How Should Efficacy Be Evaluated in Randomized Clinical Trials of Treatments for Depression?

Michael E. Thase, M.D.

The present system of conducting studies of promising antidepressant therapies has evolved through the collaborative efforts of government, industry, and academicians and is costly and inefficient. At least one third of the published clinical trials of approved antidepressants are negative for efficacy, which can be partly explained by the clinical and neurobiological heterogeneity of the depressive disorder and partly because of methodological inadequacies. Unfortunately, too little attention is given to ensuring the reliability of diagnoses and dependent measures, sample sizes are seldom large enough to detect modest yet honestly significant differences, and too many trials are pursued before dose-response characteristics are fully understood. At present, the only data beyond 1 year of treatment—and the only evidence about protection against recurrent depression—come during postmarketing or phase 4 of the drug development process. Moreover, efficacy data for depressed children and adolescents, bipolar depression, psychotic depression, dysthymia, and frail or medically ill elderly patients are rarely available at the time a drug is introduced. Thus, it is remarkable how little clinicians know about a new antidepressant at the time it is first approved for general use. Within a research strategy, tactics that ensure reliability, encourage attention to adherence, and lessen attrition at the outset of a study will increase the power and design sensitivity of a particular trial. Additionally, the issues of research funding—including division of the research pie—and the relationship of the Food and Drug Administration and investigators to the pharmaceutical industry and the National Institute of Mental Health need to be revisited. Finally, extension of a compound's patent life might be considered to expand the necessary postmarketing research. This article describes the process of conducting the clinical trials that support a New Drug Application, discusses issues in evaluating efficacy, and offers suggestions for modifying and improving the drug development process so that clinicians can better judge new drugs.

(J Clin Psychiatry 1999;60[suppl 4]:23–31)

The antidepressants now available are the result either of rational, targeted synthesis to achieve a specific effect on 1 or more neurotransmitter systems or of serendipitous observation of therapeutic effects. No currently approved medication obtains intent-to-treat response rates of better than 60%—actually, intent-to-treat rates of 40% to 50% are commonplace—and 10% to 20% of antidepressant trials are ended prematurely because of adverse effects.[1,2] Thus, because of such imperfections, coupled with the public health significance of the depressive disorders, there is still a great need for novel antidepressant agents. Indeed, each 1% of the "market share" for approved antidepressants is worth millions of dollars to the manufacturer.

It is, however, not easy to develop a successful new antidepressant. New compounds must be taken through a series of preclinical studies, which include determination of drug safety and analogue effects in various rodents and large mammals. Various animal models of depression examine the length of time an animal persists in a swim task, reversal of or resistance to learned helplessness, or response to maternal deprivation. When rodents cannot escape from a restricted swim space, they eventually cease their efforts to escape and become immobile, suggesting a state of despair.[3] Learned helplessness relates closely to some of the important aspects of clinical depression, particularly cognitive aspects such as a negative self-concept, negative interpretations of one's experiences, and a negative view of the future. Disruption of attachment bonds in humans almost invariably leads to grief reactions that can precipitate clinical depression in vulnerable individuals. By evaluating the effects of drugs on these experimental paradigms in animals, investigators can better understand specific aspects of human depression.

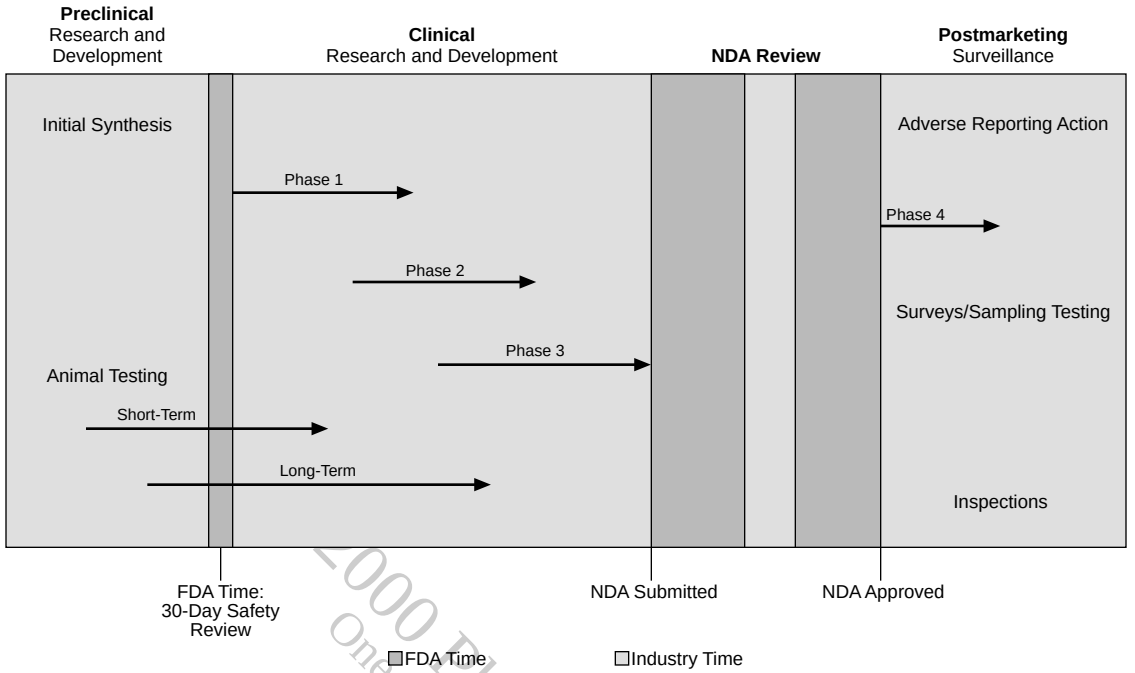
From the University of Pittsburgh School of Medicine, Pittsburgh, Pa.

Supported in part by National Institute of Mental Health grants MH-30915 (Mental Health Clinical Research Center) and MH-41884.

Presented at the symposium "Assessing Antidepressant Efficacy: A Reexamination," held January 20–21, 1998, Phoenix, Ariz., which was supported by an unrestricted educational grant from Wyeth-Ayerst Laboratories.

Reprint requests to: Michael E. Thase, M.D., Western Psychiatry Institute & Clinic, 3811 O'Hara Street, Pittsburgh, PA 15213.

Figure 1. New Drug Development*



*Adapted from reference 7. Abbreviations: FDA = Food and Drug Administration, NDA = New Drug Application.

The present system leading to approval of new drugs has evolved through the collaborative efforts of government, industry, and academicians. It is important to keep in mind that psychopharmacology is still in its relative youth. The first federal money for psychopharmacologic studies was awarded in 1956, less than 50 years ago. New prescription drugs were not required to show evidence of efficacy prior to marketing until enactment of the Kefauver-Harris amendments in 1962, and the Food and Drug Administration (FDA) was charged with implementing the efficacy requirements of the amendments within the federal government. The FDA subsequently established general standards for clinical investigations and developed regulations for Investigational New Drugs and New Drug Applications (NDAs).[4–8] The first FDA guidelines for randomized clinical trials (RCTs) of psychotropic compounds were proposed in 1970 and revised in 1977 and 1978. Additional refinements in the late 1980s and early 1990s included efforts to improve representation of women of childbearing potential and a mechanism for studying orphan drugs, i.e., compounds that no longer were covered by a patent but may have promising clinical effects for people suffering from relatively rare disorders.

The FDA program of drug development proceeds through 3 formal but partly overlapping phases, with the third phase ideally culminating in an NDA for approval of a beneficial compound for general use (Figure 1).[7] This article will describe the process of conducting the clinical trials that support an NDA, discuss issues in evaluating efficacy of antidepressants, and offer suggestions for modifying and improving the drug development process so that clinicians can better judge new antidepressants.

THE DRUG DEVELOPMENT PROCESS

The 3 FDA Phases

Compounds that show promising antidepressant effects—e.g., blockade of reuptake of serotonin or norepinephrine—in vitro and/or in animal models are extensively studied for safety in various small and moderately sized mammals such as mice, rats, pigs, and dogs. Those of sufficient promise proceed to human studies. Phase 1 is specifically devoted to human safety studies of normal or healthy volunteers. Initial doses are guided by results from animal studies, such that 1/100 or 1/50 of a dose known to be toxic in a relevant animal model usually is taken as the threshold dose. Doses are then escalated within and across subjects so that safety and tolerability can be assessed at a low risk for untoward reactions. As dose escalation studies proceed, subjects are often randomly assigned to a specific fixed dose and assessed on various behavioral or toxicologic paradigms. For example, sedative effects are tested by studies of performance and vigilance, as well as by studying additive effects in combination with alcohol. Placebo control is typically used in only 20% to 25% of a given sample in this phase.[5] If a compound continues to look promising and not particularly toxic, phase 1 usually ends with detailed studies of pharmacokinetics, bioavail-

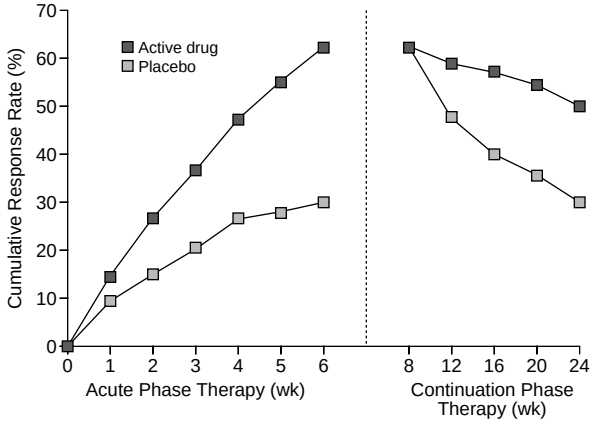
ability, and metabolism. This type of study now routinely includes in vitro and, if necessary, in vivo estimations of potential for drug interactions through inhibition of the various isozymes of the cytochrome P450 system.

Phase 2 is the stage of drug development in which clinical utility or efficacy of the compound is initially investigated. The phase ordinarily begins with open-label studies and ends with demonstration of efficacy in double-blind, placebo-controlled RCTs. Of course, compounds already approved for other indications may reenter phase 2 almost imperceptibly as creative clinical investigators begin to collect case series of off-label clinical use. For novel compounds, dose ranging, the determination of minimum and maximum effective doses, should ideally occur in the early part of phase 2 so that both fixed-dose and dose-titration paradigms can be utilized in the RCTs. In fact, the registration process in some countries includes demonstration of a compound's null dosage, i.e., the dosage that is unlikely to surpass placebo in efficacy. Compounds that demonstrate safety, tolerability, and efficacy against placebo in small phase 2 studies proceed to the third phase of investigation. During phase 3, the definitive studies of safety and efficacy of a new compound for a specific disorder are conducted.

The FDA now requires that phase 3 studies be carried out with representative samples of subjects that include men and women aged 18 to at least 60 years from various regions of the country. Thus, it is extremely unlikely that a new compound could be approved for general use in the United States on the sole basis of studies conducted in Europe, Canada, Mexico, or South America. At least 2 pivotal trials must be positive for a drug to successfully complete phase 3: the term *positive* refers to a drug benefit that is significantly greater than that of placebo on a key outcome selected prospectively (i.e., before the trial was begun), and *pivotal* refers to the size and representativeness of the trial. Because billions of dollars are usually at stake in the approval of a promising new antidepressant, most research sponsors will initiate a number of pivotal phase 3 trials to increase the likelihood that results from 2 trials will be positive. However, an FDA application is not based solely on the successful trials; therefore, if multiple trials are initiated, they must all be part of the FDA application for new drug approval.

In addition to placebo control, many phase 3 studies of antidepressants also include an active comparator condition—such as 1 of the tricyclic antidepressants in studies conducted in the 1980s and 1 of the selective serotonin reuptake inhibitors (SSRIs) in studies conducted in the late 1990s. An active comparator condition affords a research design several advantages. First, the efficacy and tolerability of the novel compound can be compared against those of a known standard. Second, the integrity or sensitivity of the research design can be cross-checked by examining the effectiveness of the comparator against the placebo condition. One can be more confident that a study failed because of methodological problems—rather than a poor drug—if the standard comparator also failed to surpass placebo. Third, potential, clinically important differences between the novel compound and the standard comparator might be revealed. This feature has obvious implications for both patient care and the design of subsequent studies, as well as future plans for marketing.

Figure 2. Hypothetical Acute- and Continuation-Phase Results of an Effective Antidepressant



During the course of phase 3, the research sponsor is now required by the FDA to collect at least 1 year of continuation phase treatment data because these medications are intended for long-term use. The FDA does not yet *require* placebo-controlled evidence of relapse prevention or prevention of recurrence in phase 3, although researchers have recently begun to conduct such studies as a final part of the approval process (i.e., phase 3b). Numerous studies document that the magnitude of the drug-placebo difference during a controlled continuation phase study is comparable to, if not larger than, that observed during an acute phase trial (Figure 2), which suggests that an active antidepressant effect may be determined by demonstrating either a beneficial effect by initiating the medication or a loss of effect (among responders) after its discontinuation. The use of placebo-controlled studies of novel treatments of conditions for which there are already a number of effective medications has recently come under fire. Specifically, some consider it to be unethical to subject people suffering from an illness such as depression or panic disorder to a placebo therapy—actually to a 50% or 33% chance of receiving a placebo therapy—when there are proven standards of therapy. This well-intentioned position is based on the assumption that studies contrasting the novel compound and a standard comparator should suffice, namely, that showing equivalence is synonymous with showing efficacy. Indeed, such 2-group equivalence studies are now commonplace in Europe. The logic supporting the exclusive use of comparator-controlled studies

is flawed, however, and fails to take into account that (1) the average drug-placebo difference in RCTs is only about 4 points on the Hamilton Rating Scale for Depression (HAM-D) or an 18% to 20% difference in global response rates; (2) approximately one third of all published RCTs of proven antidepressants have failed to show drug-placebo differences; and (3) it is much more difficult to show that 2 antidepressants are not statistically different from each other than it is to show that an active compound is more effective than a placebo.[1,2] As a result, studies comparing 2 active antidepressants need to be at least twice the size of drug-placebo comparisons, which means that twice as many people would be subjected to the unproved therapy. Larger studies also mean greater costs, and, in all likelihood, a longer wait until the novel compound is available for general use.

The anti-placebo position also fails to take into account that consenting adults are capable of making informed choices about the pros and cons of enrolling in a placebo-controlled study and that the risks of receiving a placebo are lessened by exclusion of the most desperately ill patients, ongoing and careful monitoring, and provision of an open-label standard antidepressant for all nonresponders at the end of double-blind therapy. Not infrequently, patients say during debriefing that they had hoped to receive the placebo, either to avoid the side effects of the active compound or to "prove" that they did not require an antidepressant after all. Such ambivalence could be dealt with proactively if the 2-group comparator-controlled RCT was the best way to test new antidepressants, but it is not, and researchers are grateful to those open-minded consenting patients and the Institutional Review Boards that permit the more appropriate placebo-controlled phase 2 and phase 3 RCTs to continue to be conducted.

The culmination of phase 3 is a formal NDA to the FDA. New drug applications are reviewed by a panel of experts in the field. The product is ordinarily approved if there is evidence of efficacy and reasonable safety from pivotal trials. On rare occasions, a standard is applied to indicate that the medication provides a different benefit from drugs already available. More commonly, product labeling reflects particular side effects or risks that are greater for the new compound than existing medications within the same grouping. Although theoretically possible, it would be extremely unusual for an NDA to be turned down because the drug, albeit effective, offers no substantial advantages over compounds already available.

Phase 4 takes place after the drug has been approved for general use and is available by prescription to the general public. An essential element of phase 4 is the confirmation of drug safety and the determination of rare or uncommon effects so as to improve drug labeling and clinicians' monitoring of potential uncommon toxicities. The *Rule of Three*[9] is commonly used to approximate the number of drug exposures needed to detect a rare event: if an untoward event has an actual occurrence in 1 of 10,000 patient exposures, the drug will need to be administered to at least 3 times that many patients before one can be confident that the event will be observed. Hence, priapism was not observed until well after trazodone was approved, and drug fevers and Guillain-Barré syndrome were not observed until after the release of nomifensine and zimelidine, respectively. Even more patients will have to be observed for less common adverse events, such as aplastic anemia, and as many as 10 million drug prescriptions may have to be written before an extremely rare adverse event is observed.

Another component of phase 4 research falls under the rubric of pharmacoepidemiology—that is, the general use and tolerability of the compound and the number of patients for whom it is (and is not) prescribed, as assessed by the number of prescriptions that are refilled. Many clinical researchers consider this part of phase 4 research to be the area in which the practical utility of efficacy is determined in less well-selected groups of patients. A services research industry, which conducts many of the phase 4 research studies in the United States, has arisen from pharmacoepidemiologic studies. In contradistinction to the efficacy studies of phases 2 and 3, some researchers refer to phase 4 investigations as effectiveness studies. In contrast to the more tightly controlled FDA efficacy studies, these studies generally have relatively simple outcome measurements, and, in the absence of randomization of treatments, definitive interpretations are more difficult. To illustrate one common pitfall, many clinicians reserve use of newly-introduced treatments for patients with more complicated or treatment-resistant conditions. As a result, a novel therapy undoubtedly appears less useful than a first-line compound in early phase 4 studies.

Patient Selection

The primary purpose of most clinical trials is to identify effective treatments, but it is unlikely that any single treatment will be effective for all patients.[10] Therefore, the selection of patients to participate in a clinical trial is of paramount importance. Many inclusion and exclusion criteria are available, which investigators use to choose a study sample, and the large number of criteria may reflect a lack of knowledge or agreement on definitions to describe the depressive syndrome. Generally, 5 to 10 patients are screened for every 1 or 2 patients enrolled in a clinical trial of an antidepressant. Indeed, the average depressed patient in a psychiatrist's private practice would probably be ineligible to participate in a phase 3 clinical trial, often because of comorbid medical or psychiatric illness. Because the FDA imposes no requirements for pretreatment patient assessment or assessment of antidepressant efficacy in certain comorbid or complex patient subgroups, speed and efficiency typically streamline the assessment formats. Many phase 2 and phase 3 trials use no standard-

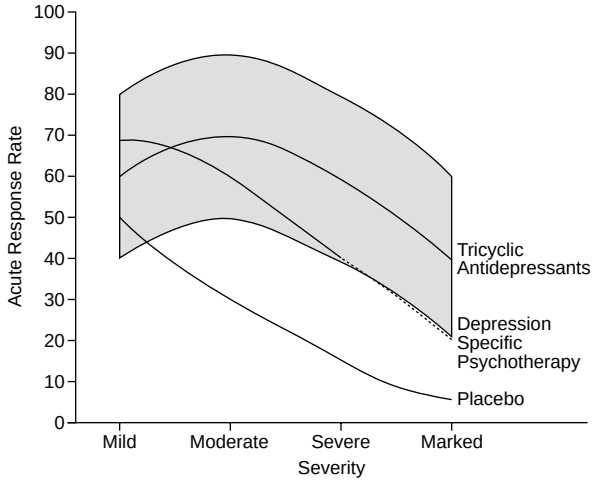
ized diagnostic assessments and difficult-to-treat patients or those with complex conditions are simply not enrolled. High-volume sites may even operate by use of a "don't ask/don't tell" strategy. If a questionable exclusion criterion is apparent or even debatable, it is in the best interest of the research entrepreneur to enroll the patient without clarifying the presence or absence of the ambiguous exclusion criterion. Fortunately, young women can now participate in industry-sponsored clinical trials unless they are pregnant, breastfeeding, or—if sexually active—unwilling to use an acceptable form of birth control.[11]

A new compound entering the market has usually been studied in only 1000 to 3000 patients,[12] a potentially inadequate database for detecting uncommon side effects. Thus, surprises may await clinicians who prescribe newly released medications for their patients. At the time of initial marketing, medication effects are untested in patients who have some of the most important and classic subtypes of depression, including bipolar and psychotic forms of depression, as well as dysthymia. It is generally not known if an antidepressant is more or less effective for melancholia or atypical depression, if efficacy is affected by common medications (including estrogen replacement therapy), or if the agent is effective and well tolerated by elders, especially those with significant cardiovascular disease. Studies of depressed patients with alcoholism or drug addiction are generally nonexistent at the time of FDA approval, even though comorbid depression and substance abuse is highly prevalent, and antidepressants are prescribed widely to such patients. Controlled studies may have been conducted against 1 or more comparators (including a market leader) but systematically controlled studies of patients unresponsive to the most widely used antidepressants are rare. Moreover, there are usually few (if any) efficacy data available for depressed children and adolescents. Accordingly, a remarkable number of unknown issues surround a new drug as it enters the marketplace.[13]

Judging Efficacy

The large number of failed clinical trials of approved antidepressants can be partially explained by the clinical and neurobiological heterogeneity of the depressive disorder. Effect sizes vary somewhat, depending on the severity of the population; however, the golden zone to detect efficacy of most antidepressants is probably in the moderate to moderately severe range of depression where the typical response rates, after adjusting for attrition, are approximately 50% for the active drug and 20% to 30% for placebo (Figure 3).[14] If the depression is severe, as in psychotic depression, antidepressant effect sizes drop considerably, although placebo response rates may approach zero. Although antidepressants are thus effective in psychotic depression relative to placebo, depressed patients manifesting hallucinations or delusions are usually excluded from trials conducted in the United States because the standard of care for psychotic depressions is combined therapy with an antidepressant and an antipsychotic, or electroconvulsive therapy. On the other hand, placebo response rates of 50% to 70% are not uncommon in mild depression.[1,2] A ceiling effect sometimes compromises interpretation of studies of mild depression because of the lack of design power to show improvement when placebo response rates are high.

Figure 3. Efficacy Zone in Relation to Severity*



*From reference 14, with permission.

Because phase 2 and phase 3 trials usually exclude patients with both common psychiatric and serious medical comorbidities, the resulting data may also fail to prepare clinicians to work with more complicated patients. Results from RCTs also may inflate expectations unrealistically. For example, antidepressant response rates are sometimes documented as low as 10% when drugs are administered in unstructured primary care situations, such as in a 1990 survey of California Medicaid patients[15] or in the more recent study of Schulberg et al.[16] in which nortriptyline was the standard antidepressant pharmacotherapy and usual care was the control condition. In the Schulberg et al. study, the tricyclic antidepressant was robustly effective when provided by primary care physicians according to a standardized protocol, yet a broad array of potential antidepressant therapies yielded dismal results when provided as usual care. It should be a given that an antidepressant effect is neither reliable nor reproducible when noncompliance and poor clinical management prevail. No medication, no matter how potent, is effective if it is left in the bottle.

For many (if not most) patients, major depressive disorder is a recurrent and potentially chronic illness that warrants long-term, preventive treatment. An acute phase response to a new agent thus will often lead to an indefinite course of maintenance phase therapy. However, clinicians

usually have access to only maintenance phase efficacy data on compounds that have passed through the 3 FDA phases of drug development. At present, the vast majority of controlled clinical trials data beyond 1 year of treatment—and the only means to demonstrate prevention or protection against recurrent depression—come from studies of tricyclics, monoamine oxidase inhibitors, and lithium salts. There are only a few published studies with a duration of therapy greater than 1 year involving antidepressants introduced since 1988.[17–19] Thus far, there have been no major surprises, but issues such as loss of efficacy over time, late-emerging weight gain, and untoward antidepressant discontinuation syndromes continue to be topical concerns.

Other topics that must typically wait until after an antidepressant is introduced concern familiarity and applicability. For example, as clinicians gain more extensive experience with the new antidepressant, they are better able to determine the best dosage across an age span; differences in dosage or tolerability as a function of gender, severity, or comorbidity; typical duration of treatment before dose adjustment or discontinuation is necessary; and utility (or lack of utility) of various means of augmentation and/or combinations of the new antidepressant with known treatments. Phase 4 occasionally serves as a marketing-oriented period for targeted studies of the new antidepressant against specific comparators. During this time, clinical research proliferates in an effort to establish alternate uses and indications for the new antidepressant in conditions such as posttraumatic stress disorder, chronic pain disorder, and sleep disorder. Recent success of antidepressants in the treatment of DSM-IV obsessive-compulsive disorder,[20] panic and other anxiety disorders,[20–22] and eating disorder[23,24] are other examples of alternate indications for a new compound.

ISSUES PERTAINING TO FUNDING

The urgency that corporate sponsors feel to move medications through the approval process dictates that multicenter trials be used; for a study of 300 patients, as many as 30 sites may enter 5 to 20 patients each within 6 months of enrollment. Since more rigorous, hypothesis-driven clinical research programs basically are unable to deliver rapid enrollment of large numbers of patients, treatment research services (derisively called research mills by some individuals) have emerged and flourished. Phase 2 and phase 3 studies are now usually conducted at these high-volume clinical venues, and most such studies have few or only a minority of academic sites. However, almost without exception, the research mills outrecruit the academic programs and account for the lion's share of data known about antidepressant compounds at the time of marketing. Further, any potential advantages for a greater participation of academic centers remain just that—potential advantages. Although there are few if any hard data indicating that academic centers demonstrate larger or more consistent drug-placebo differences or, for that matter, more reliable diagnoses or assessments of dependent measures, such differences might be demonstrated if a systematic study were to be undertaken.

Parsimony and urgency to get a drug to the marketplace inhibit risk taking in terms of the industry developing and implementing alternate research designs. Industry-sponsored trials also are conducted in clear deference and obedience to FDA standards; however, caution breeds conformity and stagnates creativity, such that investigations are so similar in design that 1998 studies resemble those conducted 10 years earlier. This practice results in large clinical studies that are designed for a single specific purpose and are often inefficient for the broader field and broader questions. Consequently, researchers are not able to reliably detect subtle differences between antidepressants when the drug is initially marketed. Studies are powered on the ability to find a 20% difference in efficacy among antidepressants, and the art of skilled clinical practice often hinges on the clinician's ability to perceive more subtle differences among drugs that deliver a 10% to 15% advantage. The 20-year controversy regarding the differential effectiveness of tricyclic antidepressants and monoamine oxidase inhibitors is a good example of this dilemma.[24,25]

One cardinal example of slavish obedience to FDA standards is the continued use of a 7- to 10-day single-blind placebo lead-in, even when no empirical evidence exists that the approach improves clinical trials or alters placebo response rates, attrition rates, or acute antidepressant response rates.[26] It is likely that a 3- or 4-week double-blind lead-in is necessary to eliminate patients who are likely to benefit from the nonspecific aspects of therapeutic support. Routine use of an extended placebo lead-in for all patients would stretch the ethical comfort zone of most investigators, however. My colleagues and I have recently developed, as a more humane alternative, a lead-in procedure in which 45 minute sessions of supportive counseling are provided to patients in addition to clinical management and single-blind placebo. This intervention yields about a 25% response rate and, based on a preliminary analysis of an ongoing study, appears to markedly decelerate subsequent placebo response rates (Thase ME, Friedman ES, Berman SR, et al. Unpublished observations, 1998).

Hypothesis-driven phase 4 studies are largely supported either by extramural grants from the National Institute of Mental Health (NIMH) or by funded studies from the manufacturer of the compound. On the one hand, the manufacturer of a new drug probably has the most to gain from additional research after the introduction of a new agent. On the other hand, new drug development is expensive; there may be as many as 20 blind alleys or false leads for every medication that successfully moves through the

FDA process. It is not fair to view such reticence simply as greed, as stockholders demand parsimony in the drug development process. Typically, there is both excitement and a sense of urgency to push a new compound through the first 3 phases of study and into the marketplace. This sense of urgency—with the patent license running out—can be likened to a ticking clock. Speedy completion of trials is necessary because the longer a medication stays in the marketplace under the protection of a patent, the greater the return on the investment. With the profit margin in mind, there is no compelling reason—with the possible exception of good will and regard for mankind—for the manufacturer to invest heavily in phase 4 studies. However, for clinical investigators pursuing federal funding, it should be appreciated that most phase 4 studies are pragmatic and not of the greatest theoretical interest, and, as a result, have a low probability of obtaining competitive federal funding.

DESIGNING MORE EFFICIENT TRIALS

Tactics for Better Research

The drug discovery process being used today has curbed development of new classes of drugs. There has been a "me-too" philosophy in the past 20 years among pharmaceutical companies, and—although sensible in terms of profit margins and security—similar preclinical screening devices have produced related anxiolytics, antidepressants, and antipsychotics that, with few exceptions, lack a pathophysiologic basis: one result is growing numbers of SSRIs and a myriad of antidepressants that work via other mechanisms now on the market. One is reminded of the story of the man looking for his keys under the lamppost because the light was better. In this case, it is likely that a key will be discovered, but it is unlikely that the key will open any new doors.

The issues of research funding—particularly division of the research pie—and the relationship of the FDA to the pharmaceutical industry, academe, and the NIMH need to be revisited. The field might uncover entirely new classes of psychotropic medications that are less burdened by clinical design problems if it operated more as a partnership (similar to oncology) among the NIMH, academic research centers, and the pharmaceutical industry. As lamented previously, certain forms of research that are extraordinarily important in clinical practice are seldom funded by NIMH. Perhaps these issues are starting to be addressed, and there is currently a greater interest in intervention and transnational research at the federal level. Hopefully, research on biology of depression will finally catch up with the clinical pharmacology so that researchers can begin to make reasoned recommendations to clinicians and their patients on the basis of data that match drug effects with specific vulnerabilities and pathophysiologic alterations. One can at least hope for such a day within the next 2 decades. The notion of universal registration of subjects should also be considered so that the so-called professional patient cannot bias clinical trials by enrolling in multiple treatment studies.

The power to detect differences of clinical interest would be immediately improved if standardization of assessments and reliability of diagnosticians and evaluators were ensured. Researchers often underestimate the variability in ratings performed, both within a given center and across multicenter sites, despite use of standard measures. Even under optimal circumstances, the diagnosis of major depressive disorder is made with only 90% interrater agreement, which delimits the validity of the trial. If a symptom-based outcome measure is administered with an intraclass correlation coefficient of 0.7, the upper limit on the validity of that trial is reduced even further. Fortunately, time and effort can remedy the problems caused by poor assessment reliability if these problems are acknowledged and the sponsor of the trial is willing to address the problem.

Better strategic coordination of phase 2 and phase 3 trials by industry would also improve outcome as well as help to move the drug toward the marketplace. Patients would be offered participation in an authentic ongoing research program, rather than fragmented 6- to 8-week treatment protocols designed to serve only 1 purpose. One way to coordinate trials of a novel antidepressant at the phase 3 level is to qualify all responders of acute phase studies as eligible to enter ongoing extension groups randomly assigned (by risk group) to longer term studies. This would provide data on continuation- and maintenance-phase effectiveness, relapse prevention, prevention of recurrence, improvements in quality of life, psychosocial functioning coincident with treatment, and late-emerging side effects. Patients who enter such studies and respond to antidepressant treatment would be allocated to a maintenance treatment trial; similarly, patients who had incomplete or partial responses to antidepressants would be excellent candidates to be enrolled in relapse prevention studies. Residual symptoms are a predictor of relapse, and continued antidepressant treatment—perhaps in higher doses in partially remitted patients—would enable investigators to determine if the medication given for a 3- to 6-month period would deliver more complete and encompassing clinical response. Placebo-controlled studies of relapse and recurrence create the opportunity for patients who become ill on placebo to be restabilized on active medication and (if the patient is interested) recycled into alternative or other research studies. For example, a patient who becomes ill on placebo during the continuation phase would be an excellent candidate to enter into a study of longer term models of treatment since it has already been established that the patient should be slowly withdrawn from antidepressants. Similarly, patients who complete a 6-month or 1-year study would seem to be ideal candi-

dates to participate in longer term trials. Again, this whole model of moving successes forward and recycling patients who relapse and suffer recurrence of depression is predicated upon being able to present a comprehensive research program to patients at the beginning of treatment.

Other opportunities to enrich or improve clinical trials include identifying placebo nonresponders and focusing studies of the new drug versus a comparator in these difficult-to-treat patients. Similarly, patients who do not respond to the comparator drug comprise an interesting group to consider for studies of the new agent against yet another dissimilar comparator or, conversely, a second member of the comparator class of compounds. Moreover, the 30% to 50% of patients who fail to respond to the new study medication could be considered for participation in controlled studies of dose escalation (greater than the typical therapeutic dose) or studies of augmentation and/or combination treatments. By comprehensive coordination, the 1000 to 3000 patients exposed to a new antidepressant in phase 2 and phase 3 studies would yield sufficient numbers of patients with melancholia, atypical depression, and depression superimposed upon dysthymia to permit meaningful analyses of these important clinical subgroups. Although my vision may be obscured by naiveté or the blinding simple logic of this type of coordinated approach, it would appear that such research programs could be conducted at little cost above that already devoted to phase 3 drug development. Such costs could be offset further by savings in subject recruitment, duplication of effort, and sample size reductions resulting from optimizing the reliability of diagnostic and dependent variable assessments.

Assessment standards appear to have been set more to meet regulatory requirements than to provide a clear understanding of the contribution of the new treatment strategy toward remission of the depressive syndrome.[8] The definitions that are commonly employed to describe the outcome of depressive disorders are often inconsistent and largely untested. In 1988, the MacArthur Foundation Mental Health Research Network on Depression organized a task force to examine the ways in which change points in the course of depressive illness had been described and the extent to which inconsistency in these descriptions might be impeding research on this disorder.[27,28] In a recent study,[29] my colleagues and I operationalized definitions for the following critical outcomes: response, remission, recovery, relapse, and recurrence. The validity of these definitions was then examined in a sample of depressed patients treated with psychotherapy rather than pharmacotherapy. All 5 definitions demonstrated moderate-to-excellent validity, and we were able to empirically distinguish response from remission and relapse from recurrence, i.e., sources of frequent confusion in the literature.

Attrition can basically invalidate a clinical trial. Not uncommonly, 20% to 40% of patients drop out of randomized clinical trials. The placebo lead-in may contribute to this problem by delaying the start of active treatment. In the substance abuse field, the practice of spending time with prospective study participants to emphasize the importance of attending therapy sessions and following instructions is commonly being substituted for placebo lead-in. Parallel models of psychoeducation can also be used with prospective patients to lower attrition rates from a compromising 30% to 40% to a more manageable 10% to 15%. Within a research strategy, tactics that ensure reliability, encourage attention to adherence, and lessen attrition at the outset of a study will increase the power and design sensitivity of a particular trial.

Alternative Strategies

Other strategies that can increase the knowledge base of new antidepressants include creative modifications of randomized placebo-controlled designs to accommodate the incremental knowledge gained during a clinical trial. One such strategy is a "Play the Winner" design,[30] an adaptive sampling method whereby the allocation of subjects into a treatment cell increases as that treatment shows greater effect. This approach results in exposing fewer patients to ineffective treatment and more patients to effective treatment, thereby improving upon the a priori assessment of sample size requirements by shifting the number of patients in each condition toward a better match of effectiveness, sample size, and subject composition.

Another important and underutilized approach is meta-analysis of existing datasets. It is important to identify the mean and the standard deviation of HAM-D scores and categorical response rates at 2, 4, 6, and 8 weeks of treatment in published studies so that these data can be used in meta-analyses. Meta-analyses do not enable a series of bad studies to counteract the lack of good studies, but they do allow trends that are apparent in small studies to be assessed in a way that is consistent and reliable. A particular type of meta-analysis now being used is Bayesian meta-analysis,[31] in which the probability that a given drug will be efficacious is estimated so that the contribution of any particular study within the database (or even a new study that is not yet conducted) can then be estimated. This approach is a useful way of establishing the variability across studies and qualifying the reproducibility of a drug effect. It may also help planners to determine if additional controlled studies of a given compound are unnecessary. Bayesian meta-analysis is also a most useful means for demonstrating small effects—on the order of a 10% difference—between various antidepressant medications. One has little confidence in a 10% difference in a single study because such a small effect is unreliable. Thus, small differences are usually interpreted as being clinically insignificant. However, a consistently reproducible 10% difference among antidepressants is important, and most clinicians and their patients would opt for a treatment that

is 10% more effective than another treatment if the effect were reliable.

Another type of meta-analysis utilizes original data, pooled across studies.[32] The type of analysis is now quite feasible for new antidepressants because all data are preserved when agencies file an NDA with the FDA. Thus, in an era of computers and sophisticated data management programs, it is possible to conduct large comprehensive meta-analyses based on original patient data, rather than the mean, the standard deviation, or the categorical response rate of RCTs. This type of mega-analysis is the most powerful method for comparing the efficacy of treatments, examining correlates of response, and addressing questions pertaining to specific subgroups.[33,34]

CONCLUSION

Randomized clinical trials still provide the best means to determine the efficacy of a new treatment. However, this standard is limited and the randomized clinical trials that are conducted in support of a New Drug Application are a particularly inefficient and costly approach. However, both strategic and tactical alternatives are available to improve them.

Drug names: nortriptyline (Pamelor and others), trazodone (Desyrel and others).

REFERENCES

1. Clinical Practice Guideline Number 5: Depression in Primary Care, vol 2. Treatment of Major Depression. Rockville, Md: US Dept Health Human Services, Agency for Health Care Policy and Research; 1993. AHCPR publication 93-0551
2. Thase ME, Kupfer DJ. Recent developments in the pharmacotherapy of mood disorders. J Consult Clin Psychol 1996;64:646–659
3. McKinney WT. Animal research and its relevance to psychiatry. In: Kaplan HI, Sadock BJ, eds. Comprehensive Textbook of Psychiatry. 6th ed. Baltimore, Md: Williams & Wilkins; 1995:400–404
4. Levine J. History and role of guidelines and guidance for drug evaluation. In: Prien RF, Robinson DS, eds. Clinical Evaluation of Psychotropic Drugs. New York, NY: Raven Press; 1994:1–12
5. Lasagna L. Decision process in establishing the efficacy and safety of psychotropic agents. In: Prien RF, Robinson DS, eds. Clinical Evaluation of Psychotropic Drugs. New York, NY: Raven Press; 1994:13–28
6. Peck CC, Barr WH, Benet LZ, et al. Opportunities for integration of pharmacokinetics, pharmacodynamics, and toxicokinetics in rational drug development. J Pharm Sci 1992;81:605–610
7. New Drug Development in the United States. US Dept Health Human Services, Food and Drug Administration. Rockville, Md: HHS publication (FDA) 88-3168; 1976
8. Klerman GL, Weissman MM, Frank E, et al. Evaluating drug treatments of depressive disorders. In: Prien RF, Robinson DS, eds. Clinical Evaluation of Psychotropic Drugs. New York, NY: Raven Press; 1994:281–325
9. Laughren TP, Levine J, Levine JG, et al. Premarketing safety evaluation of psychotropic drugs. In: Prien RF, Robinson DS, eds. Clinical Evaluation of Psychotropic Drugs. New York, NY: Raven Press; 1994:185–215
10. Endicott J, Schwartz GE, Lee JH. Classification issues in patient selection and description. In: Prien RF, Robinson DS, eds. Clinical Evaluation of Psychotropic Drugs. New York, NY: Raven Press; 1994:69–83
11. Merkatz RB, Temple R, Sobel S, et al. Women in clinical trials of new drugs. N Engl J Med 1993;329:292–296
12. Hollister LE, Jones JK, Fisher S. Post-marketing surveillance of drugs. In: Prien RF, Robinson DS, eds. Clinical Evaluation of Psychotropic Drugs. New York, NY: Raven Press; 1994:217–235
13. Klein DF. Clinical psychopharmacologic practice: the need for developing a research base. Arch Gen Psychiatry 1993;50:491–494
14. Thase ME, Howland R. Refractory depression: relevance of psychosocial factors and therapies. Psychiatr Ann 1994;24:232–240
15. McCombs JS, Nichol MB, Stimmel GL, et al. The cost of antidepressant drug therapy failure: a study of antidepressant use patterns in a Medicaid population. J Clin Psychiatry 1990;51(6, suppl):60–69
16. Schulberg HC, Block MR, Madonia MJ, et al. Treating major depression in primary care practice. Arch Gen Psychiatry 1996;53:913–919
17. Montgomery SA, Dufour H, Brion S, et al. The prophylactic efficacy of fluoxetine in unipolar depression. Br J Psychiatry 1988;153:69–76
18. Stewart JW, Quitkin PJ, McGrath PJ, et al. Use of pattern analysis to predict differential relapse of remitted patients with major depression during 1 year of treatment with fluoxetine or placebo. Arch Gen Psychiatry 1998; 55:334–343
19. Keller MB, Kocsis JH, Thase MR, et al, and the Sertraline Chronic Depression Study Group. Maintenance phase efficacy of sertraline for chronic depression: a double-blind, placebo-controlled study. JAMA. In press
20. Flament MF, Bisserbe J-C. Pharmacologic treatment of obsessive-compulsive disorder: comparative studies. J Clin Psychiatry 1997;58(suppl 12): 18–22
21. Gorman JM. The use of newer antidepressants for panic disorder. J Clin Psychiatry 1997;58(suppl 14):54–58
22. Keck PE Jr, McElroy SL. New uses for antidepressants: social phobia. J Clin Psychiatry 1997;58(suppl 14):32–36
23. Jimerson DC, Herzog DB, Brotman AW. Pharmacologic approaches in the treatment of eating disorders. Harv Rev Psychiatry 1993;1(2):82–93
24. Quitkin FM, Stewart JW, McGrath PJ, et al. Columbia atypical depression: a subgroup of depressives with better response to MAOI than to tricyclic antidepressants or placebo. Br J Psychiatry 1993;163(suppl 21):30–34
25. Thase ME, Trivedi MH, Rush AJ. MAOIs in the contemporary treatment of depression. Neuropsychopharmacology 1995;12:185–219
26. Trivedi MH, Rush J. Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? Neuropsychopharmacology 1994;11:33–43
27. Frank E, Prien RF, Jarrett RB, et al. Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Arch Gen Psychiatry 1991;48:851–855
28. Prien RF, Carpenter LL, Kupfer DJ. The definition and operational criteria for treatment outcome of major depressive disorder. Arch Gen Psychiatry 1991;48:796–800
29. Riso LP, Thase ME, Howland RH, et al. A prospective test of criteria for response, remission, relapse, recovery, and recurrence in depressed patients treated with cognitive behavior therapy. J Affect Disord 1997;43:131–142
30. Laska EM, Klein DF, Lavori PW, et al. Design issues for the clinical evaluation of psychotropic drugs. In: Prien RF, Robinson DS, eds. Clinical Evaluation of Psychotropic Drugs. New York, NY: Raven Press; 1994: 29–67
31. Eddy DM, Hasselblad V, Shacter R. An introduction to a Bayesian method for meta-analysis: the confidence profile method. Med Decis Making 1990;10:15–23
32. Moher D, Olkin I. Meta-analysis of randomized controlled trials: a concern for standards. JAMA 1995;274:1962–1964
33. Thase ME, Greenhouse JB, Frank E, et al. Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. Arch Gen Psychiatry 1997;54:1009–1015
34. Thase ME. Effects of venlafaxine on blood pressure: a meta-analysis of original data from 3744 depressed patients. J Clin Psychiatry 1998;59: 502–508

DISCLOSURE OF OFF-LABEL USAGE

The author of this article has determined that, to the best of his clinical estimation, no investigational information about pharmaceutical agents has been presented herein that is outside Food and Drug Administration–approved labeling.