Commentary

Studying New Antidepressants: If There Were a Light at the End of the Tunnel, Could We See It?

Michael E. Thase, M.D.

The availability of so many newer antidepressants necessitates that clinicians make daily judgments about the relative utility of these medications. Traditionally, it has been assumed that antidepressants that have passed the review process of the U.S. Food and Drug Administration (FDA) are (more or less) comparably effective.[1] If treatment options are truly comparably effective, then factors such as convenience, tolerability, safety, and cost are typically used to determine which medications are selected first, second, and third in a sequence of choices.[2] By the end of the 1990s, 4 members of the selective serotonin reuptake inhibitor (SSRI) class, fluoxetine, sertraline, paroxetine, and citalopram, were widely accepted as first-choice options.[2] When compared to the older standard, the tricyclic antidepressants (TCAs), the SSRIs had only 1 comparative disadvantage, higher direct cost. This drawback will begin to dissipate as generic formulations of fluoxetine and other class-mates become available. The SSRIs can thus be considered the new standard of comparison, against which all new antidepressants must be measured.

It is likely that yet another new antidepressant, duloxetine, shortly will be approved for use in the United States. Available evidence indicates that duloxetine has bona fide antidepressant effects.[3] Such early findings, which obviously require more extensive replication, are buttressed by the conceptual argument that antidepressants that simultaneously and directly affect noradrenergic and serotonergic neurotransmission will be more effective than more selective medications *if* tolerability is comparable.[4,5] Recently, evidence from pooled analyses of studies of 2 other newer medications that affect both neurotransmitter systems, venlafaxine[4,5] and mirtazapine,[5,6] suggests either more rapid or greater overall efficacy than SSRIs. These findings are not universally accepted, however, and have potentially large commercial implications. Therefore, it is timely to examine both the strengths and the limitations of the methods used to compare the effects of new antidepressants. The primary aim of this review is to determine if current research methods are sufficiently precise to recognize a superior antidepressant effect *if* one actually exists.

RANDOMIZED CONTROLLED TRIALS (RCTs)

The RCT has been the standard for establishing the efficacy of new antidepressants since the 1960s. New antidepressant medications must show significantly greater effects than a double-blind pill placebo in at least 2 well-controlled or pivotal RCTs before approval by the FDA is possible. Generally, between 1000 to 2000 patients have received the new medication in early (phase 2) and later (phase 3) RCTs by the time of FDA approval.[7] This is usually sufficient to determine if the medication has an acceptable tolerability and safety profile as compared to existing standards. It is generally not known, however, if the novel

From the Department of Psychiatry, University of Pittsburgh School of Medicine, Western Psychiatric Institute and Clinic, Pittsburgh, Pa.

Supported in part by MH-30915 (Mental Health Intervention Research Center) from the National Institute of Mental Health, Rockville, Md.

Dr. Thase submitted this Commentary to reflect his discussion of the articles derived from the scientific symposium "New Antidepressants: Light at the End of the Tunnel?" which was held May 10, 2001, at the 154th annual meeting of the American Psychiatric Association in New Orleans, La.

Reprint requests to: Michael E. Thase, M.D., Western Psychiatric Institute and Clinic, 3811 O'Hara St., Pittsburgh, PA 15213-2593 (e-mail: thaseme@msx.upmc.edu).

medication is actually somewhat more or less effective than standard medications.

Why is it so hard to determine relative efficacy? One reason is that it is now apparent that the effects of antidepressant medication in RCTs have been overestimated. For example, approximately 50% of the placebo-controlled RCTs of recently approved antidepressants failed to demonstrate statistically significant effects.[8] This is because the average intent-to-treat, drug-placebo difference among published, placebo-controlled studies of newer antidepressants is only about 18% (e.g., 48% vs. 30%).[1] When unpublished studies are taken into account, a difference of only about 10% can be expected.[8] This amounts to an average advantage for the active medication of less than 2 points on the Hamilton Rating Scale for Depression (HAM-D)[9] when compared to the placebo condition.

These effects may seem surprisingly small, especially if one considers the larger effect sizes commonly observed in the first generation of placebo-controlled trials of TCAs.[1] The relevance of the results of those studies to both contemporary practice and research, however, is limited. A fairly large proportion of the early RCTs evaluated hospitalized patients and, in the 1960s, many of the depressed patients participating in those studies had never before received a trial of antidepressant pharmacotherapy. Diagnostic practices also have changed, with a broadening of the definition of the major depressive disorder. Response to placebo is typically less pronounced in more severely depressed samples, whereas prior treatment resistance generally is associated with lower response rates to both active and placebo interventions.[10] Most contemporary studies enroll highly selected ambulatory subjects, often recruited from media advertisements. It is noteworthy that more recent ambulatory studies that compared SSRIs and TCAs have observed comparable response rates.[1,8] Thus, it is the patients who participate in RCTs that have changed, not the efficacy of antidepressants.

The primary consequence of overestimating the predicted drug-placebo difference is underestimating the number of research subjects needed in order to address the aims of a particular trial. The term *statistical power* refers to the probability of observing a predetermined, between-groups difference under certain standard assumptions (e.g., use of a 2-tailed statistical test, with $\alpha = .05$). A power value of at least 0.80 is generally desired, which means that there is an 80% chance of observing a "true" effect of the predicted magnitude *if* one exists. A power value of 0.80 also means that if the hypothesized effect is true (i.e., Drug A has real antidepressant effects), there will be no more than a 20% chance of a false negative study result. In the jargon of the clinical researcher, this undesirable outcome is referred to as a type 2 error.

Drawing upon the data from the 1960s, a predicted drug-placebo difference of 30% (e.g., 60% for Drug A and 30% for placebo) would indicate that about 75 subjects per group will be needed for the study to have 0.80 power. However, if the expected difference is only 10% as reported by Khan et al.,[8] then more than 300 subjects per group will need to be enrolled. Most studies of novel antidepressants conducted in the 1980s or 1990s enrolled no more than 100 patients per condition.[1,7,8] Therefore, the average study was woefully underpowered and the probability of type II error was unacceptably high (e.g., ≈50%) in contemporary RCTs. Conducting larger studies, the best long-term solution, will demand a far greater financial commitment from the sponsors of clinical research. Such large studies also pose daunting problems with respect to feasibility, implementation, and quality control.

The difficulties encountered by researchers trying to distinguish an active antidepressant from placebo are magnified when the comparison involves 2 effective medications,[1,5] particularly when medications have overlapping mechanistic effects (e.g., a serotonin-norepinephrine reuptake inhibitor should work for most if not all of the patients who respond to an SSRI). Also, the combination of attrition, placebo-responsivity, nonadherence, and "latent" treatment resistance places a low ceiling on the potential number of patients that will actually obtain a "true" response to pharmacotherapy.[5] Again, the likely advantage of a more effective medication in a contemporary RCT may be only about 10%, so at least 300 patients per group would be needed. I am not aware of a single study contrasting a new antidepressant and a standard comparator that enrolled enough patients to have adequate statistical power to detect such a difference.

An active comparator in the typical RCT of a new antidepressant also can be used to help document assay sensitivity.[11] *Assay sensitivity* refers to the study's ability to determine if an antidepressant with known efficacy actually worked in a given patient group. By convention, if both the experimental and the standard treatments fail to surpass placebo, the study is considered to be "failed" rather than negative. Given the poor assay sensitivity of contemporary RCTs (i.e., a 50% failure rate), the manufacturer of a novel medication typically plans to conduct 5 to 8 studies to ensure that bad luck (i.e., repeated type II error) does not "kill" the development of a potentially useful medication. However, only about one half of these studies will have included an active comparator and, with an average value power of 0.50, it is unlikely that a statistically significant difference favoring one drug over another could be found in more than 1 or 2 of the comparisons. These studies thus can be called *equivalence studies* because they usually do not have the power to separate the good from the better.

The manufacturer of a new antidepressant also may conduct a number of comparative studies after the medication has been approved by the FDA. These phase 4 or post-marketing studies are performed either to collect evidence of comparability with a leading product (again, an equivalence study) or to emphasize a particular advantage, such

as a more favorable effect on sleep[12] or sexual function.[13,14] Whether or not a placebo control group is necessary in a postmarketing study is controversial because the efficacy of both compounds has already been established. However, without a placebo group, an equivalence study might just be a negative study. For example, both groups could have 35% response rates, or one drug may appear more effective than another simply because the study group was selected from a subpopulation that was responsive to one type of medication but not the other.

INTERPRETING DIFFERENCES IN STANDARD RCTs

When a significant difference is observed in an equivalence (i.e., underpowered) study, it is unlikely that the observed difference was a chance occurrence. The conventional p = .05 convention ensures that there is no more than a 1 in 20 chance of a false positive finding, which is referred to as a type I error. Significant or nonrandom effects in 1 study are not always attributable to meaningful or across-the-board differences, however. Taking into account the likelihood that such an apparent difference will be used to influence drug sales, however, physicians, as consumers of research report findings, must take a careful look at the validity or integrity of the study.

A number of factors may qualify, invalidate, or even bias a finding of differential efficacy. These issues are illustrated by the following series of questions. First, is the finding robust—does it extend across multiple measures of therapeutic benefit? A finding that is delimited to a single dependent measure is much less convincing than a broader pattern of significant effects. Second, could the study have been biased by subject selection criteria? For example, if the novel medication were found to be more effective than an SSRI in a study group selected on the basis of past treatment failure,[15] these findings may not generalize to decisions about relative efficacy for a treatment-naive population. Third, was the study group unusual in some other respect? The characteristics of the study group can have a pronounced influence on outcomes. For example, monoamine oxidase inhibitors (MAOIs) have been found to be more effective than TCAs in studies of ambulatory patients with reverse neurovegetative features,[16] yet in inpatient studies (i.e., older, more predominantly melancholic patients), the TCAs were significantly more effective than MAOIs.[17] Similarly, among a study group selected for severe insomnia, a medication with stronger sedative effects may have a significant advantage when compared to a medication that is alerting. Fourth, was the study implemented fairly? If the group treated with the optimal dose of Drug A has a significantly better outcome than a group treated with the minimum dose of Drug B, one can not be confident that the same result would hold true if dosing were comparable. Fifth, was the study double-blind and, if so, was the integrity of the blind maintained? Open-label studies are more subject to the expectancies of the clinicians, evaluators, and patients. In the open-label comparisons of various psychotherapies and antidepressants, for instance, the allegiance of the investigative team was strongly related to the outcomes.[18] Although the double-blind partly protects clinicians from such an allegiance effect,[19] the blind is not fully intact when medications that have distinctly different side effects are compared.[20] In summary, the findings of 1 study may not accurately address the question of differential efficacy. Confidence that a meaningful difference does exist ultimately requires replication of findings across diverse groups and settings.

QUANTITATIVE METHODS FOR COMPARING RESULTS ACROSS STUDIES

The principal methods for comparing treatment effects across a group of studies are meta-analyses and pooled analyses.

Meta-Analyses

A quantitative meta-analysis[21] determines the average effect size and the variability of such effects within a set of studies. Various statistical techniques also can be used to examine the impact of study or patient characteristics on the average or standardized effect.

Meta-analyses have been instrumental in establishing that the SSRIs are as effective as the TCAs[22] and better tolerated.[23] A meta-analysis also helped to confirm that, despite overall comparability, TCAs have a modest advantage over SSRIs in studies of inpatients.[24] Further, this modest advantage in hospital-based studies actually represented a mixture of 2 patterns of outcome: a larger difference in studies utilizing "dual" reuptake inhibitor TCAs (i.e., amitriptyline and clomipramine) and virtually no difference in studies of more noradrenergically selective compounds (i.e., the TCAs desipramine and nortriptyline and the related tetracyclic maprotiline).[24] Such differential effects, which can be viewed as a 3-way interaction (i.e., treatment × subclass of treatment × setting), essentially disappear within a broader mixture of comparisons.[25]

There are 3 important shortcomings to meta-analysis. First, like other statistical approaches to grouped data, the power to detect meaningful differences is dependent on both the magnitude of the effect and the number of observations. The unit of observation in a meta-analysis is the number of studies, not the number of patients in each study. Therefore, reliable results are unlikely unless the findings are consistent or a large number of studies have been performed. Confidence in the results of broader classes of comparisons (e.g., TCAs vs. SSRIs)[1,25] is typically greater than among comparisons in which only a handful of relevant studies are available (e.g., fluoxetine vs. paroxetine).[26]

A second and potentially more critical shortcoming of meta-analysis is caused by an artifact called the "file drawer" effect. This refers to the tendency for researchers to publish positive findings and to "file away" negative trials. The file drawer effect does not simply result from a devious suppression of negative results. Studies sometime fail because of clear-cut problems in design or implementation and, even when submitted for publication, negative studies tend to receive less favorable reviews than positive studies. Nevertheless, the net result is that a meta-analysis that is delimited to only the published studies will yield inflated effect sizes, with the magnitude of the inflation for a particular medication proportional to the number of negative studies that have been left behind in the file drawer.

The third limitation of meta-analysis results from arbitrary decisions about the studies or the factors that are included or excluded from the analysis itself. As noted previously, the allegiance of the investigator is associated with the outcomes of unblinded studies.[18] With respect to double-blind RCTs, potentially moderating variables such as source of funding (i.e., industry vs. federal), length of the trial, an estimate of the integrity of the double-blind, and comparability of dosing strategies could affect the results. For example, in the meta-analysis of Freemantle et al.,[25] the decision not to partition the studies into inpatient and outpatient subgroups influenced results.[24]

Pooled Analyses

The second method for examining differences between treatments based on completed studies is a pooled analysis. Also known as a "meta-analysis of original data,"[27] a pooled analysis includes the outcomes of all of the subjects that participated in a related set of studies. The unit of observation is thus the number of patients, not the number of studies. This method results in profoundly more statistical power than available for a meta-analysis, which may be critical when only a small number of comparative studies are available. Examples of relevant pooled analyses in the antidepressant literature include studies comparing (1) the effects of fluoxetine and TCAs on suicidal ideation,[28] (2) the efficacy of MAOIs and TCAs in typical and atypical depressions,[16] (3) the additive effects of psychotherapy and pharmacotherapy,[29] and (4) the efficacy of venlafaxine and several SSRIs.[4]

Pooled analyses are not foolproof: they also can be compromised by the file drawer effect, as well as by inconsistencies in study design and arbitrary decisions about inclusion and exclusion of subjects or studies. The results of 1 large study also can "overwhelm" the findings of a number of smaller studies. Nevertheless, if all data from all subjects from all studies are utilized, the chances for mischief are diminished. Statistical safeguards, referred to as homogeneity or sensitivity analyses,[30] also can be used to ensure that the results are robust. Specifically, it can be shown that evidence of a difference does not depend on 1 study, 1 outcome definition, or 1 subgroup of patients.[4] If duloxetine is really more effective than the SSRIs, a pooled analysis could confirm such a difference after completion of as few as 4 to 6 studies.

SOME THOUGHTS ABOUT DULOXETINE

Drawing upon the experience gained conducting a pooled analysis of venlafaxine and SSRIs,[4] a number of questions about the relative efficacy of duloxetine can be anticipated. First, is evidence of a difference simply the result of a superior effect for duloxetine therapy only among the subgroup of patients who have failed to respond to a previous trial with another SSRI? Answering this question will require collection of reliable information about past treatment history. Second, is the difference similar across all 4 of the SSRIs that are approved as antidepressants in the United States? This line of investigation will necessitate additional studies employing sertraline and citalopram in addition to the already completed studies of fluoxetine and paroxetine. Third, are the differences sustained after the acute phase of therapy? Extension of studies to include continuation phase therapy would be needed to confirm that an apparent advantage in efficacy is not simply due to a faster onset of therapeutic effects. Fourth, are comparable differences apparent in studies of primary care patients? This is important because a majority of antidepressant prescriptions are now written by primary care providers. Finally, is the advantage apparent when the study is not directly funded by Eli Lilly? This question may prove to be the most difficult to answer because there are few alternate funding sources for studies comparing duloxetine with other antidepressants.

A final comment concerns "proof of concept"—the theoretical rationale for an advantage in efficacy. In the case of duloxetine, proof of concept would entail demonstration that greater effects on dysfunctional neural systems mediated by serotonin and norepinephrine are associated with stronger antidepressant effects. Relevant targets could include measures of disinhibited hypothalamic-pituitary-adrenocortical function, such as a greater reduction of cerebrospinal fluid corticotropin-releasing hormone (CRH) levels or plasma cortisol levels in response to a combined CRH-dexamethasone suppression test. Functional imaging studies demonstrating a greater effect on cerebral blood flow or regional glucose metabolism of neural circuits implicated in antidepressant response represent another avenue. Availability of ligands for positron emission tomography studies of norepinephrine systems will, of course, facilitate such studies. A comparison of in vivo measurement of percentage inhibition of serotonin and norepinephrine reuptake transporter sites in relation to antidepressant efficacy would represent a good start. The attention to detail, difficulty of implementation, and cost of such neurobiological protocols undoubtedly will

prevent incorporation of mechanistic studies within large scale RCTs. Instead, smaller focused studies of enriched study groups will be needed to "prove" the concept of dual reuptake inhibition.

SUMMARY

This commentary began with the proposition that the SSRIs have become the standard of comparison for new antidepressants. It is also suggested that the conventional wisdom that all antidepressants are equally effective is no longer true. Rather, it is asserted that the same factors that have compromised the sensitivity of RCTs to detect drug-placebo differences similarly have impaired discriminations between effective and even more effective antidepressants. In order to have the power to make such a distinction, an RCT may need to enroll 300 or more patients per cell. Few studies thus have adequate statistical power. Alternatively, conclusions can be drawn from quantitative methods that combine data from groups of smaller studies. The relative merits and limitations of 2 strategies used to examine the results of comparative studies, meta-analysis and pooled analysis, were discussed. The former method is preferred when there are a large number of relevant RCTs; however, failure to include unpublished data from all relevant studies may inflate results. The latter method, unless biased by selective inclusion of studies or marked heterogeneity of results, is preferred when there are only a handful of comparable studies. Although the task of selecting among a number of good choices remains more clinical art than science, quantitative methods are now available to help determine if there are clinically meaningful differences in antidepressant efficacy.

The conventional RCT provides a low power telescope for visualizing differences between effective medications. If there were a light at the end of the tunnel, it would be difficult to see it. Until methods that can improve the ability to discriminate between an active antidepressant and a placebo are identified and implemented, the ability to see modest differences necessitates the use of statistical methods such as meta-analysis and pooled analysis of original data.

Drug names: amitriptyline (Limbitrol and others), citalopram (Celexa), desipramine (Norpramin and others), fluoxetine (Prozac and others), mirtazapine (Remeron), nortriptyline (Aventyl), paroxetine (Paxil), sertraline (Zoloft), venlafaxine (Effexor).

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. American Psychiatric Association. Practice Guideline for the Treatment of Patients With Major Depressive Disorder [Revision]. Am J Psychiatry 2000;157(suppl 4):1–45
3. Demitrack MA. Can monoamine-based therapies be improved? J Clin Psychiatry 2002;63(suppl 2):14–18
4. Thase ME, Entsuah AR, Rudolph RL. Remission rates during treatment with venlafaxine or selective serotonin reuptake inhibitors. Br J Psychiatry 2001;178:234–241
5. Thase ME, Howland RH, Friedman ES. Onset of action of selective and multi-action antidepressants. In: den Boer JA, Westenberg HGM, eds. Antidepressants: Selectivity or Multiplicity. Amsterdam, the Netherlands: Benecke NI; 2001:101–116
6. Quitkin FM, Taylor BP, Kremer C. Does mirtazapine have a more rapid onset than SSRIs? J Clin Psychiatry 2001;62:358–361
7. Thase ME. How should efficacy be evaluated in randomized clinical trials of treatments for depression? J Clin Psychiatry 1999;60(suppl 4):23–31
8. Khan A, Warner HA, Brown WA. Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: an analysis of the Food and Drug Administration database. Arch Gen Psychiatry 2000; 57:311–317
9. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56–62
10. Thase ME, Kupfer, DJ. Characteristics of treatment-resistant depression. In: Zohar J, Belmaker RH, eds. Treating Resistant Depression. New York, NY: PMA Publishing; 1987:23–45
11. Leber P. Placebo controls. Arch Gen Psychiatry 2000;57:319–320
12. Rush AJ, Armitage R, Gillin JC. Comparative effects of nefazodone and fluoxetine on sleep in outpatients with major depressive disorder. Biol Psychiatry 1998;44:3–14
13. Croft H, Settle E Jr, Houser T, et al. A placebo-controlled comparison of the antidepressant efficacy and effects on sexual functioning of sustained-release bupropion and sertraline. J Clin Ther 1999;21:643–658
14. Ferguson JM, Shrivastava RK, Stahl SM, et al. Reemergence of sexual dysfunction in patients with major depressive disorder: double blind comparison of nefazodone and sertraline. J Clin Psychiatry 2001;62:24–29
15. Poirier M-F, Boyer P. Venlafaxine and paroxetine in treatment-resistant depression: double-blind, randomized comparison. Br J Psychiatry 1999; 175:12–16
16. 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
17. Thase ME, Trivedi MH, Rush AJ. MAOIs in the contemporary treatment of depression. Neuropsychopharmacology 1995;12:185–219
18. Luborsky L, Diguer L, Seligman DA, et al. The researcher's own therapy allegiances: a "wild card" in comparisons of treatment efficacy. Clin Psychol Sci Pract 1999;6:95–106
19. Quitkin FM, Rabkin JG, Gerald J, et al. Validity of clinical trials of antidepressants. Am J Psychiatry 2000;157:327–337
20. Himmelhoch JM, Thase ME, Mallinger AG, et al. Tranylcypromine versus imipramine in anergic bipolar depression. Am J Psychiatry 1991;148: 910–916
21. Lewis G, Churchill R, Hotopf M. Systematic reviews and meta-analysis. Psychol Med 1997;27:3–7
22. Anderson IM, Tomenson BM. The efficacy of selective serotonin re-uptake inhibitors in depression: a meta-analysis of studies against tricyclic antidepressants. J Psychopharmacol 1994;8:238–249
23. Song F, Freemantle N, Sheldon TA, et al. Selective serotonin reuptake inhibitors: meta-analysis of efficacy and acceptability. BMJ 1993;306: 683–687
24. Anderson IM. SSRIs versus tricyclic antidepressants in depressed inpatients: a meta-analysis of efficacy and tolerability. Depress Anxiety 1998; 7(suppl 1):11–17
25. Freemantle N, Anderson IM, Young P. Predictive value of pharmacological activity for the relative efficacy of antidepressant drugs: meta-regression analysis. Br J Psychiatry 2000;177:292–302
26. Edwards JG, Anderson IM. Systematic review and guide to selection of selective serotonin reuptake inhibitors. Drugs 1999;57:507–533
27. Olkin I. Meta-analysis: reconciling the results of independent studies. Stat Med 1995;14:457–472
28. Beasley CM, Dornseif BE, Bosomworth JC, et al. Fluoxetine and suicide: a meta-analysis of controlled trials of treatment for depression. Br Med J 1991;303:685–692
29. 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
30. Greenhouse J, Iyenar S. Sensitivity analysis and diagnostics. In: Cooper H, Hedges L, eds. The Handbook of Research Synthesis. New York, NY: Russell Safe Foundation; 1994:383–398