# The Development of Study Exit Criteria for Evaluating Antimanic Compounds

Ellen Frank, Ph.D.; David J. Kupfer, M.D.; Alexandre Gerebtzoff, M.D.; Uwe Meya, M.D.; Fouzia Laghrissi-Thode, M.D.; Victoria J. Grochocinski, Ph.D.; Patricia R. Houck, M.S.H.; Alan G. Mallinger, M.D.; and Robert D. Gibbons, Ph.D.



***Background:*** There is increasing interest on the part of investigators and the public at large in finding ways to study and improve treatments for the seriously mentally ill without exposing such individuals to unnecessary risks. One group of particular interest in this regard are patients suffering from acute mania. We set out to define "exit" criteria or novel clinical endpoints that might help to assess the efficacy of antimanic compounds. We sought a method that would be safer, more economical, and less sensitive to nonspecific factors in the clinical environment while still allowing unambiguous assessment of efficacy.

***Method:*** From a pool of subjects being screened for or already participating in intervention studies, we retrospectively identified 76 admissions of patients with a manic or mixed episode according to DSM-IV. We fit a mixed-effects regression model to all available data obtained using the Bech-Rafaelsen Mania Scale from admission to day 28 of treatment. Using the estimated model coefficients, we obtained empirical Bayes (EB) estimates of each subject's trend coefficients based on (1) all available data and (2) data through day 11 of treatment for mania.

***Results:*** We found a high correlation (r = .67) between EB estimates of final response at day 28 and actual day 28 scores on the Bech-Rafaelsen scale based on scores through day 11. When subjects were categorized as full, partial, or nonresponders according to their final Bech-Rafaelsen score, we were able to show that only 2 of the 23 predicted nonresponders became full responders, 27 of the 31 predicted full responders became full responders, and 16 of the 22 predicted partial responders became partial or full responders.

***Conclusion:*** We conclude on the basis of this chart review study that it should be possible to define exit criteria for trials assessing the efficacy of antimanic compounds on the basis of relatively short duration exposure to experimental treatment.

*(J Clin Psychiatry 2001;62:421–425)*

*Received Jan. 19, 2000; accepted Oct. 9, 2000.* *From the University of Pittsburgh School of Medicine, Western Psychiatric Institute and Clinic, Pittsburgh, Pa. (Drs. Frank, Kupfer, Grochocinski, and Mallinger and Ms. Houck); Novartis Pharma AG, Basel, Switzerland (Drs. Gerebtzoff and Laghrissi-Thode); F. Hoffman-La Roche AG, Basel, Switzerland (Dr. Meya); and the University of Illinois at Chicago, Chicago (Dr. Gibbons).*

*Supported in part by National Institute of Mental Health grants MH-30915 (Dr. Kupfer), MH-29618 (Dr. Frank), and MH-50634 (Dr. Mallinger) and an educational grant from Novartis (Dr. Kupfer).*

*The authors thank the following collaborators who helped to compile data for this report: Lisa Arvay, B.S.N., M.B.A.; Melanie Berg, B.A.; Joan Buttenfield, B.S.N.; Kim Cristiano, M.Ed.; Gail Kepple, B.A.; Denine Sims; Jean Smith, R.N.; Jennifer Smith; Debbie Stapf, B.S.; Steve Verfaille, M.S.W.; and Kelly Forster Wells, L.S.W.*

*Reprint requests to: Ellen Frank, Ph.D., Western Psychiatric Institute and Clinic, Bellefield Towers, 3811 O'Hara St., Pittsburgh, PA 15213 (e-mail: FrankE@msx.upmc.edu).*

The evaluation of antimanic compounds has typically involved the random assignment of subjects to a placebo-controlled trial of several weeks' duration, with the result that subjects assigned to the placebo arm may experience relatively little relief of their symptoms and are withdrawn from the trial only in the face of deterioration. This classical approach, i.e., evaluating differences in mean rating scale score changes over time between groups, requires a large number of patients and lengthy trials. Furthermore, it extends unnecessarily the treatment period of subjects not responding to placebo, ineffective compounds, or ineffective doses. It is possible that clinical trials evaluating antipsychotic and antimanic compounds may be unnecessarily long in duration. All of these problems raise ethical as well as clinical concerns.

Although we subscribe to the need for randomized controlled trials to evaluate efficacy for new compounds, reassessing the types of clinical endpoints in such trials might be appropriate. Therefore, we set out to define "exit" criteria or novel clinical endpoints that might help to assess the efficacy of antimanic compounds in a way that would be safer, more economical, and less sensitive to nonspecific factors in the clinical environment while still allowing unambiguous assessment of efficacy. Our concerns with safety had to do with the reduction of unnecessary patient exposure to placebo or ineffective compounds, as well as to ongoing severe symptomatology that, in itself, may carry significant clinical risks. With respect to the efficiency of such trials, our goal was to

**Table 1. Characteristics of 49 Patients With Manic or Mixed Episodes[a]**

| Characteristic | Value |
|---|---|
| Age, y | 34.5 ± 12.0 (32.0) |
| Men, N (%) | 20 (41) |
| White, N (%) | 43 (88) |
| Never married, N (%) | 32 (65) |
| No. of manic episodes | 5.0 ± 4.8 (3) |
| No. of depressive episodes | 4.2 ± 4.0 (3) |
| Age at first manic episode, y | 24.5 ± 8.4 (22) |
| Age at first depressive episode, y | 21.5 ± 7.1 (20) |

[a]Values shown as mean ± SD (median) unless otherwise specified.

develop methods that would facilitate testing of new compounds with smaller groups of patients that could be assessed in trials of shorter duration. Furthermore, we believe that the definition of "exit" criteria, i.e., criteria for removing patients very early in a trial in the face of indicators that they are unlikely to respond later in the trial, would have the advantage of producing results that are less subject to nonspecific factors in the treatment environment. The shorter the duration of exposure to the clinical environment, particularly that of an inpatient unit with a high staff-to-patient ratio and multiple therapeutic activities, the less likely it is that the patient is responding to factors in the environment rather than to the compound being tested. Theoretically, this approach could lead to a more robust separation between responders and nonresponders, thus requiring fewer subjects to achieve adequate power, while decreasing the length of the trial. In addition to avoiding unnecessary exposure to ineffective treatments, such an approach could provide a more favorable balance between the costs and benefits of indication trials, both for the subjects involved in such trials and for the companies sponsoring them.

In the context of studies of acute mania, we thought that study exit criteria might be defined by clinical deterioration or failure to improve by a set percentage on an appropriate mania rating scale within a set time frame. To define such exit criteria, we conducted a detailed database and chart review of relevant ongoing studies of bipolar disorder being conducted under the auspices of the Depression and Manic Depression Prevention Program at the Western Psychiatric Institute and Clinic in Pittsburgh (Pa.) and then made use of new statistical techniques to estimate later treatment response on the basis of clinical ratings of response early in the treatment of an acute manic episode.

## METHOD

Subjects were included in the analyses if they had entered 1 of 2 ongoing research protocols sponsored by the National Institute of Mental Health (Maintenance Therapies in Bipolar Disorder, MH-29618, Dr. Frank, principal investigator; Pharmacotherapy of Treatment-Resistant Mania, MH-50634, Dr. Mallinger, principal investigator). Subjects participating in either of these protocols, who were hospitalized for a manic or mixed episode according to DSM-IV during their treatment, provided the data for the present report. Data were retrospectively collected from the hospital charts during the admission stay of all subjects. Among all admissions for a manic or mixed episode, we identified 76 admissions, experienced by 49 patients for whom sufficient data were available to be included in the development of the exit criteria. Of these 49 patients, 32 experienced only 1 admission, 10 experienced 2, 4 patients experienced 3 admissions, and 3 patients experienced 4 admissions during the period covered by our chart review. Demographic and clinical characteristics of these 49 patients are displayed in Table 1. Subjects were predominantly young adults of white background with multiple previous manic and depressive episodes. Fifty-five percent of the admissions were voluntary. The mean length of stay was 3.0 ± 1.6 weeks, with a median of 2.9 weeks. Medication equivalence was obtained by using the relative oral potency values for each drug from the *Psychotropic Drug Handbook*.[1] Medication equivalence was computed for benzodiazepines and neuroleptics separately for each inpatient week. During week 1, 57 subjects took benzodiazepines with a mean equivalence of 2.98, and 67 subjects took neuroleptics with a mean equivalence of 3.26. The number of subjects receiving benzodiazepines during week 2 dropped to 47 with a mean equivalence of 2.49, and 62 took neuroleptics with a mean equivalence of 4.62 (numbers of patients take into account all 76 admissions and thus exceed 49; also, some subjects took multiple drugs).

Information included in the development of the exit criteria included twice-weekly Bech-Rafaelsen Mania Scale[2] ratings for the first 14 days of admission. The Bech-Rafaelsen Mania Scale is a 12-item instrument used to measure manic symptoms on the day of assessment. A total score of 15 or greater indicates a syndromal mania. The Bech-Rafaelsen ratings were made by research staff for any patient identified as a potential protocol participant. A subset (≈ 35%) of the Bech-Rafaelsen ratings were made retrospectively via chart review (as opposed to direct patient interview) and were completed prior to any grouping of subjects for the present report. For some calculations, as noted, the Bech-Rafaelsen ratings were summarized as the mean across each week; daily scores were used in the regression analysis.

### Statistical Methods

The changes between week 1 and week 2 were compared with paired t tests. In the case of subjects with more than 1 admission, it was assumed that each admission was an independent observation. We fit a mixed-effects regression model (see Gibbons et al.[3] for a review of the model with application in the context of psychiatric research) to

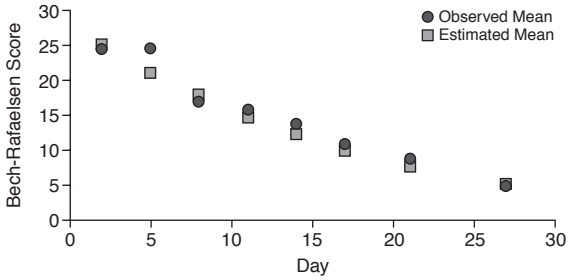
all available data from each admission through day 28. The mixed-effects regression model is an alternative technique that can be used to impute missing observations from longitudinal studies. Like the last-observation-carried-forward (LOCF) technique, missing observations are assumed to be random and ignorable.[4] The mixed-effects regression model is preferred over the LOCF since it allows us to model the change in manic symptoms over time instead of assuming a plateau of treatment effect.[5]

The primary objective of this analysis was to estimate the final Bech-Rafaelsen scores that were available for each subject (up to day 28) from all available data between day 0 and day 11. Day 11 was chosen because it was early enough to be used as a predictor of short-term outcome and far enough along in the treatment period to provide the maximum amount of data in our sample. Day 28 was chosen as the admission endpoint because 4 weeks is typical for mania treatment. The model included an intercept and linear and quadratic trend coefficients to describe the overall temporal pattern. All terms in the model were random such that the time trend for each admission had unique (i.e., person-specific) coefficients. Using the estimated model coefficients (i.e., mean trend parameters and variance-covariance matrix of the person-specific deviations from the mean trend parameters), we then obtained empirical Bayes (EB) estimates of each subject's trend coefficients[6] based on all available data for each admission and a second set of EB estimates based only on the available data from each admission through day 11. As described by Laird and Ware[6] and Gibbons et al.,[3] in using this approach each subject can be measured on a different number of days in the first 11 days and even on different days during this time period, which is a distinct advantage of the application of the mixed-effects model in this context. Using these EB estimates of the person-specific trend, we can then estimate the predicted Bech-Rafaelsen score at any point in time. The best estimate is based on all available data from each admission (i.e., through day 28). The analysis provides a measure of the accuracy of prediction of outcome at day 28 from data collected on days 1 to 11. To assess the validity of the model, we ran Pearson correlations between the EB estimated score and the actual final day score. To check the validity of the model in subjects remitting earlier in their treatment, we reran this analysis on a subset of 24 subjects whose actual final scores were recorded between day 9 and 14.

## RESULTS

Comparisons of mean ± SD clinical ratings for weeks 1 and 2 on the Bech-Rafaelsen scale showed a change from 23.6 ± 7.7 to 14.6 ± 9.3 at the end of the second week ($p = .001$). These comparisons reveal substantial differences between weeks 1 and 2 that are both clinically and statistically significant. Encouraged by these results, we went on to examine the question of whether Bech-Rafaelsen rating scale scores obtained before the end of the second week of treatment could be used to predict response status and thus lead to the development of exit criteria. The criterion for response was a score of less than 8 on the Bech-Rafaelsen scale, and the criterion for nonresponse was a score of 14 or above. We used a quadratic model to accommodate some nonlinearity toward day 28 for the Bech-Rafaelsen scores (see Figure 1). We used the Bech-Rafaelsen score through day 11 to predict those patients who would respond preliminarily to antimanic drug treatment. In this step, the nonresponders according to the Bech-Rafaelsen scale would be dropped from the trial.

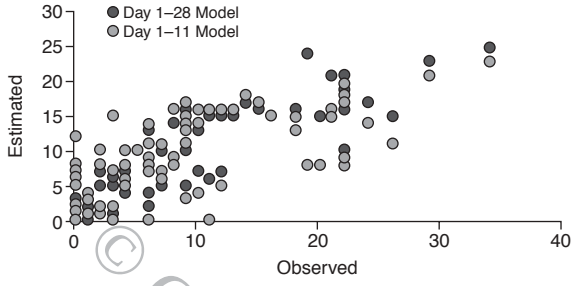
**Figure 1. Observed Versus Estimated Mean Bech-Rafaelsen Mania Scale Scores Over Time Using Day 1–11 Model**




### Analysis of Bech-Rafaelsen Mania Scores

The analysis of the available Bech-Rafaelsen data through day 28 from the 76 admissions revealed that there was very little person-specific variability in the quadratic trend coefficients; therefore, the quadratic term was treated as a fixed parameter in the model. The mean intercept was 27.99, linear trend was –1.42 (Bech-Rafaelsen units per day), and the quadratic term was 0.022 (squared Bech-Rafaelsen units per day). The mean ± SD variance in intercepts was 52.83 ± 7.26), the mean variance in linear trends was 0.06 ± 0.25, and the correlation between intercepts and slopes was $r = –.52$. Figure 1 reveals that the model fits the average level response process extremely well. Restricting the EB estimates to only the available data during the first 11 days showed some degradation in predictive value ($r = .67$ between the day 11 EB estimates of the final score and the actual final score and $r = .86$ with the EB estimate that used all available data through day 28). Figure 2 shows observed versus estimated Bech-Rafaelsen Mania Scale final score using day 1–11 and day 1–28 models. Using all available data, the correlation between the EB estimated score and the actual final day score (up to day 28) was $r = .87$. In a subset of 24 subjects with final score recorded between days 9 and 14, the correlation between the observed values and the 28-day estimate was $r = .69$;

**Figure 2. Observed Versus Estimated Bech-Rafaelsen Mania Scale Final Score Using Day 1–11 and Day 1–28 Models**




the correlation with the 11-day estimate was r = .89. This indicates that the model fits the observed data quite well at an individual subject level as well.

To further examine predictive validity, we categorized subjects' admissions into 3 groups based on status at day 28: responders, Bech-Rafaelsen scores of 0 to 7; partial responders, Bech-Rafaelsen scores of 8 to 13; and nonresponders, Bech-Rafaelsen scores of 14 or more (Figure 3). Inspection of Figure 3 reveals that using the predicted final Bech-Rafaelsen score from data obtained during the first 11 days of treatment, only 2 of 23 predicted nonresponders proved to be responders (marginal homogeneity test, exact p = .0232). Conversely, of the 31 predicted responders, none were actually nonresponders. Twenty-seven of the 31 predicted full responders became full responders. Sixteen of the 22 predicted partial responders became partial or full responders. Combining the groups of responder and partial responder, among 53 subjects predicted to be in one of these categories, only 6 actually did not respond. Interestingly, we found no significant differences among the 3 groups on demographic or clinical history characteristics, suggesting that early treatment response data alone may be sufficient to predict later response.

**Figure 3. Estimated Versus Actual Response Defined by Bech-Rafaelsen Mania Scale Scores for Patients Comprising 76 Admissions[a]**

| Estimated Final Bech-Rafaelsen Response Based on Data From Days 1–11 | Actual Final Bech-Rafaelsen Response: Nonresponder (≥ 14) | Partial Responder (8–13) | Full Responder (0–7) | Total |
|---|---|---|---|---|
| Nonresponder (≥ 14) | 12 | 9 | 2 | 23 |
| Partial responder (8–13) | 6 | 4 | 12 | 22 |
| Full responder (0–7) | 0 | 4 | 27 | 31 |
| Total | 18 | 17 | 41 | 76 |

[a]Values represent number of patients in each response group; ranges in parentheses designate Bech-Rafaelsen scores that correspond to the response groups. Marginal homogeneity test, exact p = .0232.

## DISCUSSION

This chart-review study sought to determine whether it might be possible to define exit criteria for conducting future trials assessing the efficacy of antimanic compounds. Such exit criteria would enable the early identification and removal of subjects unlikely to demonstrate a subsequent response (and ultimately a full remission) to the treatment to which they were assigned in a randomized controlled trial. Initial results from our efforts to develop such exit criteria for trials of antimanic compounds appear very promising. Also interesting was the fact that when grouped by full response, partial response, and nonresponse, the subject groups did not differ from each other in demographic or clinical history characteristics. Perhaps most important and encouraging, we found that early response based upon data at day 11 predicts final response and remission.

Also noteworthy is the very high sensitivity and acceptable specificity of the early response/ultimate remission prediction. Particularly encouraging, from our point of view, is the very low likelihood (2 of 76 admissions examined) that someone predicted to be in the nonresponse group will go on to a clear remission of symptoms as a full responder. Thus, removing someone early in a trial is unlikely to deny him or her access to a treatment that would have ultimately brought about a clear remission of symptoms.

Our results suggest that relatively simple, easily obtained information allows one to define patients unlikely to respond before the end of week 2 of treatment and that there is a strong relationship between failure to respond early in treatment and subsequent failure to remit. Thus, patients whose early treatment is not characterized by a trajectory of rapid decline in Bech-Rafaelsen scores could confidently be removed early in trials.

There are a number of caveats in interpreting these results. First, the study population was relatively small in size. Second, the data come from a tertiary referral setting in an academic medical center with a relatively high nurse-to-patient ratio that may not be representative of all settings in which antimanic compounds are evaluated. In this report, about 35% of the ratings were derived from a retrospective chart review. However, it should be noted that the ratings relevant to the exit criteria are easily completed and represent only a formal method for recording the minimum clinical monitoring necessary for the care of manic or mixed patients. Finally, it should be noted that there was no standardization of the pharmacotherapy received by the patients in this investigation, although the

treatment environment was highly comparable across all treatment episodes studied. Nonetheless, we must acknowledge the possibility that different criteria would have emerged had all patients received the same compound on a standardized dosing schedule.

The criteria we propose would, of course, benefit from prospective testing in a new study of a single antimanic compound or, ideally, in a series of studies of a variety of compounds for demonstrating drug-placebo differences. We note also that the use of stringent algorithms for dose escalation and for the use of rescue medication could make these parameters more useful in distinguishing early failure of treatment.

A further caveat is the complexity of studying bipolar disorder in general. To date, we do not have a good understanding of the relationship between early response and subsequent sustained remission. While investigators have explored such relationships in unipolar disorder,[7] no such investigations of bipolar disorder have been carried out. The situation is complicated by the fact that when a patient evidences a remission of mania as measured by a mania rating scale, he or she may or may not be free of mood symptoms and, indeed, may be progressing into depression. Our criteria begin to approach this critically important feature of mania treatment.

The general model we propose for exit criteria might well be applied to other psychotic conditions. The further development of such methods will require innovative working relationships among investigators, the pharmaceutical industry, and the U.S. Food and Drug Administration to bring safe, efficacious drugs to the market rapidly.



## REFERENCES

1. Perry PJ, Bruce A, Liskow BI. Psychotropic Drug Handbook. 7th ed. Washington, DC: American Psychiatric Press; 1997
2. Bech P, Bolwig TG, Kramp P, et al. The Bech-Rafaelsen Mania Scale and the Hamilton Depression Scale. Acta Psychiatr Scand 1979;59:420–430
3. Gibbons RD, Hedeker D, Elkin I, et al. Some conceptual statistical issues in analysis of longitudinal psychiatric data: application to the NIMH Treatment of Depression Collaborative Research Program dataset. Arch Gen Psychiatry 1993;50:739–750
4. Little R, Rubin DB. Statistical Analysis With Missing Data. New York, NY: John Wiley & Sons; 1987
5. Lachin JM. Statistical considerations in the intent-to-test principle. Control Clin Trials 2000;21:167–189
6. Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics 1982;38:963–974
7. Quitkin FM, McGrath PJ, Stewart JW, et al. Chronological milestones to guide drug change: when should clinicians switch antidepressants? Arch Gen Psychiatry 1996;53:785–792