Neuropsychopharmacology (2005) 30, 445–460

www.neuropsychopharmacology.org

Comparing the Effects of Antidepressants: Consensus Guidelines for Evaluating Quantitative Reviews of Antidepressant Efficacy

Jeffery A Lieberman*[,1], Joel Greenhouse[2], Robert M Hamer[3], K Ranga Krishnan[4], Charles B Nemeroff[5], David V Sheehan[6], Michael E Thase[7] and Martin B Keller[8]

[1]University of North Carolina, Chapel Hill, NC, USA; [2]Carnegie Mellon University, Pittsburgh, PA, USA; [3]University of North Carolina, Chapel Hill, NC, USA; [4]Duke University Medical Center, Durham, NC, USA; [5]Emory University School of Medicine, Atlanta, GA, USA; [6]University of South Florida College of Medicine, Tampa, FL, USA; [7]University of Pittsburgh Medical Center, Pittsburgh, PA, USA; [8]Brown University School of Medicine, Providence, RI, USA

With increasing numbers of treatment options available for patients with major depression over the last decade and the growing body of evidence describing their efficacy and safety, clinicians often find it difficult to determine the best and most appropriate evidence-based treatment for each patient. Systematic reviews utilizing statistical methods that synthesize and evaluate data from a number of studies have become increasingly more available over the past decade. We review major findings and lessons learned from salient examples of quantitative analyses of antidepressant research and provide recommendations for meta-analysts, journal and grant reviewers, and research 'consumers' (ie, clinicians) for conducting, reporting, and evaluating such analyses.
Neuropsychopharmacology (2005) **30,** 445–460, advance online publication, 12 January 2005; doi:10.1038/sj.npp.1300571

Keywords: meta-analysis; randomized controlled trials; evidence-based medicine; quality control; research design; antidepressive agents

INTRODUCTION

On a daily basis, clinicians are bombarded with 'evidence' that one treatment is more effective, safer, faster acting, or simply better for their patients than another. For physicians, and psychiatrists in particular, there perhaps is no therapeutic area that is more competitive than the promotion of antidepressants. An antidepressant with superior efficacy should translate into improved patient care, including a reduction in the morbidity and mortality associated with untreated depression. Moreover, evidence of differential efficacy would have significant commercial advantages for the manufacturer of a patented antidepressant, and the monetary incentives to increase or protect shares of this multibillion-dollar market are considerable. Because presentation of clinical research findings that suggest a therapeutic advantage for one antidepressant compared with other marketed agents has the potential to be so profitable, the results of industry-sponsored research, in particular, are often viewed with great skepticism. It is critical to determine whether an apparent difference (or lack thereof) between treatments is real or the result of bias, flawed study design, or, at worst, misrepresentation of research findings.

According to the Centre for Evidence-Based Medicine (Phillips *et al*, 2001), the various types of evidence can be arranged hierarchically according to strength, with anecdotal case reports and expert opinion (without explicit critical appraisal) at the least definitive end of the spectrum, and evidence derived from randomized controlled trials (RCTs) at the most definitive end. Moreover, there are at least two dimensions to consider when evaluating strength of evidence (Figure 1; Green and Byar, 1984): (1) the 'degree of belief' in the evidence generated by each category and (2) the likelihood that the observed results could be explained by an alternative hypothesis. The progression of the dotted line demonstrates that as the accumulation of scientific knowledge progresses, hypotheses tend to be tested using more rigorous experimental designs (eg, RCTs). When there are several RCTs, methods of research synthesis (eg, systematic reviews) can be used to elucidate central tendencies across studies, thereby fostering the generation of new hypotheses that can be tested via additional RCTs.

From a clinician's standpoint, it can be a daunting task to independently evaluate and integrate the vast body of available published evidence when seeking to answer a

*Correspondence: Dr JA Lieberman, Department of Psychiatry, 7025 Neurosciences Hospital, University of North Carolina at Chapel Hill, CB # 7160, Chapel Hill, NC 27599-7160, USA, Tel: +1 919 966 8990, Fax: +1 919 966 8994, E-mail: jlieberman@unc.edu
Received 22 January 2004; revised 19 July 2004; accepted 12 August 2004
Online publication: 20 August 2004 at http://www.acnp.org/citations/Npp08200404029/default.pdf

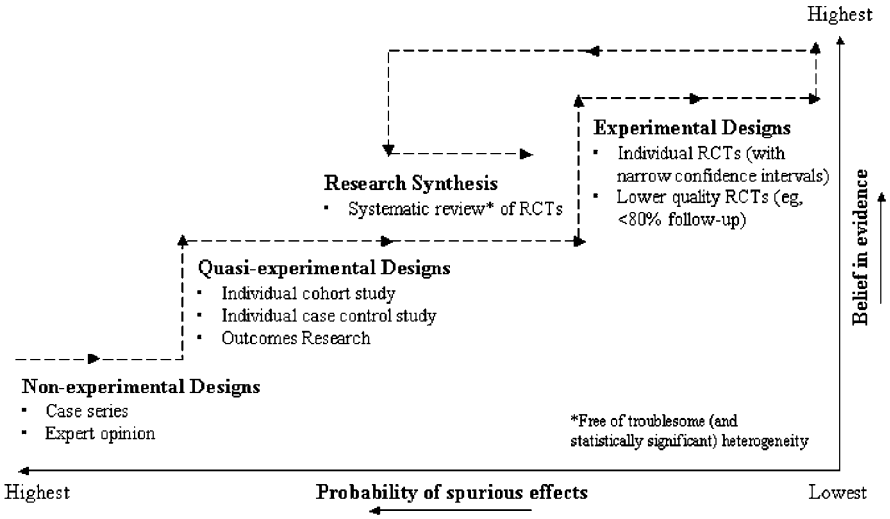

Figure 1 Evaluating sources of evidence. The *y*-axis represents 'degree of belief' in the evidence generated by each category (as recommended by the Centre for Evidence-Based Medicine); the *x*-axis represents the 'likelihood of a spurious effect' (ie, how likely the observed results could be explained by an alternate hypothesis); the dotted line represents the typical direction/order in which the accumulation of scientific knowledge about a question of interest progresses (adapted from Green and Byar, 1984).

clinical question involving the comparative effectiveness of different treatments. As has been the case for nearly 50 years, the RCT remains the gold standard of treatment comparisons. Thus, the optimal state of affairs would be the existence of a series of large, well-designed RCTs involving the full range of treatments, patients, providers, etc. that uniformly yielded the same findings. However, in the absence of a very large, randomized, double-blind, placebo-controlled trial, quantitative methods that integrate findings from the best available evidence can be a useful (albeit limited) means of providing pooled estimates of treatment benefits and/or risks. Moreover, some questions, by their nature, such as the comparative efficacy or safety of one class of medication *vs* another class are often only addressed by means of a synthesis of research, given the complexity and substantial cost entailed in designing a single study to address such questions.

A systematic review attempts to synthesize existing evidence to determine a central tendency across studies using systematic approaches to minimize biases and random errors (Egger *et al*, 2001). It may or may not employ statistical methods to summarize the results of individual studies (ie, a meta-analysis). Because even the most well-designed analysis is not without limitations, the strength of evidence provided by systematic reviews of RCTs remains controversial. We review the advantages and disadvantages of two categories of evidence, the RCT and systematic reviews of RCTs (using meta-analytic approaches), as methods to compare antidepressant treatment effects.

RANDOMIZED CONTROLLED TRIALS IN THE COMPARISON OF ANTIDEPRESSANT EFFICACY

By the mid-1960s, the reference standard in the assessment of therapeutic efficacy in psychiatry, as in other areas of medicine, had become the randomized, double-blind, placebo-controlled trial. However, most RCTs of antidepressants are designed to detect differences in efficacy between an active drug and a placebo. Several factors specific to the assessment of treatments for depression (ie, the frequency of failure to demonstrate a difference between treatments, the study sample size and power to reliably detect differences, selection bias, and patient attrition) must be considered to appreciate the difficulty of reliably detecting a difference between active treatments using contemporary RCT methodology.

An issue that has sparked debate as to the effectiveness of antidepressants in recent years is the fact that a substantial number of clinical trials fail to demonstrate a difference between active treatment groups and placebo (Khan *et al*, 2000; Kirsch *et al*, 2002; Kirsch and Sapirstein, 1998). It has been argued, however, that this phenomenon is largely due to an inflation of placebo response rates and, as a result, an underestimation of the actual expected difference between an effective antidepressant and a double-blind placebo (Thase, 2002). Specifically, whereas evidence from earlier trials revealed active drug–placebo differences of about 30% (Klein *et al*, 1980; Morris and Beck, 1974; Davis *et al*, 1993), group differences of a considerably smaller magnitude are observed in most contemporary studies (about 10–20% in many placebo-controlled trials) (Montgomery, 1999; Thase, 2002; Walsh *et al*, 2002). As a result, in the 'modern' era, even relatively large RCTs (eg, 80–100 patients per arm) are not sufficiently powered to detect a drug–placebo difference of this size.

The ability to distinguish between two active treatment arms is even more difficult than finding differences between drug and placebo. Specifically, if the expected drug–placebo difference is viewed as the upper 'boundary' of potential difference, the likely advantage of a more effective treatment may be as little as 5–10% over an established antidepressant. If one is attempting to design a trial to detect reliably a

difference of such a modest magnitude, much larger numbers of patients (ie, at least 300 subjects per treatment arm) must be enrolled to achieve an acceptable level (ie, >80%) of statistical power (Montgomery, 1999; Thase, 1999).

To date, no published comparative study of newer antidepressants has enrolled a large enough group of patients to have the power to detect *reliably* the differences between two effective treatments (Thase, 2002). One exception to this is the NIMH-sponsored Sequenced Treatment Alternatives to Relieve Depression (STAR*D) project, which will enroll 5000 patients in a comparative treatment trial (Fava *et al*, 2003). Unfortunately, owing to the cost and resources required to conduct studies of sufficient size, the average RCT evaluating antidepressant effects is woefully underpowered. For example, the average number of patients per treatment group across studies in a recent review of 186 RCTs examining the efficacy and tolerability of amitriptyline in comparison to other antidepressants was only approximately 40 subjects (Barbui and Hotopf, 2001). In an analysis of pivotal studies (ie, well-designed, well-controlled studies on which the FDA bases decisions about the efficacy of investigational drugs) for seven newer antidepressants (Khan *et al*, 2000), the average *N* per treatment group was not much larger, with approximately 65–75 patients per group. So constructed, the average study comparing two effective antidepressants would have less than 20% power to find a real, albeit modest (ie, 10%), difference in response rates. Stated another way, the likelihood of a false-negative finding (ie, a type 2 error) would be four times greater than the chance of observing a statistically significant difference.

Why have specific treatment effects apparently declined across the past several decades? There may be a selection bias at work that differs from that of a generation ago. In addition to sample size, the number of centers, number of treatment arms, dosing (eg, flexible dosing *vs* fixed), and different expectation biases can all potentially influence results. In the 1960s, more trials evaluated hospitalized patients, who generally are less responsive to placebo and who appear to have a more robust response to antidepressants (Joyce and Paykel, 1989; Khan *et al*, 2002b; Thase *et al*, 1995). Beyond the issue of in-patient/outpatient status, older studies were more likely to enroll patients with bipolar, psychotic, and recurrent melancholic subtypes of depression. The efficacy of antidepressant interventions was less well known (which may have lowered expectations) and fewer potential participants had ever received an effective course of pharmacotherapy. Contemporary trials, on the other hand, may be enrolling a somewhat different population: highly selected ambulatory patients who are often contacted through the mass media. These subjects may be less severely depressed and are rarely treatment-naïve. Attempts to lessen these problems by restricting enrollment to patients with relatively high levels of pretreatment severity have often, in fact, accentuated them by inadvertently causing an inflation of entry depression scores. Many clinical trials use entry criteria based in part on a minimum score on the same instrument used to evaluate efficacy. Investigators may be motivated, consciously or not, to slightly increase baseline scores in order to enter subjects into the trial. Said scores may then decrease by that same amount once the subject is entered, thus contributing to what appears to be a placebo effect (if not analyzed appropriately).

Another factor influencing the apparent effectiveness of antidepressants is the so-called file drawer effect: the bias introduced by the tendency to publish positive but not negative studies. This bias is most evident when comparing reviews of published studies (Williams *et al*, 2000) with reports that are based on data sets that have been submitted to the FDA for regulatory review (Khan *et al*, 2000; Kirsch *et al*, 2002). On the basis of studies conducted for the registration of new antidepressants from fluoxetine to citalopram, for example, the effects of antidepressants appear to be only about half the size (relative to placebo) once the unpublished studies are taken into account.

The problem of patient attrition also results in an attenuation of possible differences that might exist between treatment groups. Premature attrition results in the loss of outcome data from as many as 30% (Anderson and Tomenson, 1994; Hotopf *et al*, 1997; Montgomery *et al*, 1994; Song *et al*, 1993) of the subjects enrolled in an RCT. There is also a tendency for differential attrition of subjects. Specifically, placebo-treated subjects are more likely to discontinue a trial because they fail to experience an improvement or experience a worsening in their symptoms; patients on active medication may do so because of transient adverse effects that render treatment intolerable, even if effective (Thase M *et al*, 2001). At present, there is no one ideal method to address the problems that attrition poses to analysis of trial results. An analysis of study completers may inflate treatment effects because dropouts tend to have poorer outcomes overall. The current standard used in studies funded by the pharmaceutical industry, that is, an intent-to-treat (ITT) analysis using the last-observation-carried-forward (LOCF) method, is also problematic. Because the final observation before attrition is 'carried forward' for the remaining weeks of the study and substituted for the missing values, early dropouts who are nonresponders skew the results toward lack of efficacy, and treatment effectiveness may therefore be underestimated (as is often the case in depression studies) or overestimated (although this is more common in studies of bipolar mania and relatively rare in depression and anxiety studies) if the time to discontinuation is different between the groups (ie, placebo group dropping out early). Furthermore, the advantages of randomization are also reduced because attrition is not a random process (Leon, 2001). While the conservative LOCF method is deemed by many to be misleading and outmoded (Lavori, 1992; Greenhouse and Junker, 1992), it nevertheless is the standard that the FDA has required for nearly 30 years. Mixed model approaches and the mixed model repeated measures method may have advantages over traditional methods (Mallinckrodt *et al*, 2003), but they are not consistently applied and have not been fully compared. In short, missing data that are not missing completely at random (ie, for reasons unrelated to either the treatment assignment or the response variable) poses problems that are not adequately addressed by traditional statistical methods, which largely ignore the information in the 'missingness' of the data, thus introducing unknown bias.

Contemporary approaches to handling missing data (Mallinckrodt *et al*, 2003, 2001; Entsuah, 1996) are highly preferable to the aforementioned conventional methods commonly used in studies designed for purposes of regulatory approval.

QUANTITATIVE APPROACHES TO INTEGRATING FINDINGS ACROSS STUDIES

Although the RCT remains the gold standard to compare prospectively the differential effects of two treatments, there is increasing recognition that the usual RCT because of its limited sample size is often not sufficiently sensitive to detect reliably differences in relative efficacy (and avoid type 2 errors). Very large RCTs could likely answer these questions, but they are often very (albeit not prohibitively) expensive, and thus unlikely to be industry funded. Moreover, replication of findings across multiple studies helps to ensure that an observed difference (or lack thereof) is real and not the result of bias, flawed study design, or a chance finding (ie, type 1 error). Thus, the practice of evidence-based medicine must entail review and synthesis of the *body* of best available evidence, a task that can prove daunting given the enormous increase in the amount of available evidence over the past several decades (Olkin, 1995). As a result, a basic search of the literature often yields a mind-boggling number of studies relevant to one's particular query.

In an ideal case, a set of studies would consistently demonstrate the same outcome (eg, treatment A is superior to treatment B), in which case, one could simply review the available evidence and confidently assert that, indeed, the former is more effective. This is often not the case, however, with contemporary studies of antidepressants. As described above, it is likely that a high degree of type 2 (false-negative) statistical error contributes to substantial inconsistency in results across studies.

How then does one determine what empirical relationships (if any) have been revealed in such a set of studies? Clearly, a means of making sense of the vast amounts of existing data is warranted. The traditional strategy for evaluating results across individual studies has been the qualitative review, in which the author summarized relevant studies in narrative terms and attempted to integrate findings in a subjective or qualitative manner. Such reviews are often inadequate, however, for integrating inconsistent findings across large numbers of studies (Hunter and Schmidt, 1990). Even experts in a field cannot possibly qualitatively synthesize extensive amounts of data and provide balanced recommendations to guide practice. Over the past 10–15 years, qualitative reviews have been largely supplanted, or at least considerably supplemented, by quantitative approaches such as meta-analysis (ie, an analysis of a group of analyses), which integrate findings from existing data using statistical methods that attempt to approximate that applied to analyses of primary data. This manuscript focuses specifically on two approaches to meta-analysis: those that use summary data from available RCTs and those that pool the original data from a related set of trials.

CHRONOLOGY OF META-ANALYSIS IN ANTIDEPRESSANT RESEARCH

As previously discussed, meta-analyses (when conducted properly) may sometimes permit conclusions about efficacy to be drawn with a greater degree of confidence than is possible with qualitative reviews. Despite clear limitations, most would agree that meta-analyses have been influential in advancing antidepressant research in several key areas (Table 1).

In the field of psychiatry, meta-analysis was first utilized in the context of evaluating the effects of certain forms of psychotherapy (Smith and Glass, 1977). Subsequently, these methods have been used to explore differences in efficacy between psychotherapy and pharmacotherapy (DeRubeis *et al*, 1999; Dobson, 1989; Gaffan *et al*, 1995; Steinbrueck *et al*, 1983), quantify the additive effects of psychotherapy and pharmacotherapy (Conte *et al*, 1986; Thase *et al*, 1997), and to compare the effects of treatment during the acute episode and as prevention of recurrence (Davis *et al*, 1993).

Although the publications occurred somewhat belatedly in the respective life cycles of the tricyclic antidepressants (TCAs) and monoamine oxidase inhibitors (MAOIs), both pooled (Quitkin *et al*, 1993) and conventional (Thase *et al*, 1995) meta-analyses were used to address questions of differential therapeutics that had persisted for more than 30 years. Specifically, the MAOIs were found to be significantly less effective than the TCAs in studies of depressed inpatients (Thase *et al*, 1995), roughly comparable to the TCAs in studies of unselected ambulatory study groups (Thase *et al*, 1995), and significantly more effective in studies focused on atypical depression (Quitkin *et al*, 1993; Thase *et al*, 1995).

Following the introduction of the selective serotonin reuptake inhibitors (SSRIs) in the late 1980s, the question of whether the SSRIs should replace the TCAs as first-line treatment of depression became paramount. Perhaps the most influential example of meta-analyses in mood disorders research is the series by Anderson, Freemantle, and colleagues, which demonstrated that SSRIs are, as a class, comparably effective and significantly better tolerated than TCAs (Anderson, 1998, 2000; Anderson and Tomenson, 1994; Song *et al*, 1993). They also allowed for the detection of differences in subgroup comparisons not apparent in qualitative reviews or single primary studies. For example, despite the overall comparability of these antidepressant classes, when the subgroup of studies of hospitalized or severely depressed patients was examined, significant differences in favor of TCAs were found (Anderson, 1998; Anderson and Tomenson, 1994), which directly paralleled the distinction between the TCAs and MAOIs. Further analysis that evaluated separately studies that used dual-acting TCAs (clomipramine and amitriptyline) and those that used more selective TCAs (imipramine, desipramine, maprotiline) revealed that this effect derived largely from studies of two tertiary amine TCAs, clomipramine or amitriptyline (Anderson, 1998). Although the studies did not investigate the specific mechanism underpinning this difference, it has been presumed that advantage of the tertiary amine TCAs in severe depression was resulted from inhibition of the reuptake of both norepinephrine (NE) and serotonin (5-HT) (Anderson, 1998). Indeed, in-

Table 1 Chronology of Meta-Analysis in Antidepressant Research[a]

Author, year	Type[b]	Summary of major findings
Smith and Glass, 1977	Meta	Psychotherapy and counseling are established as effective treatments, with substantial effect sizes
Steinbrueck *et al*, 1983	Meta	Psychotherapy is at least as effective as drug therapy
Conte *et al*, 1986	Meta	Combination pharmacotherapy+psychotherapy appreciably more effective than placebo, but only slightly more effective than either therapy alone, or therapy+placebo
Dobson, 1989	Meta	Cognitive therapy produced greater degree of change compared with waiting list or no-treatment control, pharmacotherapy, behavior therapy, and other psychotherapies
Beasley *et al*, 1991	Pooled	No increased risk suicidal ideation or suicidal acts with fluoxetine *vs* TCAs or placebo. Improvement of suicidal ideation with fluoxetine did not differ from TCAs and was significantly greater than with placebo
Davis *et al*, 1993	Meta	Therapeutic effect of antidepressant maintenance therapy similar in magnitude to acute effects
Quitkin *et al*, 1993	Meta	Phenelzine was superior to imipramine in all trials, except those that lacked patients with atypical symptoms
Song *et al*, 1993	Pooled	No significant differences in efficacy between TCAs and SSRIs. Similar proportions of dropouts due to lack of efficacy; slightly more dropouts in TCA group due to side effects (18.8%) compared with SSRIs (15.4%)
Anderson and Tomenson, 1994	Meta	Overall, SSRIs and TCAs were comparably effective; significant differences found in various subgroups: studies of severely depressed patients, in-patients, and dual-acting tertiary amine TCAs (ie, clomipramine and amitriptyline)
Montgomery *et al*, 1994	Pooled	SSRIs and TCAs similarly effective; TCAs associated with more discontinuations due to side effects
Gaffan *et al*, 1995	Meta	Cognitive behavioral therapy was superior to pharmacotherapy or other psychotherapy, but findings may have been biased by researcher allegiance
Thase *et al*, 1997	Pooled	Combination therapy (IPT+pharmacotherapy) not more effective *vs* psychotherapy (IPT or CBT) alone in patients with milder depression, but highly significant advantage observed in patients with severe, recurrent depression
Anderson, 1998	Meta	Overall, TCAs, particularly dual reuptake inhibitors, were more effective than SSRIs; amitriptyline was the only individual TCA more effective than SSRIs; significantly fewer discontinuations due to AEs with SSRIs
DeRubeis *et al*, 1999	Meta	In severely depressed outpatients, overall effect sizes favored CBT over pharmacotherapy, but there was no significant advantage for either treatment option
Anderson, 2000	Pooled	Overall efficacy comparable for TCAs and SSRIs; TCAs have an efficacy advantage over SSRIs in hospitalized patients; SSRIs have modest tolerability advantage over most TCAs
Freemantle *et al*, 2000	Meta	No difference in efficacy observed between antidepressant drugs selective for serotonin reuptake and those that act at more than one pharmacological site
Barbui and Hotopf, 2001	Meta	Amitriptyline is less well tolerated compared with tricyclic/heterocyclic and SSRI antidepressants, but associated with a slightly greater proportion of responders
Montgomery, 2001	Meta	Efficacy with paroxetine comparable to TCAs, including clomipramine; significant tolerability advantage for paroxetine
Quitkin *et al*, 2001	Meta	Similar rates of response for mirtazapine and SSRIs. Mirtazapine may have an advantage in time to onset of action
Thase M *et al*, 2001	Pooled	Significant efficacy advantage for mirtazapine compared with SSRIs (fluoxetine or paroxetine) at weeks 2–4; significantly decreased time to remission with mirtazapine
Thase ME *et al*, 2001	Pooled	Significantly greater remission rates achieved with venlafaxine *vs* SSRIs (fluoxetine, paroxetine, fluvoxamine)
Smith *et al*, 2002	Meta	Overall significant efficacy advantage for venlafaxine compared with other antidepressants, including SSRIs as a class, as determined by standardized effect size
Nemeroff *et al*, 2003	Pooled	Significantly greater remission rates achieved with venlafaxine *vs* SSRIs (fluoxetine, paroxetine, sertraline, citalopram, fluvoxamine) as a class; significantly greater likelihood of remission against fluoxetine (but not other SSRIs) when assessed individually.
Thase *et al*, 2003	Pooled	Significantly greater remission rates achieved with duloxetine *vs* SSRIs (fluoxetine, paroxetine)

IPT: interpersonal therapy; CBT: cognitive behavioral therapy; AEs: adverse events.
[a]List is not all inclusive.
[b]Meta = meta-analysis of RCTs; pooled = pooled analysis of original patient data.

patient studies comparing SSRIs with more selectively noradrenergic agents, including maprotiline, nortriptyline, and desipramine, demonstrated no efficacy advantage for the TCAs.

Several comparisons utilizing both meta-analysis of RCTs and pooled analysis of original data have been conducted comparing newer antidepressants to determine if there may also be meaningful differences in efficacy across classes. Different analyses have yielded different answers to this question depending on the design of the analysis, the studies included for analysis, and the decisions of the meta-analysts. For example, Freemantle *et al* (2000) concluded that they were unable to detect an advantage for 'dual-acting' antidepressants compared with single-acting medications such as the SSRIs. However, the investigators did not include the interaction of drug type and study type (ie, in-patient/outpatient status) in their meta-analysis separately, despite the previous findings of one of the authors

(Anderson, 1998). They also had limited access to unpublished data for some but not all agents, which is problematic because of the tendency for selective publication of positive findings (see Identification and inclusion of all relevant studies). The same group of researchers subsequently published a second meta-analysis of both published and unpublished studies ($n=29$ studies for efficacy assessments) directly comparing one dual-acting agent to other types of antidepressants (Smith *et al*, 2002). In this case, the authors concluded that there were significant differences between venlafaxine and the other antidepressants studied, although the overall effect size difference was small (standardized effect size −0.14). Hence, similar meta-analyses yielded different findings depending on how the analysis was conducted (Smith *et al*, 2002; Freemantle *et al*, 2000).

Results from pooled analyses of original data comparing the serotonin-norepinephrine reuptake inhibitors (SNRIs) venlafaxine (Nemeroff *et al*, 2003; Thase ME *et al*, 2001) and duloxetine (Thase *et al*, 2003) with some SSRI treatments suggested a therapeutic advantage for the SNRIs. However, the majority of SSRI-treated patients in the studies included in these analyses received either fluoxetine (in the venlafaxine studies) or paroxetine (in the duloxetine studies), and no studies comparing the SNRIs to escitalopram were included, thus limiting generalizability to the SSRIs *as a class*.

Another critical and statistical consideration when interpreting these findings is the issue of dose comparability. The dosing in these analyses was generally low for all study drugs, and particularly for the comparator SSRI, where the dose was more frequently fixed at the minimum recommended dose compared with the SNRI treatment group. A sufficiently large head-to-head comparison of an SNRI *vs* an SSRI utilizing the full therapeutic dose range for both treatment arms remains to be published. The importance of an assessment of dose comparability when interpreting the results of meta-analyses is discussed in more detail below (see Assessment of dose comparability). In addition, the aforementioned analyses included only data from the SNRI manufacturers. Studies sponsored by other companies have been conducted comparing the SNRI venlafaxine to SSRIs (for example, a sertraline–venlafaxine study is completed), but the primary data set is not available to be included in the aforementioned pooled analysis (Nemeroff *et al*, 2003).

Pooled analyses of studies comparing two other antidepressants presumed to have multiple mechanisms of action, mirtazapine (three studies, Quitkin *et al*, 2001; four studies, Thase M *et al*, 2001) and bupropion (seven studies, Thase *et al*, 2003), with SSRIs suggested that these agents were effective compared to SSRI treatment overall, with possible differences in time to onset of action for the former and a tolerability advantage in terms of sexual dysfunction in favor of the latter.

Finally, a recent meta-analysis of published RCTs and observational studies, commissioned and funded by the United Kingdom Department of Health, evaluated the efficacy and safety of electroconvulsive therapy (ECT) for treatment of depression (UK ECT Review Group, 2003). Although many of the trials included for analysis were older and many were relatively small, the authors concluded that there was a reasonable evidence base supporting the use of ECT, validating its role as an important treatment option for the management of severe depression. Specifically, results showed that ECT was an effective short-term treatment for depression, and suggested an efficacy advantage compared with antidepressant pharmacotherapy. Efficacy differences were observed on the basis of the form of ECT and the dose administered, so no one strategy could be recommended for all clinical situations (ie, one size did not fit all).

The evolution of evidence and, at times, inconsistency of conclusions across different analyses (Jadad *et al*, 1997) underscore an important object lesson regarding methodological limitations. While quantitative statistical methods provide a useful means to synthesize existing information, the results can be heavily dependent on decisions of the meta-analysts and the quality and similarity in designs of the studies selected for analysis (Cappelleri *et al*, 1996; Ioannidis *et al*, 1998; Kjaergard *et al*, 2001; Moher *et al*, 1999b; Greenhouse and Iyengar, 1994). As demonstrated by the previously described examples of antidepressant meta-analyses, inclusion of newly available or previously excluded data can yield results that differ from previous analyses. It should be noted that this phenomenon is a function of the technique rather than the therapeutic area. Similarly conflicting results have also been observed outside the realm of antidepressant research. When comparing meta-analysis of schizophrenia treatments for example, the question of the comparative efficacy and safety of atypical *vs* conventional antipsychotics has had several different answers depending on the specifics of the analysis. A systematic review and meta-analysis of RCTs conducted as a basis for formal development of guidelines for the treatment of schizophrenia in the United Kingdom (Geddes *et al*, 2000) suggested that many of the perceived benefits of atypical antipsychotics may have been due to the use of relatively high doses of the comparator drug used in the trials. Hence, without clear evidence of overall superior efficacy or tolerability, the panel asserted that it was inappropriate to advocate the first-line use of atypicals. However, a more recent systematic review comparing schizophrenia treatments casts doubt on these findings and the consequent recommendations. Leucht *et al* (2003) compared newer generation antipsychotics specifically with low-potency conventional drugs (eg, chlorpromazine). Consistent with previous findings, optimal doses of low-potency conventional antipsychotics (ie, mean doses <600 mg/day of chlorpromazine or its equivalent) had no higher risk of extrapyramidal side effects than new generation drugs. However, the results did not replicate the correlation between comparator dose and advantage of the new generation drugs suggested by Geddes *et al* (2000), who focused on a single high-potency antipsychotic, haloperidol. Rather, new drugs were found to be moderately more effective than conventional ones, irrespective of dose.

All other factors being equal, the greater the number of related, well-designed studies available for analysis, the closer a central tendency can be estimated. But ultimately, just as no one RCT is definitive, no one analysis can put a question entirely to rest (LeLorier *et al*, 1997). In other words, the capacity of meta-analysis to discern valid patterns of results is dependent on the quality of the studies that are included, the quality of the data derived from them, and the similarity or dissimilarity of the designs.

CONSENSUS RECOMMENDATIONS FOR CONDUCTING, REPORTING, AND EVALUATING META-ANALYSES

Given the clear influence of meta-analysts' decisions on the observed outcomes of their investigations and the growing use of meta-analysis to make sense of large bodies of often discrepant studies, it is incumbent on experts in the field to develop guidelines to improve not only the quality of meta-analyses but also the ability of research consumers to discern what is and is not revealed in a given analysis. With these intended goals, we explored specific issues related to the adequate and rigorous conduct and informed, critical evaluation of meta-analyses. Although a comprehensive step-by-step guide to meta-analysis was beyond the scope of this manuscript, many excellent resources are available for researchers providing guidelines for conducting a meta-analysis (Cooper and Hedges, 1994, Egger *et al*, 2001, Hunter and Schmidt, 1990), and a number of organizations have established guidelines for proper reporting of results and/or sets of criteria that can be systematically assessed in judging the quality of individual studies (eg, the CONSORT group (Begg *et al*, 1996; Moher *et al*, 2001)) and meta-analyses (eg, the QUOROM group (Moher *et al*, 1999a) and the MOOSE group (Stroup *et al*, 2000)). Our recommendations are intended to build upon those set forth by others, with an emphasis on the analysis of trials of antidepressant drug treatments. Although the relative lack of large comparative studies in psychiatry compared with other fields of medicine (eg, heart disease) renders our approach particularly applicable to psychiatry, these recommendations generally can be applied to other fields of medicine as well.

The Cochrane Collaboration (1999), an organization formed for the purpose of synthesizing evidence, tracks and appraises research studies according to systematic review criteria. They hold that meta-analysts must be particularly vigilant about choosing studies that employed reliable, established methods for randomizing participants to experimental conditions, and blinding investigators as to the groups to which participants are assigned to guard against the influence of investigator bias on outcomes assessments. A crucial aspect of the Cochrane group's appraisal process involves an assessment of the quality of evidence entered into the review. It is important to note, however, that although quality assessment criteria provide a useful guideline of factors to consider, they should not be interpreted as a 'score' or definitive guarantee of quality. Additional often unaccounted for variables, such as variation between studies in patient and investigator compliance with study protocols, likely influence quality as well. While there is some evidence that studies of poor methodologic quality (according to quality rating criteria) are associated with bias and inflated treatment effects (Moher *et al*, 1998), the degree to which these factors influence treatment effects has not been consistently demonstrated. A recent review evaluating quality measures in 276 RCTs included in meta-analyses concluded that such measures are not reliably associated with the strength of treatment effect across studies and different therapeutic areas (Balk *et al*, 2002). For this reason, we have chosen not to assign values or rank to our recommendations (with the exception that dosing comparability be given primacy over other factors). Evaluation of quality on the basis of established criteria, however, can serve to highlight factors warranting particular scrutiny.

Outlined and discussed below (and summarized in Table 2) are recommendations for researchers, research reviewers (ie, journal editorial boards, journal editors, grant reviewers), and research 'consumers' (ie, clinicians) on specific issues related to conducting, reporting, and evaluating results of meta-analyses of RCTs and pooled analyses of original patient data. We discuss the meta-analytic approach, primary and secondary end points, inclusion criteria, pooling on design, dosing comparability, combining data from placebo- and nonplacebo-controlled studies, and additional factors.

Choice of Meta-Analytic Approach: Meta-Analysis of RCTs *vs* Pooled Analysis of Original Data

In a conventional meta-analysis of RCTs, all relevant studies are identified, selected, and abstracted according to predetermined inclusion and exclusion criteria. Outcomes typically are summarized and expressed in a standardized format for the group of studies (see Cooper and Hedges, 1994; Egger *et al*, 2001, or Hunter and Schmidt, 1990 for a comprehensive description of technique). As is the case with RCTs, the statistical power to detect differences between treatments is dependent on both the number of observations (in the case of conventional meta-analysis, '*n*' is the number of studies) and the magnitude of the effect. Thus, power is relatively limited with a meta-analysis of RCTs unless there are a large number of relevant studies or the average effect is large (which is unlikely in the case of antidepressant studies).

Another kind of meta-analysis that is gaining currency in the area of antidepressant research, as well as in other areas of medicine, is a method commonly referred to as pooled analysis of original data, which differs from a conventional meta-analysis of studies in that *original* individual patient data from a series of studies are pooled, rather than summary statistics such as effect size. The major differentiating factor between the two techniques is that the unit of observation is the outcome of each individual patient rather than the individual study. Thus, when compared with a conventional meta-analysis of RCTs, a pooled analysis usually will impart greater statistical power (Thase, 2002), a distinction providing a particular advantage when the number of studies is small or when modest between-treatment effects are expected.

Although the process may be laborious, modern information systems can facilitate the electronic transfer of large data sets from one research group to another. Nevertheless, identification and compilation of all original data from a comprehensive set of studies (including unpublished completed trials when the findings are known) can prove difficult. However, this should become an increasingly tenable step in the evaluation of new antidepressants because summary statistics for all original data provided to the FDA when a New Drug Application is submitted is now available through the Freedom of Information Act. While the summary statistics does not include raw data, it nevertheless provides a useful means to identify the

Table 2 Recommended Features of Systematic Meta-Analytic Reviews[a]

Feature	Recommendation
Meta-analytic approach	Pooled analysis of original data preferred over meta-analysis when original patient data are available
Primary and secondary end points	*A priori* selection Authors should be able to provide evidence of the preplanned protocol they developed before the analysis was conducted
Inclusion criteria[a]	At a minimum, studies should be randomized, double-blind controlled trials with a sample size of at least 30 patients per treatment arm
	Additional assessment of the quality of individual studies should be systematically evaluated where applicable
	Authors should clearly state and provide rationale for inclusion and exclusion of studies
Pooling on design	Studies included in a meta-analysis should be reasonably similar with respect to design characteristics
	Investigators should conduct and present the results of sensitivity analyses to test the robustness of the findings and ensure the results are not dependent on any one study, study type, or idiosyncratic definition of outcome
Dosing comparability[a]	Should be given highest preferential consideration
	Studies should be included for analysis only if it has been assessed and determined that the design permitted and resulted in fair and comparable dosing of the active treatments
	Mean/modal doses at end point for all active treatments should be reported
	Tolerability data should be included as a proxy measure of dosing comparability
Combining data from placebo and nonplacebo-controlled studies[a]	Results should be reported as (1) both combined, (2) placebo-controlled studies alone and (3) nonplacebo-controlled studies alone
Identification and inclusion of relevant studies[a]	A signed affidavit from the company or agency sponsor should be provided stating that all relevant studies (published and unpublished) meeting the inclusion criteria have been included for analysis (with description of if and why other studies were excluded) A 'funnel plot' analysis should be performed and provided for review during the submission process
Additional factors and/or limitations	The clinical significance of the findings should be assessed Authors should highlight the limitations of their analysis and describe how these limitations might affect the results Source of funding and/or potential conflicts of interest must be clearly stated

[a]Essential features.

existence of unpublished studies, as well as the parties to contact regarding acquisition of the original data. The development of an international comprehensive registration database of clinical trials (Dickersin and Rennie, 2003) would greatly enhance the identification of unpublished studies, although various barriers to creation of such a register (eg, lack of funding, industry resistance, lack of awareness of the necessity) impede a uniform effort to do so.

Given the loss of information that occurs when the unit of observation is the entire study rather than the individual patient, many experts maintain that pooled analysis of original patient data is superior to traditional meta-analysis of RCTs (Olkin, 1995; Thase, 2002; Clarke and Stewart, 2001). This potential advantage is most evident when the aim of the research is to investigate interactions between selected patient characteristics and specific treatments. Although the basic goal of meta-analysis is the assessment of global effectiveness of a treatment, it is also important to understand how efficacy varies across a number of parameters, particularly patient variables such as age, gender, depressive subtype, and so forth. It is possible to miss a clinically important relationship between a subgroup of patients and a subgroup of antidepressants if an investigator fails to model a particular interaction or is unable to do so given the available data. Access to original data increases the feasibility of these assessments. In light of these important potential advantages, we are in agreement that pooled analysis of original data is the preferred approach to a traditional meta-analysis of RCTs when original patient data from a related set of studies are available.

Selection of Primary and Secondary End Points

The retrospective nature of meta-analyses renders the approach vulnerable to exploitation of type I error: it is possible that multiple variables could be examined and only positive findings selectively reported. Such willfully selective reporting of results can promote misleading conclusions about the nature, timing, and scope of the efficacy and/or safety associated with a given treatment, and hence, has the potential for a significant, negative impact on clinical practice. Thus, it is essential that meta-analyses be undertaken with parameters that are prospectively defined. Moreover, meta-analysts should be able to provide evidence of the analytic plan developed before the actual analysis was conducted, including *a priori* selection of parameters to be evaluated with specification of the *a priori* per protocol primary *vs* secondary outcome measures and statistical analysis plan.

To encourage this practice, The Cochrane Collaboration (1999) created the Prospective Meta-Analysis Methods Group, which aims to (1) provide a mechanism to enable the registration of prospective meta-analyses; (2) provide a means of evaluating protocols submitted for registration;

(3) develop appropriate methodological standards for prospective meta-analyses; and (4) provide advice and support to those embarking on (or contemplating) a prospective meta-analysis. They define a prospective meta-analysis as one that identifies, evaluates, and determines study eligibility before the results of the studies become known. This approach can therefore help to overcome some of the problems of retrospective meta-analyses by enabling hypotheses to be specified *a priori*, ignorant of the results of individual trials and mandating that prospective application of selection criteria and *a priori* statements of intended analyses, including subgroup analyses, be made before the analysis is conducted.

An essential component of the registration process of a meta-analysis entails use of a protocol detailing the specific hypotheses/objectives to be tested, eligibility criteria for trial design (eg, requirements for randomization, minimum sample size, blinding), eligibility criteria for the patient population, eligibility criteria for the treatment comparisons, definition of outcomes, details of subgroups, the analysis plan, details of trials identified for inclusion, efforts made to identify ongoing trials, etc. While the stipulation that trials should be included only if their results were unknown at the time they were identified and added to the prospective meta-analysis precludes evaluation of previously published studies, a preplanned protocol including the criteria listed above would indeed help to minimize many of the potential biases and criticisms of this approach.

Study Inclusion/Exclusion Criteria

Although there often is an attempt to include every possible relevant study in a meta-analysis, controversy exists as to whether all studies (regardless of design flaws) should be included or only studies that meet certain minimum standards. If the set of studies included in a given analysis are not reasonably comparable and of a certain minimal quality, it will be difficult to place a great deal of confidence in the validity of its findings. 'Garbage in/garbage out' is a pejorative phrase used to imply that no amount of statistical sophistication can overcome the inclusion of low-quality or 'flawed' studies in a meta-analysis. For example, inadequate randomization and double blinding may contribute to exaggerated estimates of treatment effect and have been purported to be potential factors contributing to discrepancies between the results of large randomized trials and the results of meta-analyses of several smaller trials (Kjaergard *et al*, 2001). It therefore seems prudent for the studies included to meet certain minimal standards of investigational rigor, particularly with regard to use of adequate control conditions and sufficient numbers of subjects to ensure normal distribution of data. Specifically, we recommend that, at a minimum, studies included for analysis should be randomized, double-blind, controlled trials with a sample size of at least 30 patients per treatment arm. In many circumstances, 30 observations are sufficient to allow some summary statistics, such as proportions or means, to have sampling distributions that are approximately normally distributed. Additional assessment of the quality of the individual studies (see the recommendations suggested by the QUOROM group; Moher *et al*, 1999a) should be systematically undertaken and discussed when applicable. In general, these criteria apply to meta-analyses of studies as well as meta-analyses of individual patient data.

At the opposite end of the spectrum, the exclusion of studies on the basis of putative design flaws or other factors, may itself constitute a veiled form of biased sampling, or 'cherry-picking.' Investigators who perform meta-analyses sometimes unknowingly make decisions about what studies to include or exclude that actually favor their implicit assumptions (Luborsky *et al*, 1999). Thus, a decision to include a certain set of studies and exclude others, even for reasons that are apparently 'above board,' may actually impart bias.

Study inclusion criteria, including the minimal standards of quality for incorporating the different studies, should be clearly defined when reporting a meta-analysis. Similarly, the rationale for the exclusion of any studies should be stated. If the results of the current analysis build upon or extend the findings of a previous analysis, any overlap of included studies should be acknowledged. It is also preferable that the studies included for analysis be tabulated in the manuscript (eg, summarizing design characteristics, outcomes, mean dose, dose range in both the protocol and in the results) and listed separately in an appendix.

Pooling on Design

Meta-analysts should seek to minimize and quantify the impact of both clinical or methodologic heterogeneity (inclusion of studies that differ substantially in patients, treatments, design, or outcome) and statistical heterogeneity (variations in the directions and degrees of results among individual studies that are more than would be expected by chance) in order to provide conclusive answers to the hypotheses being tested (Phillips *et al*, 2001; Thompson, 2001; Greenhouse and Iyengar, 1994). Clearly, the issues are inter-related and the degree of the former influences the latter. The extent of clinical or methodologic heterogeneity (ie, the degree to which the studies are not entirely comparable) not only increases the likelihood of discrepancies in the results of the individual studies (thus increasing the degree of statistical heterogeneity), but it also increases the potential for invalid generalizations: for example, a meta-analysis that combined studies of children with studies of elders would encourage extrapolation of the results to mid-life, despite the absence of evidence in this population. While it is particularly problematic to make comparisons across fundamentally different types of research paradigms (eg, combining results from nonrandomized or open-label studies with those derived from double-blind RCTs), even where studies are fit to the same general design template, protocol differences that *prima facie* seem relatively insignificant have the potential to affect the outcome of the analysis (eg, drug dosage and whether dosing was fixed or flexible; Khan *et al*, 2003).

A meta-analysis limited to studies that are identical with respect to design (eg, diagnosis, inclusion and exclusion criteria, assessments, blinding, and length) would yield the least potential for worrisome heterogeneity. There are, however, often too few studies available to apply such selectivity to analyses of antidepressant treatments. Moreover, inclusion of studies with similar but not identical

design characteristics permits an assessment of how the variable in question might affect outcome. Hence, while some control over heterogeneity could be accomplished via exclusion of studies with discrepant designs, methods do exist to evaluate both clinical/methodologic heterogeneity and statistical heterogeneity. When assessing the latter, it is important to remember that the *extent* of statistical heterogeneity is more important than evidence of statistically significant variation in results (Thompson, 2001). A lack of a statistically significant finding should never be interpreted as evidence that no differences exist. As such, it is essential to investigate carefully the influence of specific differences among the studies rather than a reliance on merely the presence or absence of statistical heterogeneity.

Sensitivity analyses enable an assessment of the impact of various differences in the individual studies on the overall outcome. In essence, the data are re-analyzed, using alternative assumptions or methods (Greenhouse and Iyengar 1994; Thompson, 2001). For instance, the analysis is repeated, but now with different, perhaps more restrictive inclusion criteria, which would result in the removal of studies not meeting these criteria from the data set to ensure that the factor in question did not impact the findings of the aggregate data set (see, for example, the approach taken by Thase ME *et al* (2001) and Nemeroff *et al* (2003)). Similar methods can and should be used to assess the sensitivity of the results to the inclusion of individual studies or the use of a particular outcome measure.

In short, a thorough investigation of the impact of any differences in the studies included for analysis and other, potentially relevant decisions of the meta-analysts must be undertaken. A robust effect should not be dependent on any one study, one study type or design characteristic, or use of any one idiosyncratic definition of outcome (Greenhouse and Iyengar, 1994).

Assessment of Dose Comparability

The issue of dose comparability should be given *preferential consideration* in evaluating the quality and validity of a meta-analysis and the individual studies included in such an analysis. In the absence of established comparable dosing between active drug comparison groups, it will be difficult to draw any firm conclusions regarding the validity of the outcomes observed.

Unfair dosing is arguably a particular concern in studies sponsored by the pharmaceutical industry comparing two active medications. In the most extreme case, the minimum therapeutic dose of a competitor drug is selected for comparison with a more optimal dose of the sponsor's drug. If repeated across studies, the impact of this type of design bias will only amplify the difference between the two drugs in the meta-analysis. This is of particular importance in studies comparing two active antidepressant treatments in the absence of a placebo-control arm, which are frequently employed in phase IV postmarketing clinical trials (ie, studies undertaken after an antidepressant has received FDA approval) and in 'ex-US' studies, as some regulatory authorities do not have the same stipulations about use of placebo control as the US FDA.

Ideally, each study included for analysis should have permitted dosing across the full FDA-approved dose range (for the specific indication and population of interest) for each treatment group. Barring this scenario, a general rule of thumb when evaluating dosing fairness and comparability would be that if the study employed minimal doses of drug A, it should be compared with minimal doses of drug B, moderate doses should be compared with moderate doses, and so forth. One specific approach to such an assessment could entail a comparison of the mean doses of the active treatments relative to the maximum FDA-approved dosage or the maximum dosage accepted as therapeutic in clinical practice. For example, if the mean dose of drug A was 25% of the maximum approved dosage and the mean dose of drug B was 75%, clearly, the comparability of dosing would be questionable. In contrast, a study that employs low therapeutic doses of both active treatments might be considered 'fair and comparable' if both arms are proportionately low (eg, the mean doses of drugs A and B are 45 and 50%, respectively, of the maximum recommended doses), although it could also be reasonably argued that such studies do not allow an adequate comparison between treatments and should be excluded from analysis.

It is important to note that strategies based on approved dose guidelines are not without limitations. For some antidepressants, approved dose ranges are relatively wide, which could result in poor tolerability if doses are forcibly titrated to the upper limit (eg, the recommended initial and maximal doses of venlafaxine were reduced after the drug's introduction because of tolerability issues). Conversely, for other medications, such as moclobemide and mirtazapine, clinicians often use doses above the upper end of the regulatory-defined maximum in order to attain optimal efficacy. For yet another antidepressant, reboxetine, the recommended dose range is so narrow (ie, 8–10 mg/day) that most attempts to establish parity will not be satisfactory.

Tolerability data and discontinuation rates due to lack of efficacy and side effects also can serve as useful proxy measures of dosing comparability. For instance, if the incidence of a known dose-related side effect is considerably less than expected (and could not reasonably be explained by other factors), this could be an indicator of an insufficient dose.

To enable assessments such as those described above, we recommend that mean, median, modal, and per protocol dose ranges for all active treatments be reported for the pooled data set and all of the individual studies included in the analysis. Studies that do not utilize fair and comparable dosing of the active treatments should be removed first in a sensitivity analysis to determine the effect of this variable on the outcome.

Combining Data from Placebo- and Nonplacebo-Controlled Studies

The randomized, double-blind, placebo-controlled trial remains the gold standard of treatment comparisons, and it has been persuasively argued by a number of leading experts in psychiatry that the placebo element of the RCT is indispensable in the conduct of antidepressant efficacy trials (Khan *et al*, 2002a, 2000; Walsh *et al*, 2002). However, as stated above, some European countries do not permit

inclusion of a placebo-controlled group. The decision of whether to combine data from both placebo- and nonplacebo-controlled studies (ie, active-comparator only trials) is debatable. A basic tenet of meta-analysis (as previously noted) is that studies must be pooled or combined on the basis of similarity in design, and it could be argued that combining data from placebo- and nonplacebo-controlled trials violates this principle. One reason for this is that studies that include a placebo-controlled group generally enroll a less severely depressed patient group (ie, the prospects of receiving a placebo is a disincentive to some). Another difference is that, when there is no chance of receiving placebo, both patients and their doctors tend to have higher expectations of benefit and, not surprisingly, about a 10% greater likelihood of responding to study treatment.

In contrast, it could be argued that it is inefficient to exclude a large number of otherwise well-controlled studies, and that the impact of this variable can be explored via sensitivity analysis (as described above). Further, although inclusion of a placebo control reduces the absolute magnitude of treatment effects, it does so for both comparison treatments. Hence, the inclusion of data from studies with and without placebo control arms imparts more variance, but in return, more studies are available for analysis.

In light of its well-established effect on outcome, the influence of this variable (ie, presence or absence of a placebo control arm) in a given meta-analysis must be carefully analyzed. As such, we recommend that the results be reported as (1) combined (ie, all studies), (2) placebo-controlled studies alone, and (3) nonplacebo-controlled studies alone.

Identification and Inclusion of all Relevant Studies

As noted above, potential problems with the validity of a meta-analysis arise when, consciously or unconsciously, investigators engage in the selective inclusion of favorable studies. This practice is tantamount to scientific misconduct if studies with less favorable results were knowingly and intentionally excluded (ie, 'cherry-picking'). This is of particular importance for analyses sponsored by the manufacturer of the drug evaluated. In such cases, it is crucial to avoid even the appearance of impropriety. To ensure that this does not occur, some form of binding documentation, such as a signed affidavit from the studies' sponsor, should be provided stating that all relevant studies (published and unpublished) have been made available for review. The authors should also be provided the raw data (preferably) or the study analysis report with all appendices and tables referenced in the report. Studies excluded because insufficient data were available for analysis should be acknowledged and the potential impact of their inclusion (were it possible) on the findings should be addressed. Finally, when comparing the effects of one drug to another drug or class of drugs, attempts should be made to include *all* relevant studies, including those conducted by parties other than the manufacturers of the study drug. It is incumbent on the meta-analysts to ensure that such attempts are undertaken, and company sponsors should not only be willing but eager to obtain data from other companies that include their drug for inclusion in the meta-analyses.

Electronic databases that are readily accessible sources of published studies include MEDLINE, EMBASE, CINAHL, HealthSTAR, HSTAT, PsycINFO, and the Cochrane Library. In order to ensure unbiased sampling of the available published evidence, comprehensive literature searches, including the laborious process of hand-searching all relevant journals, have traditionally been recommended given the inefficiency of the methods commonly employed in a simple search of the medical literature (Dickersin *et al*, 1994). For example, it has been estimated that only about 50% of published RCTs can be located with a comprehensive search of the MEDLINE database (Adams *et al*, 1994).

A recent review of meta-analyses (Egger *et al*, 2003) revealed that studies that are difficult to locate are often of lower quality, thereby potentially increasing bias rather than preventing it. The authors concluded that systematic reviews based on a search of English language literature that is accessible in the major bibliographic databases will often produce results that are close to those obtained from reviews based on more comprehensive searches. A related issue that may perhaps have more bearing on the outcome is whether the retrieved studies are a representative sample of the population of studies of interest (ie, no retrieval bias).

A form of study selection bias that may be more insidious and ultimately problematic stems from the so-called 'file drawer' effect, which refers to the tendency of researchers to publish positive findings and to 'file away' negative trials, so that what is chosen for publication is not representative of the entire extant database. Moreover, studies that report negative findings are less likely to be favorably reviewed and published (Lewis *et al*, 1997). The likelihood then is that a comparison of only published studies will overestimate effect sizes because disproportionately more negative studies will be unavailable. Khan *et al* (2002b) estimated that the file drawer effect may be responsible for the artifactual doubling of effect sizes in some instances. This phenomenon is especially problematic if the 'file drawer' is emptied for some drugs but not others (Thase, 2002). The remedy for publication bias (outside of company sponsorship) is sometimes elusive, given that it is difficult to seek what may or may not exist. Reasonable efforts should be made to contact investigators in order to obtain unpublished data and to follow up preliminary published reports or reports in the so-called 'gray literature' (Gilbody *et al*, 2000) such as conference abstracts.

As a further safeguard against study selection bias in a set of published reports, a 'funnel plot' analysis, which provides a useful indication of whether or not publication bias is operative (Egger *et al*, 1997, 2003), should be performed and provided for review during the submission process. This method takes into account that small studies with negative results are less likely to be published. The funnel plot is created by plotting effect size (bidirectionally on the horizontal axis, with an effect size of zero at the center of the axis) against study sample size (vertical axis), conventionally in ascending order of study size. Given that there will almost certainly be more variability in the outcomes of small studies, there should be scatter of these points across the effect size (horizontal) axis, lending symmetry to the plot at the base of the 'funnel.' If the plot is

asymmetrical at its base, and particularly if it is skewed in the direction of effects in favor of the treatment under investigation, it is possible that some studies with negative outcomes may have been omitted from the analysis (Gilbody *et al*, 2000). It is important to note, however, that there is a degree of subjectivity in interpretation of the plot, and asymmetry may result from factors other than publication bias, such as methodological flaws in the studies themselves (Gilbody *et al*, 2000). The technique may nonetheless be useful as a preliminary red flag of the existence of bias in the study selection. It may also be helpful to do a funnel plot on the dropout rates or other data that do not influence publication bias. Dropouts should be symmetrically distributed because discontinuation rates are rarely significantly influenced by publication bias. The finding of an asymmetrical distribution for effect size and a symmetrical distribution for dropouts in the same studies weighs in favor of publication bias.

Additional Factors to be Addressed when Reporting Results of Meta-Analyses

Meta-analysis in the context of a systematic review should include adequate discussion of the limitations of the report (including both inherent limitations of meta-analysis and those specific to the investigation at hand) and an interpretation of the relevance of the findings to the field. A commonly cited limitation of meta-analysis is the notion that a large data set provides statistical power of a sufficient magnitude to yield results that are statistically significant yet clinically meaningless. While technically true, the possibility of this occurrence does not preclude the desired outcome of results that are both statistically and clinically significant, nor does it prevent an assessment of both. As such, statistical power is not a true limitation of the methodology. Rather, the ability of a large-scale analysis to detect reliably small to moderate statistically significant differences between treatments merely necessitates that the findings be interpreted in light of the clinical relevance of the observed difference. Although the criteria for what constitutes a clinically significant difference can vary greatly on the basis of factors such as the patient population (eg, severity of the illness, treatment history) and the outcome variable (eg, reduction in symptoms *vs* reduction in mortality), there are several useful metrics to gauge the relevance of differences in antidepressant efficacy. One measure of clinical significance is a comparison of the magnitude of the difference between the study drug and active comparator with that of the difference between the active comparator and placebo. Another useful indicator of relevance to clinical practice is the number of patients needed to treat to realize a difference in efficacy or the number needed to harm for comparisons of safety/tolerability profiles (Cook and Sackett, 1995; Daly, 1998; McQuay and Moore, 1997). Finally, the choice of a clinically relevant primary outcome measure can assist in interpreting the significance of observed benefits. In the case of comparisons of antidepressant treatments, a wealth of evidence supports remission as the optimal outcome of acute-phase therapy, in terms of restoration of functioning (Hirschfeld *et al*, 2002; Judd *et al*, 2000a; Miller *et al*, 1998; Mintz *et al*, 1992) and reduction of longer-term risks of relapse and recurrence (Judd *et al*, 2000b; Paykel *et al*, 1995; Simon *et al*, 2000; Thase *et al*, 1992). Even a modest difference between treatments in the proportion of patients achieving remission is therefore likely to be associated with advantages in important 'real-world' domains.

In addition to the relative strength and clinical significance, authors should address the generalizability of the findings, and consideration should be given to factors such as concordance of the findings with studies excluded from the analysis or previous meta-analyses. Finally, given the association of industry sponsorship with proindustry conclusions (Bekelman *et al*, 2003), it is essential that meta-analysts disclose the source of funding for their work and report all potential conflicts of interest.

CONCLUSION

Olkin (1995) aptly observed, 'It is easy to do a meta-analysis; it is hard to do a good meta-analysis.' The complexities of combining, analyzing, and interpreting data from a number of studies, sometimes involving thousands of patients, are often a recipe for problems and biases that can compromise the validity of the analysis' findings. The recommendations we have set forth are intended to guide the conduct and critical evaluation of meta-analyses, particularly those comparing antidepressant treatments. The common thread across the recommendations is an attempt to increase the confidence that can be placed in the findings of a meta-analysis via evaluation of certain key variables, such as the type of analysis that was conducted, the nature and inclusiveness of the study selection criteria, the use of fair and comparable dosing practices across all the studies, and the identification and elimination, where possible, of other biases. Since a meta-analysis is an observational study and it uses data (studies) over which it has no experimental control, the results need to be weighed and interpreted with caution. We hope that adherence to these recommendations will increase the number of both high-quality meta-analyses and informed consumers of the study results.

ACKNOWLEDGEMENTS

The guidelines set forth in this manuscript were based on the proceedings of a meeting held in December 2002 in San Juan, Puerto Rico. Financial support was provided via an unrestricted educational grant from Wyeth Pharmaceuticals. We thank Diane M Sloan, PharmD, for editorial assistance in preparing this consensus report.

REFERENCES

Adams CE, Power A, Frederick K, Lefebvre C (1994). An investigation of the adequacy of MEDLINE searches for randomized controlled trials (RCTs) of the effects of mental health care. *Psychol Med* **24**: 741–748.

Anderson IM (1998). SSRIS *versus* tricyclic antidepressants in depressed inpatients: a meta-analysis of efficacy and tolerability. *Depress Anxiety* **7**(Suppl 1): 11–17.

Anderson IM (2000). Selective serotonin reuptake inhibitors *versus* tricyclic antidepressants: a meta-analysis of efficacy and tolerability. *J Affect Disord* **58**: 19–36.

Anderson IM, Tomenson BM (1994). The efficacy of selective serotonin reuptake inhibitors in depression: a meta-analysis of studies against tricyclic antidepressants. *J Psychopharm* **8**: 238–249.

Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C *et al* (2002). Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* **287**: 2973–2982.

Barbui C, Hotopf M (2001). Amitriptyline v. the rest: still the leading antidepressant after 40 years of randomised controlled trials. *Br J Psychiatry* **178**: 129–144.

Beasley Jr CM, Dornseif BE, Bosomworth JC, Sayler ME, Rampey Jr AH, Heiligenstein JH *et al* (1991). Fluoxetine and suicide: a meta-analysis of controlled trials of treatment for depression. *BMJ* **303**: 685–692.

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I *et al* (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* **276**: 637–639.

Bekelman JE, Li Y, Gross CP (2003). Scope and impact of financial conflicts of interest in biomedical research: a systematic review. *JAMA* **289**: 454–465.

Cappelleri JC, Ioannidis JP, Schmid CH, de Ferranti SD, Aubert M, Chalmers TC *et al* (1996). Large trials *vs* meta-analysis of smaller trials: how do their results compare? *JAMA* **276**: 1332–1338.

Clarke MJ, Stewart LA (2001). Obtaining individual patient data from randomised controlled trials. In: Egger M, Smith GS, Altman D (eds). *Systematic Reviews in Health Care*, 2nd edn. BMJ Publishing Group: London. pp 109–121.

Conte HR, Plutchik R, Wild KV, Karasu TB (1986). Combined psychotherapy and pharmacotherapy for depression. A systematic analysis of the evidence. *Arch Gen Psychiatry* **43**: 471–479.

Cook RJ, Sackett DL (1995). The number needed to treat: a clinically useful measure of treatment effect. *BMJ* **310**: 452–454.

Cooper H, Hedges LV (eds) (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation: New York, NY.

Daly LE (1998). Confidence limits made easy: interval estimation using a substitution method. *Am J Epidemiol* **147**: 783–790.

Davis JM, Wang Z, Janicak PG (1993). A quantitative analysis of clinical drug trials for the treatment of affective disorders. *Psychopharmacol Bull* **29**: 175–181.

DeRubeis RJ, Gelfand LA, Tang TZ, Simons AD (1999). Medications *versus* cognitive behavior therapy for severely depressed outpatients: mega-analysis of four randomized comparisons. *Am J Psychiatry* **156**: 1007–1013.

Dickersin K, Rennie D (2003). Registering clinical trials. *JAMA* **290**: 516–523.

Dickersin K, Scherer R, Lefebvre C (1994). Identifying relevant studies for systematic reviews. *BMJ* **309**: 1286–1291.

Dobson KS (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *J Consult Clin Psychol* **57**: 414–419.

Egger M, Davey Smith G, Schneider M, Minder C (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**: 629–634.

Egger M, Juni P, Bartlett C, Holenstein F, Sterne J (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol Assess* **7**: 1–76.

Egger M, Smith GS, Altman D (eds) (2001). *Systematic Reviews in Health Care*, 2nd edn. BMJ Publishing Group: London.

Entsuah R (1996). Etrank: a ranking procedure for handling missing data in clinical trials: application to venlafaxine extended-release depression clinical trial. *J Biopharm Stat* **6**: 457–475.

Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA *et al* (2003). Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. *Psychiatr Clin N Am* **26**: 457–494.

Freemantle N, Anderson IM, Young P (2000). Predictive value of pharmacological activity for the relative efficacy of antidepressant drugs. Meta-regression analysis. *Br J Psychiatry* **177**: 292–302.

Gaffan EA, Tsaousis I, Kemp-Wheeler SM (1995). Researcher allegiance and meta-analysis: the case of cognitive therapy for depression. *J Consult Clin Psychol* **63**: 966–980.

Geddes J, Freemantle N, Harrison P, Bebbington P (2000). Atypical antipsychotics in the treatment of schizophrenia: systematic overview and meta-regression analysis. *BMJ* **321**: 1371–1376.

Gilbody SM, Song F, Eastwood AJ, Sutton A (2000). The causes, consequences and detection of publication bias in psychiatry. *Acta Psychiatr Scand* **102**: 241–249.

Green SB, Byar DP (1984). Using observational data from registries to compare treatments: The fallacy of omnimetrics. *Stat Med* **3**: 361–370.

Greenhouse JB, Junker BW (1992). Exploratory statistical methods, with applications to psychiatric research. *Psychoneuroendocrinology* **17**: 423–441.

Greenhouse T, Iyengar S (1994). Sensitivity analysis and diagnostics. In: Cooper H, Hedges LV (eds). *The Handbook of Research Synthesis*. Russell Sage Foundation: New York, NY. pp 383–398.

Hirschfeld RM, Dunner DL, Keitner G, Klein DN, Koran LM, Kornstein SG *et al* (2002). Does psychosocial functioning improve independent of depressive symptoms? A comparison of nefazodone, psychotherapy, and their combination. *Biol Psychiatry* **51**: 123–133.

Hotopf M, Hardy R, Lewis G (1997). Discontinuation rates of SSRIs and tricyclic antidepressants: a meta-analysis and investigation of heterogeneity. *Br J Psychiatry* **170**: 120–127.

Hunter JE, Schmidt FL (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Sage Publications: Newbury Park, CA.

Ioannidis JP, Cappelleri JC, Lau J (1998). Issues in comparisons between meta-analyses and large trials. *JAMA* **279**: 1089–1093.

Jadad AR, Cook DJ, Browman GP (1997). A guide to interpreting discordant systematic reviews. *CMAJ* **156**: 1411–1416.

Joyce PR, Paykel ES (1989). Predictors of drug response in depression. *Arch Gen Psychiatry* **46**: 89–99.

Judd LL, Akiskal HS, Zeller PJ, Paulus M, Leon AC, Maser JD *et al* (2000a). Psychosocial disability during the long-term course of unipolar major depressive disorder. *Arch Gen Psychiatry* **57**: 375–380.

Judd LL, Paulus MJ, Schettler PJ, Akiskal HS, Endicott J, Leon AC *et al* (2000b). Does incomplete recovery from first lifetime major depressive episode herald a chronic course of illness? *Am J Psychiatry* **157**: 1501–1504.

Khan A, Khan S, Brown WA (2002a). Are placebo controls necessary to test new antidepressants and anxiolytics? *Int J Neuropsychopharmacol* **5**: 193–197.

Khan A, Khan SR, Walens G, Kolts R, Giller EL (2003). Frequency of positive studies among fixed and flexible dose antidepressant clinical trials: an analysis of the food and drug administration summary basis of approval reports. *Neuropsychopharmacology* **28**: 552–557.

Khan A, Leventhal RM, Khan SR, Brown WA (2002). Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol* **22**: 40–45.

Khan A, Warner HA, Brown WA (2000). 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* **57**: 311–317.

Kirsch I, Moore TJ, Scoboria A, Nicholls SS (2002). The emperor's new drugs: an analysis of antidepressant medication data submitted to the US Food and Drug Administration. *Prev Treat*

5 (Article 23). Available on the World Wide Web: http://www.journals.apa.org/prevention/volume5/pre0050023a.html

Kirsch I, Sapirstein G (1998). Listening to Prozac but hearing placebo: a meta-analysis of antidepressant medication. *Prev Treat* **1** (Article 0002a). Available on the World Wide Web: http://www.journals.apa.org/prevention/volume1/pre0010002a.html

Kjaergard LL, Villumsen J, Gluud C (2001). Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* **135**: 982–989.

Klein DF, Gittelman R, Quitkin F, Rifkin A (1980). *Diagnosis and Drug Treatment of Psychiatric Disorders: Adult and Children*. Williams & Wilkins: Baltimore, MD.

Lavori PW (1992). Clinical trials in psychiatry: should protocol deviation censor patient data? *Neuropsychopharmacology* **6**: 39–48; discussion 49–63.

LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med* **337**: 536–542.

Leon AC (2001). Measuring onset of antidepressant action in clinical trials: an overview of definitions and methodology. *J Clin Psychiatry* **62**: 12–16; discussion 37–40.

Leucht S, Wahlbeck K, Hamann J, Kissling W (2003). New generation antipsychotics *versus* low-potency conventional antipsychotics: a systematic review and meta-analysis. *Lancet* **361**: 1581–1589.

Lewis G, Churchill R, Hotopp M (1997). Systematic reviews and meta-analysis. *Psychol Med* **27**: 3–7.

Luborsky L, Diguer L, Seligman DA (1999). The researcher's own therapy allegiances: a 'wild card' in comparisons of treatment efficacy. *Clin Psychol Sci Pract* **6**: 95–106.

Mallinckrodt CH, Clark SW, Carroll RJ, Molenbergh G (2003). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *J Biopharm Stat* **13**: 179–190.

Mallinckrodt CH, Clark WS, David SR (2001). Accounting for dropout bias using mixed-effects models. *J Biopharm Stat* **11**: 9–21.

McQuay HJ, Moore RA (1997). Using numerical results from systematic reviews in clinical practice. *Ann Intern Med* **126**: 712–720.

Miller IW, Keitner GI, Schatzberg AF, Klein DN, Thase ME, Rush AJ *et al* (1998). The treatment of chronic depression, part 3: psychosocial functioning before and after treatment with sertraline or imipramine. *J Clin Psychiatry* **59**: 608–619.

Mintz J, Mintz LI, Arruda MJ, Hwang SS (1992). Treatments of depression and the functional capacity to work. *Arch Gen Psychiatry* **49**: 761–768.

Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF (1999a). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet* **354**: 1896–1900.

Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A *et al* (1999b). Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess* **3**: 1–98.

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M *et al* (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* **352**: 609–613.

Moher D, Schulz KF, Altman D (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* **285**: 1987–1991.

Montgomery SA (1999). The failure of placebo-controlled studies. ECNP Consensus Meeting, September 13, 1997, Vienna. European College of Neuropsychopharmacology. *Eur Neuropsychopharmacol* **9**: 271–276.

Montgomery SA (2001). A meta-analysis of the efficacy and tolerability of paroxetine *versus* tricyclic antidepressants in the treatment of major depression. *Int Clin Psychopharmacol* **16**: 169–178.

Montgomery SA, Henry J, McDonald G, Dinan T, Lader M, Hindmarch I *et al* (1994). Selective serotonin reuptake inhibitors: meta-analysis of discontinuation rates. *Int Clin Psychopharmacol* **9**: 47–53.

Morris JB, Beck AT (1974). The efficacy of antidepressant drugs. A review of research (1958–1972). *Arch Gen Psychiatry* **30**: 667–674.

Nemeroff CN, Entsuah R, Willard LB, Demitrack MA, Thase ME (2003). Comprehensive pooled analysis of remission data: venlafaxine *vs* SSRIs. Presented at the American Psychiatric Association Annual Meeting, May 2003, San Francisco, CA.

Olkin I (1995). Meta-analysis: reconciling the results of independent studies. *Stat Med* **14**: 457–472.

Paykel ES, Ramana R, Cooper Z, Hayhurst H, Kerr J, Barocka A (1995). Residual symptoms after partial remission: an important outcome in depression. *Psychol Med* **25**: 1171–1180.

Phillips B, Ball C, Sackett DL, Badenoch D, Straus S, Haynes B *et al* (2001). Levels of evidence and grades of recommendation (Centre for Evidence-Based Medicine web site). Accessed March 2003, at http://www.cebm.net/levels_of_evidence.asp#notes.

Quitkin FM, Stewart JW, McGrath PJ, Tricamo E, Rabkin JG, Ocepek-Welikson K *et al* (1993). A subgroup of depressives with better response to MAOI than to tricyclic antidepressants or placebo. *Br J Psychiatry* **21**(Suppl): 30–34.

Quitkin FM, Taylor BP, Kremer C (2001). Does mirtazapine have a more rapid onset than SSRIs? *J Clin Psychiatry* **62**: 358–361.

Simon GE, Revicki D, Heiligenstein J, Grothaus L, VonKorff M, Katon WJ *et al* (2000). Recovery from depression, work productivity, and health care costs among primary care patients. *Gen Hosp Psychiatry* **22**: 153–162.

Smith D, Dempster C, Glanville J, Freemantle N, Anderson I (2002). Efficacy and tolerability of venlafaxine compared with selective serotonin reuptake inhibitors and other antidepressants: a meta-analysis. *Br J Psychiatry* **180**: 396–404.

Smith ML, Glass GV (1977). Meta-analysis of psychotherapy outcome studies. *Am Psychol* **32**: 752–760.

Song F, Freemantle N, Sheldon TA, House A, Watson P, Long A *et al* (1993). Selective serotonin reuptake inhibitors: meta-analysis of efficacy and acceptability. *BMJ* **306**: 683–687.

Steinbrueck SM, Maxwell SE, Howard GS (1983). A meta-analysis of psychotherapy and drug therapy in the treatment of unipolar depression with adults. *J Consult Clin Psychol* **51**: 856–863.

Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D *et al* (2000). Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* **283**: 2008–2012.

Thase M, Haight BR, Richard NE, Rockett CB, Mitton M, Wang Y (2003). Remission rates following therapy with bupropion or SSRIs (poster). Presented at the American Psychiatry Association Annual Meeting, May 2003, San Francisco, CA.

Thase M, Howland RH, Friedman E (2001). Onset of action of selective and multi-action antidepressants. In: Boer JA, Westenberg H (eds). *Antidepressants: Selectivity or Multiplicity?*. Benecke NI: Amsterdam, the Netherlands. pp 101–116.

Thase ME (1999). How should efficacy be evaluated in randomized clinical trials of treatments for depression? *J Clin Psychiatry* **60**(Suppl 4): 23–31; discussion 32.

Thase ME (2002). Comparing the methods used to compare antidepressants. *Psychopharmacol Bull* **36**: 4–17.

Thase ME, Entsuah AR, Rudolph RL (2001). Remission rates during treatment with venlafaxine or selective serotonin reuptake inhibitors. *Br J Psychiatry* **178**: 234–241.

Thase ME, Greenhouse JB, Frank E, Reynolds III CF, Pilkonis PA, Hurley K *et al* (1997). Treatment of major depression with psychotherapy or psychotherapy–pharmacotherapy combinations. *Arch Gen Psychiatry* **54**: 1009–1015.

Thase ME, Simons AD, McGeary J, Cahalane JF, Hughes C, Harden T *et al* (1992). Relapse after cognitive behavior therapy of depression: potential implications for longer courses of treatment. *Am J Psychiatry* **149**: 1046–1052.

Thase ME, Trivedi MH, Rush AJ (1995). MAOIs in the contemporary treatment of depression. *Neuropsychopharmacology* **12**: 185–219.

The Cochrane Collaboration (1999). Prospective Meta-analysis Methods Group (The Cochrane Collaboration web site). Accessed March 2003, at http://www.cochrane.org/cochrane/pma.htm#DOCS.

Thompson SG (2001). Why and how sources of heterogeneity should be investigated. In: Egger M, Smith GS, Altman D (eds). *Systematic Reviews in Health Care*, 2nd edn. BMJ Publishing Group: London. pp 157–176.

UK ECT Review Group (2003). Efficacy and safety of electroconvulsive therapy in depressive disorders: a systematic review and meta-analysis. *Lancet* **361**: 799–808.

Walsh BT, Seidman SN, Sysko R, Gould M (2002). Placebo response in studies of major depression: variable, substantial, and growing. *JAMA* **287**: 1840–1847.

Williams Jr JW, Mulrow CD, Chiquette E, Noel PH, Aguilar C, Cornell J (2000). A systematic review of newer pharmacotherapies for depression in adults: evidence report summary. *Ann Intern Med* **132**: 743–756.

Disclosure Information

Jeffrey Lieberman, MD	Consultant	Abbott Laboratories, AstraZeneca, Bristol-Myers Squibb, Eli Lilly, GlaxoSmithKline, Novartis, Solvay
	Grants/research	AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Forest, GlaxoSmithKline, Janssen, Novartis, Pfizer
	Advisory boards	Abbott Laboratories, AstraZeneca, Aventis, Bristol-Myers Squibb, Eli Lilly, GlaxoSmithKline, Organon, Pfizer
	Stockholder	None
Joel Greenhouse, PhD		None
Robert M Hamer, PhD	Consultant	Wyeth
	Stockholder	Numerous pharmaceutical companies
		Wife (previously employed at Wyeth) is Wyeth stockholder
Martin B Keller, MD	Consultant/ honoraria	Bristol-Myers Squibb, Collegium, Cypress Bioscience, Cyberonics, Eli Lilly, Forest Laboratories, Janssen, Merck Inc., Organon, Otsuka, Pfizer Inc., Pharmacia, Pharmastar, Sepracor, Vela Pharmaceuticals, Wyeth Pharmaceuticals
	Grants/research	Eli Lilly, Forest Laboratories, Merck Inc., Organon, Pfizer Inc., Wyeth Pharmaceuticals
	Advisory boards	Bristol-Myers Squibb, Cephalon Inc., Cyberonics, Cypress Bioscience, Eli Lilly, Forest Laboratories, GlaxoSmithKline, Janssen, Merck Inc., Mitsubishi Pharma Corporation, Novartis, Organon, Pfizer Inc., Pharmacia, Sanofi-Synthelabo, Scirex, Sepracor, Somerset Pharmaceuticals, Vela Pharmaceuticals, Wyeth Pharmaceuticals
	Major stockholder	None
K Ranga R Krishnan, MB, ChB	Grants/research support	Novartis
	Consultant	Abbott, Amgen, GlaxoSmithKline, Johnson & Johnson, Merck, NPS, Organon, Otsuka, Pfizer, Somerset, Synaptic, Vela, Wyeth
Charles B Nemeroff, MD, PhD	Grants/research	Abbott Laboratories, AFSP, AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Forest Laboratories, GlaxoSmithKline, Janssen Pharmaceutica, Merck, NARSAD, NIMH, Pfizer Pharmaceuticals, Stanley Foundation/NAMI, Wyeth-Ayerst
	Consultant	Abbott Laboratories, Acadia Pharmaceuticals, AstraZeneca, Bristol-Myers Squibb, Corcept, Cypress Biosciences, Cyberonics, Eli Lilly, Forest Laboratories, GlaxoSmithKline, Janssen Pharmaceutica, Neurocrine Biosciences, Novartis, NPS Pharmaceuticals, Organon, Otsuka, Sanofi, Scirex, Somerset, Wyeth-Ayerst
	Speakers bureau	Abbott Laboratories, AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Forest Laboratories, GlaxoSmithKline, Janssen Pharmaceutica, Organon, Otsuka, Pfizer Pharmaceuticals, Wyeth-Ayerst
	Stockholder	Corcept, Neurocrine Biosciences
	Board of directors	American Foundation for Suicide Prevention (AFSP), Cypress Biosciences, George West Mental Health Foundation, Novadel Pharma, Heinz C Prechter Fund for Manic Depression
	Patents	Method and devices for transdermal delivery of lithium (US 6 375 990 B1). Method to estimate serotonin and norepinephrine transporter occupancy after drug treatment using patient or animal serum (provisional filing April 2001)
David V Sheehan, MD, MBA	Lectures/ presentations	Abbott Laboratories, AstraZeneca, Boehringer Ingelheim Pharmaceuticals, Boots Pharmaceuticals, Bristol-Myers Squibb, Burroughs Wellcome, Charter Hospitals, Ciba Geigy, Dista Products Company, Eli Lilly, Excerpta Medica Asia, Glaxo Pharmaceuticals, GlaxoSmithKline, Hospital Corporation of America, Humana, ICI, Kali-Duphar, Marion Merrill Dow, McNeil Pharmaceuticals, Mead Johnson, Merck Sharp & Dohme, Organon, Novo Nordisk, Parke-Davis, Pfizer Inc., Pharmacia & Upjohn, Rhone-Poulenc Rorer Pharmaceuticals, Roche Laboratories, Roerig, Sandoz Pharmaceuticals, Schering Corporation, SmithKline Beecham, Solvay Pharmaceuticals, TAP Pharmaceuticals, TGH-University Psychiatry Center, The Upjohn Company, Warner Chilcott, Wyeth-Ayerst Laboratories
	Stock shareholder	Layton Bioscience, Medical Outcome Systems
	Grant/research support	Abbott Laboratories, American Medical Association, Anclote Foundation, Bristol-Myers Squibb Pharmaceuticals Company, Burroughs-Wellcome Pharmaceutical Company, Eisai America Inc., Forest Laboratories, Glaxo-Wellcome, International Clinical Research (ICR), Janssen Pharmaceutica Products, LP, Kali Duphar Laboratories Inc.,

		Eli Lilly & Company, Mead Johnson, Merck Sharp & Dohme Ltd, National Institute of Drug Abuse, National Institute of Health (NIH), Novartis Pharmaceuticals Corp., Parke-Davis, Pfizer, Quintiles, Sandoz Pharmaceuticals Corporation, Sanofi-Synthelabo Recherche (LERS), SmithKline Beecham Pharmaceuticals, Tampa General Hospital-University Psychiatry Center, TAP Pharmaceuticals, The Upjohn Company, Warner Chilcott Pharmaceutical Company, Worldwide Clinical Trials, Wyeth-Ayerst Pharmaceutical Company, Zeneca Pharmaceuticals
	Consultant (list reflects 2000–to present)	Roche Pharmaceuticals (2003), Cephalon (2002–present), Faxmed Inc. (2001), Cortex Pharmaceutical (2001–2002), Parexel International Corporation (2001), Pharmacia (2001–2003), GlaxoSmithKline (2001–present), Sanofi-Synthelabo Research (2000–2002), AstraZeneca (2000–present), Orion Pharma (2000), National Anxiety Foundation (1992–present), USF Friends of Research in Psychiatry, Board of Trustees (1989–present)
Michael E Thase, MD	Consultant	Bristol-Myers Squibb Company, Cephalon Inc., Cyberonics Inc., Eli Lilly & Co., Forest Laboratories Inc., GlaxoSmithKline, Novartis, Organon Inc., Pfizer Pharmaceutical, Pharmacia & Upjohn, Wyeth Pharmaceuticals
	Grant/research	Cyberonics Inc., Pharmacia & Upjohn, Wyeth Pharmaceuticals
	Speakers bureau	Bristol-Myers Squibb Company, Eli Lilly & Co., Forest Laboratories Inc., GlaxoSmithKline, Organon Inc., Pfizer Pharmaceutical, Pharmacia & Upjohn, Solvay Pharmaceuticals, Wyeth Pharmaceuticals
	Major stockholder	None
	Other financial or material support	None

APPENDIX A1

BIBLIOGRAPHY OF SUGGESTED READINGS

Balk EM, Bonis PA, Moskowitz H *et al* (2002). Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* **287**: 2973–2982.

Bekelman JE, Li Y, Gross CP (2003). Scope and impact of financial conflicts of interest in biomedical research: a systematic review. *JAMA* **289**: 454–465.

Cooper H, Hedges LV (eds) (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation: New York, NY.

Egger M, Smith GS, Altman D (eds) (2001). *Systematic Reviews in Health Care*, 2nd edn. BMJ Publishing Group: London.

Egger M, Smith DG, Schneider M, Minder C (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**: 629–634.

Gaffan EA, Tsaousis I, Kemp-Wheeler SM (1995). Researcher allegiance and meta-analysis: the case of cognitive therapy for depression. *J Consult Clin Psychol* **63**: 966–980.

Gilbody SM, Song F, Eastwood AJ, Sutton A (2000). The causes, consequences, and detection of publication bias in psychiatry. *Acta Psychiatr Scand* **102**: 241–249.

Hunter JE, Schmidt FL (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Sage Publications: Newbury Park, CA.

Ioannidis JPA, Haidich AB, Pappa M *et al* (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* **286**: 821–830.

Khan A, Leventhal RM, Khan SR, Brown WA (2002). Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol* **22**: 40–45.

Khan A, Warner HA, Brown WA (2000). Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials. *Arch Gen Psychiatry* **57**: 311–317.

Klein DF (1998). Listening to meta-analysis but hearing bias. *PrevTreat* **1**: Article 0006c.

Klein DF (2000). Flawed meta-analyses comparing psychotherapy with pharmacotherapy. *Am J Psychiatry* **157**: 1204–1211.

LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med* **337**: 536–542.

Lewis G, Churchill R, Hotopf M (1997). Systematic reviews and meta-analysis. *Psychol Med* **27**: 3–7.

Luborsky L, Diguer L, Seligman DA *et al* (1999). The researcher's own therapy allegiances: a 'wild card' in comparisons of treatment efficacy. *Clin Psychol Sci Pract* **6**: 95–106.

Olkin I (1995). Meta-analysis: reconciling the results of independent studies. *Stat Med* **14**: 457–472.

Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B *et al* (2001). Levels of evidence and grades of recommendation (Centre for Evidence-Based Medicine web site). May 2001. Available at http://www.cebm.net/levels_of_evidence.asp#notes. Accessed March 2003.

Sackett DL (2000). *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone: New York, NY.

Thase ME (2002). Comparing the methods used to compare antidepressants. *Psychopharmacol Bull* **36** (Suppl 1): S1–S17.

Walsh BT, Seidman SN, Sysko R, Gould M (2002). Placebo response in studies of major depression: variable, substantial, and growing. *JAMA* **287**: 1840–1847.