Challenges in the development of clinical trials for major depressive disorder: lessons learned from trials in minor depression

Mark H. Rapaport, MD; Rachel E. Maddux, BA

This paper reviews some of the challenges faced by individuals who design and implement clinical trials of potential antidepressant medications. Particular emphasis is placed on questioning the validity of some of the theoretical assumptions that form the underpinnings of most conventional trials. Work from our group developing clinical trial methodology for minor depression is used as an example of how alternate constructs may be helpful to differentiate drug-placebo differences.

Dialogues Clin Neurosci. 2002;4:402-407.

Keywords: *clinical trial design; clinical methodology; antidepressant; major depressive disorder; depression; minor depression*

Author affiliations: Department of Psychiatry, Cedars-Sinai Medical Center, Los Angeles, Calif, USA (Mark H. Rapaport); Department of Psychiatry, University of California, Los Angeles, Los Angeles, Calif, USA (Mark H. Rapaport, Rachel E. Maddux)

Address for correspondence: Mark H. Rapaport, MD, Department of Psychiatry, Cedars-Sinai Medical Center, 8730 Alden Drive, Thalians, Room C301, Los Angeles, CA 90048, USA
(e-mail: Mark.rapaport@cshs.org)

Over the last decade, there has been increasing attention focused on the inadequacy of the current methodology employed in randomized clinical trials involving new antidepressant medications. The primary focus of this concern has centered on the need to adequately differentiate the effectiveness of new treatments from the placebo condition. There has been considerable consternation because of the increasing rate of placebo response seen in all types of trials in psychiatry, particularly trials of mood and anxiety disorders.[1-3] This growing awareness has led to a variety of different efforts that have begun to address concerns about trial design and methodology.[4-6] These include an ongoing series of workshops sponsored by the National Institute of Mental Health (NIMH) and the New Clinical Drug Evaluation Unit (NCDEU).[7] The NIMH has also hosted a series of consensus conferences over the last few years in an attempt to begin to focus attention on these concerns. Such conferences have investigated issues including placebo and placebo response and the development of new instruments for the assessment of mood and anxiety disorders. There has also been a series of international meetings, including a symposium held in Rhodes, Greece in 2000, which brought together international experts in methodology with senior staff from the NIMH and the Food and Drug Administration (FDA). The culmination of these concerted efforts was a consensus statement that was published in *Neuropsychopharmacology* in 2002.[8] The Rhodes panel identified 4 critical problem areas: (i) the nature of the patient sample; (ii) the limitations of behavioral methods and analyses used for assessing treatment-related improvement and recovery; (iii) the lack of consensus about standards for determining speed of onset and action for medications; and (iv) the failure to integrate advances into our knowledge about depression in antidepressant

development with current clinical trial design. The topics requiring greater emphasis include concerns about the validity of our current diagnostic nosology, as well as questions about how diagnoses are made. There are also questions about the best way to assess the severity of psychiatric syndromes. Our current standard is to use psychometric rating scales. However, many times these scales only reflect one dimension of a complex illness. Another critical issue is the number, as well as the length, of the evaluations to be performed. A related issue of concern is the total length of time that is given to the evaluation of the active treatments. One of the major recurrent challenges faced in medication development is ensuring that the trials are adequately powered in order to differentiate relatively subtle differences. Very often power calculations are not based on empirical data, but rather reflect the aspirations of the trial design planners.

Assumptions made about the sample for the study often end up greatly influencing the trial design. These assumptions are made in order to facilitate the use of relatively simple inferential statistical models. However, some of these assumptions reflect lack of thought about the psychiatric syndromes. One of the intrinsic assumptions made in the design of trials is that the sample being analyzed will be relatively homogeneous. We frequently attempt to control for age, ethnicity, length of illness, comorbid diagnosis, and comorbid medical factors. We frequently do not go to great lengths to determine that the subjects being evaluated truly have a similar disorder. Yet, any clinician will readily attest that patients with depression in clinical practice clearly respond differently to the same medication and, in some cases, do not respond at all.[9-14] This suggests that there is considerable heterogeneity within the group of individuals who have major depressive disorder. Furthermore, clinicians can certainly confirm that the same medication given to different individuals may produce very different side-effect profiles for each of those individuals. Even simple clinical observation suggests that we are dealing with a heterogeneous syndrome when we discuss major depressive disorder. An overview of any large clinical trial's database will demonstrate that improvement is not uniform for subjects receiving an active, effective treatment. Some individuals get markedly better, while many individuals do not improve at all during a standard antidepressant trial.

The representativeness of the sample poses another concern. After the advent of the *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition*, the concept of comorbidity was given much greater weight. Prior to that, a hierarchical approach to diagnosis was used. The emphasis on the presence of comorbid disorders led to the development of rigorous inclusion and exclusion criteria for most studies. Although there is little empirical evidence that supports the use of most of these inclusion and exclusion criteria, they have become standardized, and in many cases, quite limiting. However, it should be noted that many of these criteria seem to be developed as part of a response to perceived expectations by regulatory agencies such as the FDA and the European regulatory authorities. Nevertheless, these criteria end up limiting the representativeness of the sample being investigated. The majority of individuals suffering from the syndrome are excluded from participation in these trials. Therefore, we have limited information about the generalizability of either positive or negative results to the syndrome in general.

A factor that is rarely discussed is the lack of stability inherent in most of these syndromes. Most clinical trials use one rating scale as a primary measure of success. Therefore, the trial measures only a limited aspect of that syndrome. A second assumption that is made in the design of the trial and the treatment of the disorder is that the disorder itself will be relatively stable if no intervention is made. Unfortunately, this is a fallacious assumption. Some individuals demonstrate significant week to week variation in ratings measures, independent of any type of treatment intervention. This intrinsic fluctuation associated with the disorder makes it difficult to discern what degree of change can be attributed to either the placebo condition or the active treatment condition. This has led to a reductionistic approach to analysis of subjects and trials, where all responses and changes are essentially attributed to either the active treatment condition or the placebo condition.

Lessons learned from clinical trials investigating minor depressive disorder

One can use randomized clinical trials in minor depressive disorder as a case study to emphasize some of the challenges faced in trial design and possibly some solutions to these challenges. Minor depressive disorder is an area where there is no consensus about its conceptualization or definition. Some individuals believe that minor depression is merely a segue into major depressive disorder, while others consider minor depression an entity in itself.[15-17] Some individuals worry that investigating minor depression trivializes the core concept of major depressive dis-

order, while others consider it an important part of the spectrum of depressive syndromes.[18] Even among those who believe that minor depression is a valid concept that requires rigorous investigation, there is considerable debate about what the definition of minor depression is or should be.[19] Furthermore, there is little empirical evidence to support any of the currently employed definitions. Many of the older clinical trials investigating minor depression actually grouped patients into cohorts that contained individuals with major depressive disorder described as being mild in severity. Some of these trials did not differentiate between major depressive disorder and a diagnosis of minor depression, but merely stated that those with lower Hamilton Depression Rating Scale (HAMD) scores should be considered as having minor depression. Other trials combined patients with major depression of a milder form with Research Diagnostic Criteria (RDC) patients with minor depression. Older trials employed either tricyclic antidepressant medications or antipsychotic medications. It is not surprising, based on the side-effect profiles of these agents and the weighting of the HAMD towards somatic concerns, that it was difficult to differentiate an active treatment response from a placebo response. A second challenge that studies of minor depression emphasize is the use of rating scales that were developed at another time and for another diagnostic entity to assess minor depression. All of the older studies used the HAMD 17 as a primary outcome measure.[20] As discussed above, this rating scale, developed to assess inpatients with endogenous depression, is heavily weighted toward somatic and/or vegetative factors. This makes the HAMD a very coarse instrument to use for individuals with milder forms of depression or minor depression, since neither somatic nor vegetative symptoms are highly prominent in such patients. Furthermore, these less highly prominent symptoms tend to be transient in presentation and thus may vary greatly from week to week on a rating scale. This emphasizes the importance of carefully ensuring that the methods of assessment fit the most relevant signs of the syndrome being studied. Very often, both government and industry have been willing to commit a large percentage of limited resources to clinical trial research, without considering the appropriateness of the measures in assessing the full scope of the syndrome.

Another concern highlighted by investigations of minor depression is the lack of objective measures of either functional or quality of life impairment. This problem is also true for most studies of most psychiatric disorders. Thus, in spite of the fact that the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV)* requires functional impairment or quality of life impairment to be present in order for a diagnosis of the syndrome to be made, there have been few efforts to establish some type of criteria for quality of life or functional impairment with these disorders.[21] It has been shown in primary care studies that many people who seem to meet criteria for psychiatric syndromes have spontaneous remissions when followed longitudinally. This may well reflect the inclusion of individuals who, because of life stress, have a particular series of signs and symptoms, but in actual fact do not have the pathology associated with a lifelong syndrome. As would be expected, the result of not paying attention to these challenges when designing clinical trials is that the trials tend to be uninformative, if not misleading.

In contrast to some of the problems identified above, a consortium of investigators at the University of California, San Diego, the University of Texas Southwestern, Western Psychiatric Institute and Clinics, and Eli Lilly conducted a multisite trial of minor depressive disorder (Judd et al, manuscript submitted). In order to deal with the concerns about the diagnosis of minor depression, the following criteria were used to operationalize our definition: (i) a subject had to have dysphoria and anhedonia plus at least one additional symptom of major depressive disorder from a *DSM-IV* checklist, or dysphoria or anhedonia and two additional symptoms of major depressive disorder; (ii) a clear-cut functional disability as evidenced by a Global Assessment of Functioning (GAF) score of less than 70 and Medical Outcome Survey (MOS) subscale score of less than 75 for social functioning, and of less than 67 for emotional role functioning.[22,23] In developing these criteria, we recognized that they were rather arbitrary and thus felt it was necessary to be rigorous and precise with our definition of what the syndrome was. We deliberately decided to include individuals with a past history of major depressive disorder or dysthymia, as long as they had been in remission for at least 2 years prior to developing their current episode of minor depression. Furthermore, we required individuals to have had minor depression for a minimum of 1 month prior to entering the trial. We deliberately did not use a longer period than 1 month, since it is difficult to gather accurate retrospective information about the presence of minor symptoms. However, in order to compensate for concern that our definition of minor depression was merely a way station for individuals going into major depressive disorder or recovering from major

depressive disorder and becoming euthymic, we included a 4-week, single-blind, placebo lead-in phase to the study. This caused a second dilemma: which individuals would we exclude from the study? Since all of the rating scales we employed were validated for major depressive disorder and not minor depression, it was difficult to know how individuals would score on these measures. Furthermore, we did not have any data to suggest to us what the range of symptomatology should be for individuals with minor depressive disorder on these rating measures. Therefore, we decided to require that individuals meet the same entrance requirements for 3 out of 4 weeks in the placebo phase, including the last 2 weeks prior to randomization, in order to enter the study. This facilitated getting an accurate sense of the types of changes one would see on the ratings scales, independent of their having a bearing on whether or not individuals were able to enter the double-blind portion of the study.

Some of the other important features of this trial included the use of the Inventory of Depressive Symptomatology—Clinician Rated (IDS-C) as well as three different forms of the HAMD (17-, 21-, and 28-item), the Hamilton Anxiety Rating Scale, the GAF, and the MOS Short Form (SF-36).[20,22] The IDS-C was identified as the primary outcome measure because of the unique features of the scale. First, this scale encompasses a much broader range of depressive symptomatology, extending from various psychological symptoms through somatic symptoms. Second, this symptom scale attempts to quantify, in a uniform way, both the severity and the intensity of symptomatology. Yet, we were not comfortable merely using existing rating scales as a way of assessing response in this trial. Therefore, we also investigated the effects of active treatment on complete resolution of symptoms of depression, plus resolution of functioning. This is a remarkably high bar to attempt to overcome.

Another crucial feature, as described earlier in this article, is the length of the evaluation. Many early randomized clinical trials were 2- to 4-week placebo-controlled trials.[23] Over time, trials have extended to 6 to 8 weeks' duration. Yet, as is clearly emphasized by the work of Stassen and colleagues and others, many individuals with major depressive disorder are just beginning to reach recovery at the 8- to 12-week time points.[24] Therefore, we elected a 12-week acute trial particularly because we were interested in determining the number of individuals that met remission criteria, as well as a change in rating scale. The primary input, again, was the change in the IDS-C with the major outcome point being the ability to achieve complete remission for 1 month prior to the end of the trial.

Since we were dealing with minor depression, it created a series of opportunities that we felt we had to explore in order to gather pilot data if further investigations were warranted. One of the other major questions was: what happens if one allows individuals to undergo an extended period of time on placebo (ie, 4 months)? Will this impact response to pharmacotherapy? A second question was: is acute treatment of minor depression sufficient? Will individuals who respond acutely require continuation treatment, as is the case with major depressive disorder? Additionally, what is the course of untreated minor depression for individuals who participate in a trial? Are we placing these people at an increased risk or burden by their continued presence in the trial while on placebo? In order to gather pilot data to begin to answer these questions, individuals who completed the initial 12 weeks of the trial entered a continuation phase. The randomization of individuals for the acute and continuation phase of the trial were performed at the initial point of randomization, rather than a second re-randomization, after completion of the acute trial. Therefore, individuals in this trial were randomized both to an acute phase and maintenance treatment with either fluoxetine or placebo and to one of four continuation phase conditions: fluoxetine-fluoxetine, fluoxetine-placebo, placebo-placebo, or placebo-fluoxetine. Analysis of the continuation phase of the study was a priori specified to be exploratory, because we knew that sizes of the cells would not be sufficient to answer these questions.

There were several features during the analysis plan that were unique. First was the realization that minor depression was most likely a heterogeneous syndrome. Therefore, we acknowledged the need to investigate the relationship between minor depression and a previous history of major depressive disorder and dysthymia, and also the relationship between minor depression and a family history of psychiatric disorders. In an attempt to more thoroughly utilize the data that would be gathered in this study, we decided that a mixed regression model would be more powerful than a standard analysis of variance of statistical approach. However, since the random regression model is not as accepted in psychiatric literature, we specified in the initial data analysis plan that both types of analyses be performed. A third aspect of this study was the evaluation of the categorical end point (ie, full remission

of symptoms and return of functioning), as well as the parametric end points.

One can use the design of this trial in minor depression to address a number of the challenges that we had earlier identified. This trial is a good example of the type of consensus thinking process that can be used to enhance diagnostic rigor and assessment of severity of illness. Furthermore, it highlights the need to carefully review the items within assessment instruments, in order to ensure that the instruments are as useful as possible for the disorder. This study also highlights the type of thought that should go into determining the evaluation for a study. Since minor depression was an unknown entity at the time, one of the key questions was: how stable is this entity over time? Although 4 weeks is an arbitrary length of time, we felt for ethical and scientific reasons, that it would be an adequate length of time for an extended placebo run-in.

Conclusions

A careful and critical review of clinical trial methodologies is imperative for the field to move forward. Attention to many of the assumptions that are inclusively made when a trial is designed will be critical in enhancing the success of clinical trials. We must think closely about the diagnostic criteria used in the trial in the inclusion and exclusion criteria. Many of the currently accepted criteria limit the generalizability of the findings and have not been demonstrated in a systematic fashion to enhance differentiation of drug versus placebo responses. Yet, some important aspects of the very definition of these syndromes have been neglected, in particular, the importance of including functional disability and quality of life dysfunction as part of the definition of the syndrome. Some individuals may present with a requisite number of symptoms, but may not be as adversely affected as if they had had a profound, long-lasting syndrome. It is quite likely they are suffering from a transient constellation of symptoms due to an external stressor. A second important concern is the appropriateness of the assessments that are being used in randomized controlled trials. Very often, the assessments that are employed represent "me too" assessments, because studies done by other companies have used the measures in the past. Yet, this may not reflect our best knowledge about the disorder being studied, nor a sufficient way of bringing a new compound onto the market. Frequently, the argument for the use of such instruments is that they are supposedly mandated by regulatory agencies. However, more often than not, this is a myth that is perpetuated rather than the outcome of frank and careful discussions with the regulatory authority. A third important issue that requires some thought is assumptions about the stability of the syndrome over time. Many times, studies are designed with the assumption that randomization to placebo should lead to a relatively static or, if anything, disadvantageous course for patients. Yet, investigation of most medical syndromes suggests that there is an intrinsic waxing and waning to the course of the syndrome. Therefore, arbitrary assessment using instruments that investigate only one aspect of the syndrome may well lead to spurious results.

A last concern, but one that can greatly influence a trial, involves appropriate statistical design. Often studies are powered based on desire, rather than available data. A second concern is that often a new statistical design represents the easiest or safest design, rather than a design that is most likely to produce informative results.

In conclusion, it is clear that there is tremendous opportunity to improve the design and methodology used in randomized clinical trials. The recognition of these challenges by the NIMH, the FDA, the European regulatory authorities, as well as industry, implies that important future change is likely to occur. ❑

REFERENCES

1. Khan A, Leventhal RM, Khan S, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol*. 2002;22:1-6.
2. Bowder CL, Calbrese JR, McElroy SL, et al. A randomized , placebo-controlled 12-month trial of divalproex and lithium in treatment of outpatient with bipolar I disorder. *Arch Gen Psychiatry*. 2000;57:481-489.
3. Fisher S, Lipman RS, Uhlenhuth EH, Rickels K, Park LC. Drug effects and initial severity of symptomatology. *Psychopharmacologia*. 1965;7:57-60.
4. Katz MM, Koslow SH, Berman N, et al. Multivantaged approach to the measurement of behavioral affect states for clinical and psychobiological research. *Psychol Rep*. 1984;55:619-671.
5. Leon AC. Measuring onset of antidepressant action in clinical trials: an overview of definitions and methodology. *J Clin Psychiatry*. 2001;62(suppl 4):12-16.

Desafíos en el desarrollo de ensayos clínicos para el trastorno depresivo mayor: lecciones aprendidas de los ensayos en la depresión menor

Este artículo revisa algunos de los desafíos que afrontan aquellos investigadores que trabajan en el diseño y la realización de ensayos clínicos con medicamentos con potencial antidepresivo. Se cuestiona con especial énfasis la validez de algunos de los supuestos teóricos que forman las bases de los ensayos más convencionales. Se utiliza el trabajo de nuestro grupo que desarrolla metodología de ensayos clínicos para la depresión menor como un ejemplo de cómo los constructos alternativos pueden ser útiles para distinguir las diferencias fármaco-placebo.

Défis posés par le développement des essais cliniques sur les troubles dépressifs majeurs : leçons tirées des essais sur les dépressions légères

Cet article passe en revue quelques-uns des défis lancés à ceux qui conçoivent et réalisent des essais cliniques pour évaluer l'efficacité de nouveaux médicaments antidépresseurs. Il a été particulièrement insisté sur la mise en question de la validité de quelques-unes des hypothèses théoriques qui étayent la plupart des essais classiques. Le travail de notre groupe qui développe une méthodologie d'essai clinique pour les dépressions légères illustre comment l'utilisation de concepts différents peut permettre de distinguer les différences médicament-placebo.

6. Katz MM. Need for a new paradigm for the clinical trials of antidepressants. *Neuropsychopharmacology*. 1998;19:517-522.
7. Stahl SM, Shayegan D. Reducing measurement variability in psychopharmacology applying principles of adult education by utilizing multimedia to facilitate changes in rater behavior. New Clinical Drug Evaluation Unit Abstracts; May 28-31, 2001; Phoenix, Ariz.
8. Katz MM, Halbreich UM, Bowden CL, et al. Enhancing the technology of clinical trials and the trials model to evaluate newly developed, targets antidepressant. *Neuropsychopharmacology*. 2002;27: 319-328.
9. Boyer WF, Feighner JP. Clinical significance of early non-response in depressed patients. *Depression*. 1994;2:12-35.
10. Coryell W, Coppen A, Zeigler VE, Biggs J. Early improvement as a predictor of response to amitriptyline and nortriptyline: a comparison of 2 patient samples. *Psychol Med*. 1982;12:135-139.
11. Dunbar GC, Cohn JB, Fabre LF, et al. A comparison of paroxetine, imipramine and placebo in depressed outpatients. *Br J Psychiatry*. 1991;159:394-398.
12. Katz MM, Koslow SH, Maas JW, et al. The timing, specificity and clinical prediction of tricyclic drug effects in depression. *Psychol Med*. 1987;17:297-309.
13. Katz MM, Koslow SH, Frazer A. Onset of antidepressant action: Reexamining the structure of depression and multiple drug actions. *Depress Anxiety*. 1997;4:257-267.
14. Laska EM, Siegel C. Characterizing onset in psychopharmacological clinical trials. *Psychopharmacol Bull*. 1995;31:41-44.
15. Horwath E, Johnson J, Klerman GL, Weissman MM. Depressive symptoms as relative and attributable risk factors for first-onset major depression. *Arch Gen Psychiatry*. 1992;49:817-823.
16. Wells KB, Burnam MA, Rogers W, Hays R, Camp P. The course of depression in adult outpatients. Results from the Medical Outcomes Study. *Arch Gen Psychiatry*. 1992;49:788-794.
17. Maier W, Gansicke M, Weiffenbach O. The relationship between major and subthreshold variants of unipolar depression. *J Affect Disord*. 1997;45:41-51.
18. Angst J, Merikangas K. The depressive spectrum: diagnostic classification and course. *J Affect Disord*. 1997;45:31-39.
19. Rapaport MH, Judd LL, Schettler PJ, et al. A descriptive analysis of minor depression. *Am J Psychiatry*. 2002;159:637-643.
20. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56-62.
21. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*. Washington, DC: American Psychiatric Association; 1994.
22. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The inventory of depressive symptomatology (IDS): psychometric properties. *Psychol Med*. 1996;26:477-486.
23. Montgomery S. Are two-week trials sufficient to indicate efficacy? *Psychopharmacol Bull*. 1995;31:29-35.
24. Stassen HH, Delini-Stula A, Angst J. Time course of improvement under antidepressant treatment: a survival analytic approach. *Eur Neuropsychopharmacol*. 1993;3:127-135.