysis of the data for patients with depressive episodes produced strikingly different findings: survival analysis methods showed the combination to be significantly superior to either imipramine or lithium administered alone. Shapiro *et al* (1989) noted that survival analyses "are more powerful techniques that may detect effects not found by other analytical methods", and recommended that "survival techniques be used to analyse future maintenance trials in bipolar illness".

Survival analyses usually require the prospective selection of a primary outcome measure. It is now well recognised that there are many ways of defining outcome. Patients define outcome by emphasising convenience and safety, whereas managed-care companies define outcome by direct and indirect costs. Alternatively, investigators and pharmaceutical companies base outcome on measures most likely to separate differences in efficacy between 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.

In the early bipolar disorder maintenance studies the mean overall placebo response rate was low – only 21% (Goodwin & Jamison, 1990*b*). However, this rate was based on responder analyses of observed or completer data rather than survival analyses performed on last observations carried forward on the intent-to-treat sample of study patients. In contrast, Bowden and colleagues reported an apparently high placebo response rate of 62% (Bowden *et al*, 1997). However, this study performed survival analyses on the intent-to-treat sample (patients who received at least one dose of study medication). The 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 (Denicoff *et al*, 1994). 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 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, partlcularly those dropping out because of withdrawn consent, protocol violations and so on. As these tend to occur disproportionately in patients in the placebo group; fewer patients receiving placebo are left to be at risk of relapse, decreasing the power of the study and representing a disadvantage associated with survival analyses that employ time to relapse or recurrence as the primary outcome

https://doi.org/10.1192/bjp.178.41.s157 Published online by Cambridge University Press

measure. For example, of the 94 patients assigned to placebo in the Bowden study, 70 patients (75%) dropped out prematurely: 21 (22%) owing to relapse into mania requiring hospitalisation or a mania rating score of 16 points; 15 (16%) owing to depression; 11 (12%) owing to intolerance/poor compliance; and 23 (25%) because of other reasons, such as protocol violations, withdrawn consent, loss to follow-up and intercurrent illness (Bowden *et al*, 1997). Only 38% relapsed into episodes of mania or depression meeting strict criteria, which leaves the false impression that overall 62% of patients responded to placebo.

REGULATORY ISSUES

The substantial variation in the requirements of regulatory agencies throughout the world will also have a strong influence on study design. All regulatory agencies require a thorough examination of basic pharmacokinetics, safety and efficacy. However, significant differences exist in their requirements for concurrent pivotal data on acute and prophylactic treatment as well as for both manic and depressive phases of the illness. Recent experience with the Food and Drug Administration suggests that three factors are required for acute indications in the USA: pharmacokinetic information, two acute pivotal data-sets and safety in both acute and longitudinal settings. Regulatory agencies in Europe additionally request maintenance efficacy data and quite possibly continuation phase/intermediate duration (12 weeks) data.

Since most of the recent putative mood stabilisers were first developed as anti-epileptic drugs, it is interesting to focus on the methods by which the same drug is developed for these two different types of illness. When undergoing development for use in epilepsy, double-blind, placebo-controlled, add-on/augmentation trials are initially used to evaluate the efficacy of the putative anti-epileptic agent. Not infrequently, the drugs are blindly added to regimens of two or three other concomitantly prescribed anti-epileptic 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. As placebo response rates in patients with bipolar disorder are markedly higher than in patients with epilepsy, this explains why placebo-controlled add-on maintenance trials have not been previously conducted in bipolar disorder.

COMMERCIAL ISSUES

Bipolar disorder is a 'massively neglected problem' (Montgomery, 1997) with current treatments being effective in only about half of those affected. At least one-third of patients with bipolar disorder relapse despite good compliance, and most of the cost of the disorder can be attributed to ineffective treatment (Greenberg *et al*, 1993). Nevertheless, there is a growing appreciation within the pharmaceutical industry of the enormous unmet need in bipolar disorder and the commercial opportunity that it represents. Following the demonstration that anticonvulsant drugs were effective in bipolar disorder (Ballenger & Post, 1978; Okuma *et al*, 1979), more pharmaceutical companies are realising that all anti-epileptic drugs should be considered as putative mood stabilisers and developed as such. Issues with remaining patent life and time to market-place are particular concerns for companies with anti-epileptic drugs under development for bipolar disorder, since many years of the patent lives of these drugs have already expired during their development for use in epilepsy. However, drug development is often streamlined when basic phase I (pharmacokinetic) studies have already been completed.

Although the development of mood stabilisers has largely centred on anti-epileptic drugs, resources have also been invested in the development of atypical antipsychotic medications and selective serotonin reuptake inhibitors. Despite these welcome trends, it is clear that there is no 'ideal mood stabiliser' available within the current range of treatments for bipolar disorder. Such an agent would be effective in each phase of the illness and could be used as monotherapy. In addition, it would have acute onset of effect within hours of administration, low rates of drug discontinuation due to side-effects, no side-effects requiring additional treatment and also be safe in overdose.

Another commercial issue that affects drug development in bipolar disorder is the extent to which the pharmaceutical industry has relied on expert psychiatric consultation from the research community. Although a variety of agents are now available for both acute and long-term therapy, there is an urgent requirement for controlled studies to provide evidence for the effectiveness of the newer therapies so that they can be integrated into routine clinical practice. The increased use of expert advisory panels would have a beneficial effect in drug development.

CONCLUSION

Reviewing the methods used in the first- and second-generation maintenance studies in bipolar disorder indicates that study design has evolved greatly since the 1970s, and that each of the various designs has different strengths and weaknesses. Consequently, there is little consensus about the methods for evaluating the prophylactic efficacy of a putative mood stabiliser. The methods that have evolved the most include enrolment procedures, randomisation schemes, use of outcome measures, statistical analyses and country-specific regulatory and commercial issues.

Old and new study designs alike are burdened by the problems of discontinuation-induced relapses. If a patient is stabilised on one medication during the open stabilisation phase of a maintenance study and that medication is abruptly discontinued following randomisation, as is the case with both crossover and enriched parallel designs, patients who are randomly assigned to a placebo group can experience an iatrogenically increased risk of relapse. Consequently, the results of the study tend to be inflated, widening the apparent advantage between study treatments. This problem can be managed effectively by extending the tapering-off period and starting the time-to-relapse survival analysis when blinding and the weaning to study medication begin. The latter is required because a significant number of patients, especially those with rapid-cycling disorders, will relapse quickly and their data will therefore be unavailable to the survival analyses if the analyses do not begin until the wean is complete (Faedda *et al*, 1993).

Enrolment in the early maintenance studies may have selected preferentially patients with the more lithium-responsive forms of bipolar disorder. The most recent version of the DSM includes such atypical variants of bipolar disorder as abnormal mood states with incongruent psychotic symptoms, mixed or dysphoric mania and

https://doi.org/10.1192/bjp.178.41.s157 Published online by Cambridge University Press

rapid cycling (American Psychiatric Association, 1994). However, less is known about these variants of illness as they have not been extensively studied. In particular, the degree of subtype response to placebo is unclear. Nevertheless, when considered as a group, these atypical variants of bipolar disorder are more prevalent than the classic forms of the illness. Therefore, new studies should include these more common patterns of presentation, to improve generalisability to routine clinical practice.

Because of the growing number of putative mood stabilisers available for clinical use in bipolar disorder, it has become increasingly difficult to enrol the more severely ill patient in controlled maintenance studies. Accordingly there has been a selection bias toward a less ill – and therefore less representative – study population. This change in enrolment patterns has become a major barrier to the unsuccessful completion of any maintenance study. Since placebo response rates in these less-impaired patients are higher, it becomes increasingly difficult to detect differences between putative mood stabilisers and placebo (Fisher *et al*, 1965; Bowden *et al*, 1997; Uhlenhuth *et al*, 1997). Use of randomised add-on designs such as those employed in recent trials for new drug approvals in epilepsy would help relieve this problem.

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

In conclusion, it is clear that more longitudinal research is needed, with the effectiveness of the treatment studied on a long-term basis rather than solely during acute episodes of mania and depression. Bipolar disorder is a complex illness characterised by disparate phases and multiple patterns of presentation. Rapid-cycling and mixed states such as dysphoric mania and hypomania are particularly difficult to study over short periods, as they migrate unpredictably throughout the natural course of the illness. Achieving adequate methodological rigour without sacrificing the study's overall feasibility has become an important scientific focus. In order to overcome the virtual absence of NIMH-funded long-term studies in bipolar disorder, there will be a need for increased flexibility and compromise.

CLINICAL IMPLICATIONS

- There is minimal consensus about methods of evaluating the prophylactic efficacy of a putative mood stabiliser.
- Controlled studies are needed to determine the effectiveness of newer therapies both as acute and as long-term therapy.
- Future studies should also focus on atypical variants of bipolar disorder such as rapid cycling.

LIMITATIONS

- In recent years study enrolment has been biased towards a less ill and therefore less representative population.
- Studies using crossover or parallel designs can be heavily influenced by discontinuation-induced relapses.

JOSEPH R. CALABRESE, MD, DANIEL J. RAPPORT, MD, MELVIN D. SHELTON, PhD, SUSAN E. KIMMEL, MD, Mood Disorders Program, Department of Psychiatry, Case Western Reserve University, Cleveland, Ohio, USA

Correspondence: Dr Joseph R. Calabrese, Mood Disorders Program, Department of Psychiatry, Case Western Reserve University, School of Medicine, 10900 Euclid Avenue, Cleveland, OH 44106, USA. Tel: +1 216 844 2850; fax: +1 216 844 2875; e-mail: jrc8@po.cwru.edu

ACKNOWLEDGEMENTS

Sources of funding for clinical grants: National Institute of Mental Health; Abbott Laboratories; CIBA-GEIBY; E-Merck; Forest Labs; Glaxo-Wellcome Pharmaceuticals; Lilly Research Laboratories; MacArthur Foundation; National Alliance for Research in Schizophrenia and Affective Disorders; Parke-Davis Pharmaceuticals; Robert Wood Johnson Pharmaceutical Research Institute; Sandoz Pharmaceuticals Corporation; SmithKline Pharmaceuticals; Stanley Foundation; Tap Holdings, Inc.; UCB Pharma; and Wyeth Ayerst Pharmaceuticals.

Consulting Agreements: Abbott Pharmaceuticals; AstraZeneca; Elan Pharmaceuticals; Eli Lilly; Glaxo-Wellcome; Janssen Cilag; Novartis; Parke Davis; Robert Wood Johnson Pharmaceutical Research Institute; Shire Labs; SmithKline; TAP Holdings; Teva Pharmaceuticals; and UCB Pharma.

REFERENCES

American Psychiatric Association (1970) *Diagnostic and Statistical Manual of Mental Disorders* (2nd edn) (DSM–II). Washington, DC: APA.

— (1980) *Diagnostic and Statistical Manual of Mental Disorders* (3rd edn) (DSM–III). Washington, DC: APA.

— (1994) *Diagnostic and Statistical Manual of Mental Disorders* (4th edn) (DSM–IV). Washington, DC: APA.

Baastrup, P. C. (1964) The use of lithium in manic depressive psychosis. *Comprehensive Psychiatry*, **5**, 396–408.

—, Poulsen, J. C., Schou, M., *et al* (1970) Prophylactic lithium: double blind discontinuation in manic–depressive and recurrent-depressive disorders. *Lancet*, **2**, 326–330.

Ballenger, J. C. & Post, R. M. (1978) Therapeutic effects of carbamazepine in affective illness: a preliminary report. *Communications in Psychopharmacology*, **2**, 159–175.

Benjamin, L. S. (1963) Statistical treatment of the Law of Initial Values (LIV) in automic research: a review and recommendation. *Psychosomatic Medicine*, **25**, 556–566.

Bowden, C. L., Swann, A., Calabrese, J. R., *et al* (1997) Maintenance clinical trials in bipolar disorder: design implications of the divalproex–lithium placebo study. *Psychopharmacology Bulletin*, **33**, 693–699.

Candela, R. L., Brooks, P. W. & Murray, L. G. (1972) A controlled evaluation of lithium prophylaxis in affective disorders. *Psychological Medicine*, **2**, 308–311.

Coppen, A., Noguera, R., Bailey, J., *et al* (1971) Prophylactic lithium in affective disorders: controlled trial. *Lancet*, **2**, 275–279.

Denicoff, K. D., Blake, K. D., Smith-Jackson, E. E., *et al* (1994) Morbidity in treated bipolar disorder: a one

https://doi.org/10.1192/bjp.178.41.s157 Published online by Cambridge University Press

year prospective study using daily life chart ratings. *Depression*, **2**, 95–104.

Dion, G. L., Tohen, M., Anthony, W. A., *et al* (1989) Symptoms and functioning of patients with bipolar disorder six months after hospitalization. *Hospital and Community Psychiatry*, **39**, 652–656.

Endicott, J., Spitzer, R. L., Fleiss, J. L., *et al* (1976) The Global Assessment Scale. A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry*, **33**, 766–771.

Faedda, G. L., Tondo, L., Baldessarini, R. J., *et al* (1993) Outcome after rapid vs gradual discontinuation of lithium treatment in bipolar mood disorders. *Archives of General Psychiatry*, **50**, 448–455.

Feighner, J. P., Robins, E., Guze, S. B., *et al* (1972) Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry*, **26**, 57–63.

Fieve, R. R., Kumbaraci, T. & Dunner, D. L. (1976) Lithium prophylaxis of depression in bipolar I, bipolar II, and unipolar patients. *American Journal of Psychiatry*, **133**, 925–929.

Fisher, S., Lipman, R. S., Uhlenhuth, E. H., *et al* (1965) Drug effects and initial severity of symptomatology. *Psychopharmacologia*, **7**, 57–60.

Fleiss, J. L. (1986) Analysis of data from multiclinic trials. *Controlled Clinical Trials*, **7**, 267–275.

Fyro, B. & Petterson, U. (1977) A double-blind study of the prophylactic effect of lithium in manic depressive disease. *Acta Psychiatrica Scandinavica*, **262**, 17–22.

Goodwin, F. K. & Jamison, K. R. (1990*a*) Maintenance medical treatment. In *Manic Depressive Illness*, pp. 665–724. Oxford: Oxford University Press.

— & — (1990*b*) Medical treatment of manic episodes. In *Manic Depressive Illness*, pp. 603–629. Oxford: Oxford University Press.

Greenberg, P. E., Stiglin, L. E., Finkelstein, S. N., *et al* (1993) The economic burden of depression in 1990. *Journal of Clinical Psychiatry*, **54**, 405–418.

Hamilton, M. (1960) A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, **23**, 56–62.

Hullin, R. P., McDonald, R. & Allsopp, M. N. E. (1975) Further report on prophylactic lithium in recurrent affective disorders. *British Journal of Psychiatry*, **126**, 281–284.

Kaplan, E. L. & Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Louis, T. A., Lavori, P. W., Ballar, J. C., *et al* (1984) Crossover and self-controlled designs in clinical research. *New England Journal of Medicine*, **310**, 24–31.

Maj, M., Pirozzi, R., Magliano, L., *et al* (1998) Long-term outcome of lithium prophylaxis in bipolar disorder: a 5 year prospective study of 402 patients at a lithium clinic. *American Journal of Psychiatry*, **155**, 30–35.

Mander, A. J. (1986) Is there a lithium withdrawal syndrome? *British Journal of Psychiatry*, **149**, 498–501.

Markar, H. R. & Mander, A. J. (1989) Efficacy of lithium prophylaxis in clinical practice. *British Journal of Psychiatry*, **155**, 496–500.

Melia, P. I. (1970) Prophylactic lithium: a double-blind trial in recurrent affective disorders. *British Journal of Psychiatry*, **116**, 621–624.

Mendlewicz, J. (1984) Lithium discontinuation in bipolar illness: a double blind prospective controlled study. In *Current Trends in Lithium and Rubidium Therapy* (ed. G. V. Corsini), pp. 135–141. Lancaster: MTP Press.

Montgomery, S. (1997) The treatment of bipolar disorder. *Journal of Bipolar Disorders*, **1**, 20–22.

Okuma, T., Inanaga, K., Otsuki, S., *et al* (1979) Comparison of the antimanic efficacy of carbamazepine and chlorpromazine: a double-blind controlled study. *Psychopharmacology (Berl.)*, **66**, 211–217.

Post, R. M., Frye, M. A., Leverich, G. S., *et al* (1998) The role of complex combination therapy in the treatment of refractory bipolar illness. *CNS Spectrums*, **3**, 66–86.

Prien, R. F., Caffrey, E. M. & Klett, C. J. (1973*a*) Prophylactic efficacy of lithium carbonate in manic depressive illness: report of the Veterans Administration and National Institute of Mental Health collaborative study group. *Archives of General Psychiatry*, **28**, 337–341.

—, Klett, C. J. & Caffey, E. M. (1973*b*) Lithium carbonate and imipramine in prevention of affective episodes: a comparison in recurrent affective illness. *Archives of General Psychiatry*, **29**, 420–425.

—, Kupfer, D. J., Mansky, P. A., *et al* (1984) Drug therapy in the prevention of recurrences in unipolar and bipolar affective disorders. *Archives of General Psychiatry*, **41**, 1096–1104.

Shapiro, D. R., Quitkin, F. M. & Fleiss, J. L. (1989) Response to maintenance therapy in bipolar illness: effect of index episode. *Archives of General Psychiatry*, **46**, 401–405.

Stallone, F., Shelley, E., Mendlewicz, J., *et al* (1973) The use of lithium in affective disorders, III: a double blind study of prophylaxis in bipolar illness. *American Journal of Psychiatry*, **130**, 1006–1010.

Suppes, T., Baldessarini, R. J., Faedda, G. L., *et al* (1991) Risk of recurrence following discontinuation of lithium treatment in bipolar disorder. *Archives of General Psychiatry*, **48**, 1082–1088.

Swann, A. C., Bowden, C. L., Morris, D., *et al* (1997) Depression during mania: effect on response to lithium or divalproex. *Archives of General Psychiatry*, **54**, 37–42.

Uhlenhuth, E. H., Matuzas, W., Warner, T. D., *et al* (1997) Methodological issues in psychopharmacological research: growing placebo response rate: the problem in recent therapeutic trials. *Psychopharmacology Bulletin*, **33**, 31–39.

Young, R. C., Biggs, J. T., Ziegler, V. E., *et al* (1978) A rating scale for mania: reliability, validity and sensitivity. *British Journal of Psychiatry*, **133**, 429–435.

https://doi.org/10.1192/bjp.178.41.s157 Published online by Cambridge University Press