Published in final edited form as:
Clin Ther. 2011 December ; 33(12): B49–B61. doi:10.1016/j.clinthera.2011.11.019.

Solving the Antidepressant Efficacy Question? Effect Sizes in Major Depressive Disorder

Paul A. Vöhringer, MD[a][Assistant Professor of Psychiatry] and **S. Nassir Ghaemi, MD[b][Professor of Psychiatry, Director]**

[a]Hospital Clinico Universidad de Chile, Facultad de Medicina Universidad de Chile, Santiago; Research Fellow, Mood Disorders Program, Tufts Medical Center, Boston

[b]Mood Disorders Program, Tufts Medical Center, Tufts University School of Medicine, Boston

Abstract

Background—Numerous reviews and meta-analyses of the antidepressant literature in MDD, both acute and maintenance, have been published, some claiming that antidepressants are mostly ineffective, others that they are mostly effective, either in acute or maintenance treatment.

Objective—To review and critique last and most notable antidepressant MDD studies, and conduct our own reanalysis of the FDA database studies specifically analyzed by Kirsch and colleagues.

Methods—We gather effect estimates of each MDD study. In our reanalysis of Kirsch and colleagues, we correct those analyses for a statistical floor effect, so that relative (instead of absolute) effect size differences are calculated.

Results—Our reanalysis shows that antidepressant benefit is seen not only in severe depression, but also in moderate depression while confirming lack of benefit of antidepressants over placebo in mild depression. Relative antidepressant versus placebo benefit increased linearly from 5% in mild depression to 12% in moderate depression to 16% in severe depression.

The claim that antidepressants are completely ineffective, or even harmful, in maintenance treatment studies involves unawareness of the enriched design effect, which, in that analysis, was used to ensure placebo efficacy. The same problem exists for the standard interpretation of those studies, though: they do not prove antidepressant efficacy either, since they are biased in favor of antidepressants.

Conclusions—In sum, we would conclude that antidepressants are effective for acute depressive episodes that are moderate to severe, but not in mild depression. One can turn the attention-getting conclusions of the review by Kirsch and associates around: Instead of concluding that antidepressants are ineffective acutely except for the most extreme depressive episodes, correction for the statistical floor effect proves that antidepressants are effective acutely except for the mildest depressive episodes. These considerations only apply to acute depression, though. For

Corresponding author for proof and reprints: Dr. Nassir Ghaemi, Tufts Medical Center, Department of Psychiatry, 800 Washington St., Boston, MA 02111. Phone: 617-636-5735. Fax: 617-636-7795. nghaemi@tuftsmedicalcenter.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest statement: SNG has received - Research grant (Pfizer), Research consulting (Sunovion). PAV has nothing to disclose.

maintenance efficacy, antidepressant long-term efficacy is not proven, but neither do those data support the conclusion that they are harmful.

Keywords

Effect size; Antidepressants; Meta analysis; Randomized Clinical Trials

Introduction

Much controversy has surrounded recent meta-analyses and randomized clinical trials (RCTs) of antidepressant efficacy in major depressive disorder (MDD), including in the non-scientific media. In this review, we will use the concept of effect sizes to make clinical and scientific sense of what has become a cultural debate.

We examine here the most prominent RCTs or meta-analyses of RCTs published in the last five years for both acute and maintenance efficacy of antidepressants in MDD. The summary of review of those studies is provided in Table 1.

In acute depression RCTs, one set of studies involves reanalysis of the US FDA database of randomized clinical trials (RCTs) conducted by pharmaceutical companies. The major non-pharmacuetical industry study is the NIMH-sponsored Sequenced Alternatives for Treatment Resistant Depression (STAR*D) project[1]. The pharma trials have been analyzed and reanalyzed by different authors, with the most media attention given to the analysis by Kirsch and colleagues[2]. Other published analyses are also important[3].

Maintenance RCTs for prevention of depressive episodes have been analyzed in the Cochrane database[4], with most of these studies conducted by pharmaceutical companies. The most prominent and highly-marketed and cited recent study of the topic was a 2 year RCT with the antidepressant venlafaxine[5]. A recent reanalysis of the maintenance RCT studies has also examined the impact of antidepressant discontinuation, concluding that antidepressants may cause long-term biological harm. [6]. The STAR*D study also provides data for analysis regarding maintenance prevention of depressive episodes in MDD[1].

Patients and methods

Weanalyzed recent prominent RCTs and meta-analyses which addressed antidepressant efficacy in major depressive disorder (MDD). We examined how assessment of effect sizes could clarify the controversies surrounding acute and maintenance efficacy of antidepressants in MDD. Effect estimates given by these studies are reported along with their 95% confidence intervals (CI) when available.

Results

11 prominent RCTs or meta-analyses of RCTs (2006–2011) are described (Table 1). Each study is described, in Table 1, regarding main aspects of its study design, clinical characteristics, and outcomes. Below we describe those results in more detail, divided in two sections – acute and maintenance studies - and interpret them using effect size concepts. In Table 2, we reanalyze the results of one prominent meta-analysis [2] to correct for a statistical floor effect in mild depression. In so doing, we find that the claim that antidepressants are only effective in severe depression, but not in moderate or mild depression, is wrong. They are also effective in moderate depression, as explained below.

Discussion

Acute depression

Analyses of the FDA database—The pharmaceutical industry is obligated to submit all its data, positive or negative, regarding studies of drugs which receive FDA approval. Through the Freedom of Information act, scholars have begun to get access to these FDA records. Previous systematic reviews of such studies of antidepressants in MDD have shown that many negative studies have gone unpublished. Turner and colleagues showed that about 94% of the published literature on antidepressants in MDD demonstrates efficacy (positive studies), but when the unpublished FDA database is included, only 51% of all such studies (published and unpublished) are positive. The standardized effect size fell from about 0.37 to 0.31 after including the negative unpublished studies, both effects being in the mild range [7].

The same year as the above analysis, another was published by Kirsch and colleagues [2] with a smaller sample of the FDA database (less than half the size of the analysis by Turner's group). It confirmed an unstandardized effect size of 0.32, similar to the prior analyses by Turner and associates. The key difference was that this meta-analysis [2] focused on a clinical significance criterion set in the UK by the National Institute for Clinical Excellence (NICE): a 3 point difference on the Hamilton Depression Rating Scale (HDRS), or a 0.5 standardized effect size difference. As shown in the table 1, the results of this reanalysis fell short of those effect size cut offs, except for severe depression. In follow-up popularizations, the first author of that meta-analysis [2] has interpreted his analysis as indicating that antidepressants, in general, do not have clinically meaningful effects in MDD, in general. In the scientific paper, the authors were more circumspect though still critical; they attributed antidepressant benefit only to “the most extremely depressed patients” [2], although a HDRS cut-off of 28 is not, in clinical practice, descriptive of “the most extremely depressed” patients. Many such patients have HDRS scores in the 30s or above. In this meta-analysis, the drug-placebo difference varied based on severity of illness, approximating zero at a HDRS of 24, and reaching about 3 points at a HDRS of 28. The authors note that this effect was due to changes in placebo response, which fell with increasing severity, rather than antidepressant response, which was consistent. While this was noted, the authors never grappled with its meaning. It would seem that mild depression is highly placebo responsive while severe depression is not. The authors seem to presume as if this means that antidepressants are not more effective in severe depression; in fact, they are. The loss of placebo “response” may not be the loss of a response of anything at all; placebo reflects, in part if not in whole, the natural history. Severe depression just does not go away rapidly; if one does not treat it, it remains. Antidepressants treat it. They are effective. The authors ignore this fact because they ignore the importance of the natural history in assessing treatment effects.

Our reanalysis of the FDA meta-analyses: Correction for a floor effect disproves claims of antidepressant inefficacy—A key statistical issue in such comparisons of milder versus more severe depression, when using absolute effect sizes, is a floor effect. With a lower baseline HDRS score, the same drug-placebo effect (eg, 50% decrease in scores) would produce smaller absolute differences (eg, 20 to 10 HDRS points, a 10 point difference) compared to a higher baseline HDRS score (30 to 15 HDRS points, a 15 point difference). In this meta-analysis, the drug placebo difference, when adjusted for baseline severity of illness, increased in nonstandardized effect size from 0.32 to 0.40. In other words, some of the apparent lack of benefit of antidepressants in milder depression may be an artifact of this floor effect. The authors reported this result in a table, but did not comment on it [2].

Another way to address this problem would be to report the *relative* (not absolute) drug-placebo difference relative to the baseline severity of depression. This was not reported in the above analysis. We provide such an analysis, for the first time, in this paper as follows:

Table 2 shows percent differences in drug effect, with the absolute change in the drug group divided by the baseline HDRS score. Using this relative effect measure, antidepressants were somewhat less effective in milder depression (HDRS with baseline scores at 24 or less) versus severe depression (baseline HDRS 28 or above)relative antidepressant versus placebo benefit increased linearly from 5% in mild depression to 12% in moderate depression to 16% in severe depression. The studies used in the meta-analysis [2] had a weighted mean baseline HDRS of 25.5{Horder, 2010 #15737}. Using that baseline and the absolute improvement rates near those reported in the study (9.6 for drug, 7.8 for placebo) but widened to meet the NICE criterion of 3 points or greater difference (i.e., ≥ 10 for drug versus ≤ 7 for placebo), we can calculate that the NICE criterion would have been met with relative drug improvement of 39.2% (10/25.5) versus relative placebo improvement of 27.5% (7/25.5), for a drug-placebo relative difference of 11.7%. With this definition of the NICE criterion, antidepressants still do not meet that definition in mild depression (HDRS<24), but they do meet it for both moderate (24 <HDRS > and severe depression.

In this reanalysis, we used the same severity cut-offs as used by the authors of the meta-analysis: HDRS scores below 24, versus 24–28, versus above 28; we label these three groups as mild, moderate, and severe, respectively. Despite analyzing their data in these three groupings, the authors of the meta-analysis claimed to use the American Psychiatric Association's criteria for severity of symptoms (based on HDRS scores): Mild (HDRS = 8–13), Moderate (HDRS = 14–18), Severe (HDRS = 19–22), and Very severe (HDRS>22) [9]. In so doing, they ignore the obvious fact that symptoms differ from episodes: the typical major depressive episode (MDE) produced HDRS scores of at least 18 or above. Thus, by using symptom criteria, all MDEs are by definition severe or very severe. Clinicians know that some patients meet MDE criteria and are still able to work; indeed others frequently may not even recognize that such a person is clinically depressed. Other patients are so severe they function poorly at work so that others recognize something is wrong; some clinically depressed patients cannot work at all; and still others cannot even get out of bed for weeks or months on end. Clearly, there are gradations of severity within MDEs, and the entire debate in the above meta-analysis is about MDEs, not depressive symptoms, since all patients had to meet MDE criteria in all the studiesincluded in the meta-analysis (conducted by pharmaceutical companies for FDA approval for treatment of MDEs).

The question, therefore, is not about severity of depressive symptoms, but severity of depressive *episodes*, assuming that someone meets DSM-IV criteria for a major depressive episode. On that question, a number of prior studies have examined the matter with the HDRS and with other depression rating scales, and the three groupings shown in table 2 correspond rather closely with validated and replicated definitions of mild (HDRS <24), moderate (HDRS 24–28), and severe (HDRS>28) major depressive episodes [10–12].

In other words, *if one corrects for the statistical floor effect* (which was also shown in the data reported by the authors in a regression model correcting for baseline severity of illness), then the claim that antidepressants are only effective in the most extreme depressive conditions is disproven. Antidepressants are also effective in moderate, as well as in severe, depression.

In sum, one can turn the conclusions of Kirsch and associates around: They said that antidepressants are generally ineffective except in “the most severely depressed patients”

[2]. The reality is that this analysis proves that antidepressants are generally effective except in the mildest depressive episodes.

Other reanalyses of the FDA meta-analyses: Correction of pooling methods increases effect size to clinical significance—Horder and colleagues [9] also reanalyzed the dataset in the above meta-analysis [2], and noted two errors in calculation of pooled effect size differences. In the original meta-analysis, the authors pooled all the antidepressant effect sizes (drug effect pre and post treatment), and then they pooled all the placebo effect sizes (pre and post treatment). Then they subtracted these two pooled effect sizes. This is statistically incorrect. Pooled differences need to be assessed within each study to maximally incorporate the benefits of randomization within each study; thus for a first study, the difference between drug and placebo should be calculated; for a second study, the same difference should be calculated; and so on. The pooled effect size for the meta-analysis should be the sum of each effect size difference between drug and placebo for each study, divided by the number of studies. Horder and colleagues corrected the calculation using this approach to pooling effect size differences. They also used the absolute effect size difference on the HDRS, since all the studies used the same scales. They correctly note that there is no need to use a standardized effect size measure (like Cohen's d) when all studies use the same outcome (HDRS); standardized effect sizes are used to allow for an attempt to equalize different outcomes (such as HDRS compared to different depression rating scales). By standardizing, the mathematical manipulations introduced may alter one's results somewhat, making them both less interpretable and less valid. Finally, Horder and colleauges used the most valid measure of meta-analytic effect – the random effects model, as opposed to fixed effects as in the original review [2]. Fixed effects models assume that all studies have similar variablities; when one is comparing studies with different drugs, in different patient populations varying by severity of illness, which the authors showed was an important predictor of response, then the fixed effects assumption is not valid. The random effects assumption entails the view that studies differ from each other in important respects; fixed effects model only correct for sample size, and assume no other kinds of error, while random effects model introduce a second correction for presumed error.

When making these three corrections – a) pooling drug-placebo difference study by study, b) using the absolute HDRS effect size difference only, c) using a random effects model for the meta-analytic summary – Horder and colleagues found a much higher effect size, HDRS difference of 2.70, quite near the NICE cut-off of 3, as opposed to the clearly low HDRS difference effect size of 1.80 in the original meta-analysis.

Other reanalyses—Fountoualakis and Moller [13] also reanalyzed the same meta-analysis that has received great attention [2]. They made one correction, a weighting of the mean difference in each study for sample size. In so doing, they find a slightly larger effect size of 2.18, but not large enough to meet the NICE criterion. They also report that when examined by drug type, venlafaxine and paroxetine meet the NICE criterion of 3 point improvement, while nefazodone and fluoxetine do not. We would add that the nefazodone studies all involve mild depression (no baseline HDRS above 25), and thus lack of benefit may reflect mildness of depression per se (where natural history leads to rapid recovery, as discussed below), rather than inefficacy of the drug itself.

Other discussions of these meta-analyses have included a commentary by Ioannidis [3], who concludes, like the group critical of antidepressants [2], that these agents are largely ineffective. Ioannidis adds a quantitative simulated analysis of a situation where, if one assumes that the true effect size is small(like 0.20), then, with moderate or larger variability (due to small sample sizes), reported effect sizes would always be larger than the real effect size of 0.20. In other words, Ioannidis points out that most effect sizes are probably inflated estimates of the real effect sizes, especially if studies are not large. This is relevant, he thinks, to the antidepressant literature. Thus, if we are debating how close we are to the NICE threshold of 3 points, and the issue is whether 2.70 is close enough compared to 1.80, Ioannidis suggests that we should adjust, conceptually, for somewhat lower effect sizes than these exact numbers.

His review has been challenged by Davis and colleagues [14]. They emphasize that however one analyzes the antidepressant literature, the effect size of benefit with antidepressants over placebo is not zero. An effect size of 0.31 is a small effect size, but it is still an effect, in some people. They point out that oncology studies, for instance, support the use of treatments with much smaller effect sizes, because those conditions are otherwise terminal. They emphasize that because of the notable morbidity and mortality of severe depression, at least, *any* drug benefit is valuable. They base their conclusions on a narrative review of prior meta-analysis and major RCTs; their discussion of maintenance efficacy studies is uncritical, as we'll see.

STAR*D—These sometimes rancorous debates about the pharmaceutical-industry studies are limited by the fact that these are pharmaceutical-industry studies. These studies were conducted for FDA registration, to market drugs for profit, not to find out the truth in any economically disinterested fashion. This is why the huge NIMH-sponsored double-blind, randomized STAR*D study is of major importance in addressing the question of antidepressant efficacy. It was conducted by academic sites that carefully organized and conducted their studies to meet NIMH standards, not by for-profit research groups that tried to meet pharmaceutical-industry standards. The latter setting often involves paying patients money to participate, sometimes at rather high rates, and there are well-known concerns about the “fudging” of data to meet recruitment goals. Further, the FDA database involves pooling many different studies with different drugs in different study subjects, sometimes in different countries. The heterogeneity introduced by such differences is the bane of such large meta-analyses. Such heterogeneity is a type of confounding bias, making the results of these huge meta-analyses somewhat doubtful, since the pooled results of studies are not randomized. Only the data within each study is randomized, and thus free of confounding bias.

This is not a minor issue, and one that many of the debaters ignore. A meta-analysis can never be taken at face value, because it is not randomized; meta-analyses are always observational, and thus biased, to greater or lesser degree. All things being equal, a large single RCT is more valid than a meta-analysis, because the former is randomized and the latter is not.

Thus, a single huge RCT, like STAR*D, is more valid, based on confounding bias concerns, than the huge FDA meta-analyses of multiple RCTs. The main limitation of STAR*D is the absence of placebo controls, which does not allow us to definitively use it to answer the question whether antidepressants are better than placebo. However, if antidepressants were nothing but placebos, we would be able to legitimately expect rather low response rates in STAR*D, especially in severe depression.

In STAR*D, the main purpose was to see which antidepressant treatments were effective in those who failed to remit initially with a single antidepressant trial. The antidepressant chosen was citalopram, a typical SRI, and it was given open-label initially, so as to identify non-responders who were then randomized to various steps of other treatments. Perhaps not too surprisingly, initial response openly to citalopram was about 50%, and initial remission about 30% [15]. The remaining subjects were then randomized to three sequential stages of treatment. They continued down the tree of options if they failed to remit in any phase, and as long as they were willing to stay in the randomized studies. In the second stage of treatment (either switching to a different antidepressant or augmenting with one), a similar rate of acute response was seen (about 50%). However by stages 3 and 4, despite using agents previously shown to be most effective (like tricyclic antidepressants and monoamine oxidase inhibitors or lithium augmentation), acute response rates ranged around 20%. Further, by stages 2 and onward, remission and response rates were about the same (i.e., better response was not seen with a more liberal definition of improvement than used for remission). As the authors of STAR*D comment in one paper, one can read these results as good news in the sense that one can conclude, with multiple phases of treatment, that about 60% or so of patients will respond acutely (>50% improvement in depressive symptoms) [1]. This seems much higher than one would expect from natural history.

Readers should clearly understand: after the initial citalopram treatment phase, STAR*D was a double-blind randomized study (though without a placebo arm). All stages 2 onward discussed above involved randomized, not observational, data, and are as valid as any standard randomized clinical trial.

The results are not definitive, though, given the FDA database analyses, since the mean initial HDRS score was 21.8 in STAR*D, consistent with mild depression. In that group, one would expect much spontaneous recovery by natural history or the non-specific benefits of a placebo response. This possibility cannot be ruled out.

Maintenance efficacy of antidepressants in MDD

Biases of the enriched design for maintenance efficacy—Before examining analyses of maintenance studies in MDD, we should understand how such studies are designed so as to appreciate why they are mostly biased in favor of antidepressants.

Most maintenance studies of antidepressants begin, before the study begins, with an acute major depressive episode, frequently treated with the antidepressant being studied. Then, patients who respond to the antidepressant are entered into the maintenance study. Those who do not respond or tolerate the antidepressant are not included; this already biases the study in favor of the antidepressant. Then patients are followed for one to two years. The majority of patients relapse in the first six months of follow-up, however. This does not prove maintenance efficacy, because the maintenance phase of treatment in MDD does not begin until one year after the acute episode ends, which is when we know that the natural remission of an acute depressive episode happens [16–18]. Thus, one year or longer is the relevant time frame to assess prevention of new episodes; there is a clear consensus on this issue in the depression literature. Even if one did not want to focus on one year, it would at least be reasonable to say that at least 6 months or longer after the acute episode is needed to assess maintenance efficacy.

Most maintenance RCTs fail, as we will see, to pass this simple test.

This problem has been much discussed in the bipolar disorder literature [18], and we have related it previously to the maintenance studies of neuroleptics in bipolar disorder [19]. The early literature with lithium included both prophylaxis and relapse prevention methodologies. In the prophylaxis design, “all comers” are included in the study: in other words, any patient who is euthymic, no matter how that person got well, is eligible to be randomized to drug versus placebo or control, including those with recent manic or depressive episodes. In the relapse prevention design, typically only those patients who acutely respond to the drug being studied are then eligible to enter the randomized maintenance phase. Those who responded to the drug are then randomized to stay on the drug or be withdrawn from it (usually abruptly, sometimes with a taper) and switched to placebo. The “prophylactic” and “relapse prevention” designs are obviously not addressing the same questions about drug efficacy. In the lithium studies in which the relapse prevention design was used (i.e., only initial lithium responders to acute treatment were included), there was evidence of lithium withdrawal following acute treatment in the placebo group [20, 21]. The problem is that by design those who reach the “maintenance” phase and are treated with placebo are in fact persons who responded acutely to the study drug (lithium) and then get abruptly discontinued. Thus, if the placebo relapse rate is very high and almost exclusively limited to the first 1–2 months after study initiation, then one is observing a withdrawal effect involving a relapse back in to the same acute episode that had just been treated rather than a new episode, i.e. a recurrence. In other words, the relapse prevention design methodology confounds prevention of relapse back into the index episode with prevention of a new episode.

Besides the above problem of withdrawal relapse, a key aspect of the relapse prevention design is that it is definitely biased in comparison with active controls, and it is very likely biased against placebo as well. This is simply seen by realizing that though such studies are randomized, they are only randomized after preselecting all subjects to be randomized as responsive to *only one* of the two arms of the study. Thus, randomization is, in effect, instituted after the study has already been biased in favor of one of the two treatments.

To put it simply, if some people like chocolate ice cream and others like vanilla ice cream, and we preselect only those who like chocolate ice cream to be randomized to again receive chocolate ice cream versus vanilla ice cream, we will find that most chocolate ice cream-lovers will continue to prefer chocolate ice cream. This does not prove that chocolate ice cream is superior to vanilla ice cream.

The same principle will apply to studies in which patients are preselected to respond to the study drug, and later randomized to stay on study drug or receive placebo. Again, the study would be biased in favor of study drug, and would not prove inherent superiority of study drug over placebo.

A truly randomized study would have to either preselect subjects to be responsive to both treatments being studied, or, as in the traditional prophylaxis study, make no preselection at all.

These inherent biases of the enriched maintenance design will be key to analyzing meta-analyses of the maintenance antidepressant efficacy literature. None of those reviews, save one, understand the relevance of the enriched design, and thus they draw incorrect conclusions, both for and against antidepressants.

Maintenance RCTs—The standard review of maintenance efficacy of antidepressants often involves reference to the Cochrane collaboration meta-analysis of published studies. In that report, 10 studies with SRIs (n=2080) and 15 with TCAs (n=881), mostly with one year follow-up, showed maintenance benefit versus placebo. The longest follow-up with modern antidepressants was two years with venlafaxine [4]. An obvious problem with simply stating the results this way is that this meta-analysis does not address the issue of publication bias. If the acute antidepressant studies are any indicator, it is likely that some negative maintenance studies with antidepressants in MDD exist, but are unpublished, and they would reduce this reported effect size.

A more important issue is the problem of enriched maintenance designs, which bias studies in favor of drug enrichment (or placebo, if analyses are enriched in the opposite direction, as discussed below). The only analysis of antidepressants RCTs in MDD which has addressed the problem of enrichment is a recent paper by Briscoe and El-Mallakh [22]. They address the problem of enrichment by limiting data analysis to 6 months or longer after the acute depressive episode. By so doing, they exclude those who relapsed soon after the maintenance study started, right after the end of the acute episode. Those who received antidepressant and were switched to placebo would relapse rapidly in the first few months of the maintenance treatment; this discontinuation effect is an artifact of the enriched design, and would not, in this view, demonstrate true recurrence of a new episode, but rather immediate relapse into the same episode that had been present in prior weeks. Only 5 RCTs provided data on relapse rates before and after 6 months. Limiting analyses to those studies, the researchers found that, as expected given the biases of the enriched design, the majority of relapses (about 2/3) occurred in the first six months of follow-up; these are not new episodes of depression, but withdrawal relapse into the same acute episode that had just occurred a few weeks or months earlier, before the maintenance study began. In the one-third of relapses occurring after 6 months, and thus testing the proposition of whether new episodes were truly being prevented, 4 of 5 studies found no benefit with antidepressants over placebo.

The venlafaxine PREVENT maintenance study—Many authors cite a recent long, large study of venlafaxine (VNL) as evidence for antidepressant maintenance efficacy in MDD [5].

This study purports to show major benefits with VNL for maintenance treatment of MDD but it really reflects what we might call “super-enrichment”: the study repeatedly picks out those who respond to venlafaxine and rerandomizes them to VNL or placebo, thus repeatedly selecting a smaller and smaller group of highly-VNL responsive patients. By two years, this small group is indeed very responsive to VNL, but hardly generalizable to a patient who might newly be prescribed VNL.

The specific data are as follows: In that study, 1096 MDD patients initially received, for acute depression, venlafaxine (VNL) or fluoxetine. 715 responders were enrolled in 6 month blind continuation on the same treatment. 258 (35.9%, 258/715) of those acute responders remained well at 6 months and entered maintenance phase A for one year treatment (randomized to VNL vs placebo)[23]. 131 responders (83 VNL, 48 placebo) in maintenance phase A entered phase B for a second year of maintenance (VNLresponders were re-randomized to VNLvs placebo; placebo responders stayed on placebo, fluoxetine responders stayed on fluoxetine).

In the first year of maintenance treatment in 258 responders, 23% of VNL-treated patients relapsed vs 42% with placebo. Thus 77% of the VNL group (n=83) stayed well for one year after already preselecting those who had stayed well for 6 months (n=258), selected after initially responding to treatment for an acute episode (n=715), as described in the previous paragraph. This is only 11.6% (83/715) of initial sustained responders.

Only 12.5% of placebo responders at one year relapsed at two years, but, in rerandomized VNL responders (another super-enrichment on top of all the prior enriched selection phases), 44.8% of the placebo group relapsed at 2 years vs 8.0% with VNL. Or, as the pharmaceutical industry marketing emphasized, 92% of venlafaxine patients remained well at 2 years follow-up. This 92% is seems like a huge number. But, because of super-enrichment, it represents the repeated selection of a tiny group of highly VNL-responsive patients: It is 92% of the 11.6% above (those who responded at one year), which is *10.7% of original sustained responders.* Once dropouts are included, patients still treated at two years, after the initial sample of over 1000 patients, was 15 subjects with placebo and 31 subjects with VNL – 4.2% of the original sample.

Antidepressant discontinuation meta-analysis—The most recent review of the maintenance MDD literature represents a unique analysis [8]. The authors essentially conducted an enriched study of placebo response; in other words, they selected the data they would analyze based on a sample enriched for placebo responders, and biased against drug response. Then they concluded that drugs were ineffective and even harmful.

All they really proved – once again – is that the enriched maintenance design is biased against whatever one wants to bias it against.

This is the converse of the standard enriched design maintenance study, as described above, which is enriched for drug response and biased against placebo response. The same limitations apply in both cases: enrichment does not prove the inefficacy or harm of the treatment that is not being enriched, nor does it prove the efficacy or benefit of the treatment that is being enriched.

In this review, they collected 7 studies of maintenance treatment with antidepressants versus placebo in which initial acute treatment was provided with the two arms; in these 7 studies, the maintenance phase involved continuation of those patients who had responded to placebo acutely. In those acute placebo responders, relapse in the maintenance phase was (not surprisingly) uncommon (24.7%). In contrast, 39 trials involved acute treatment with antidepressant versus placebo, and the reviewers selected those patients who responded to antidepressants acutely, and then were randomized to receive placebo in maintenance treatment. In this group, which reflects antidepressants discontinuation after acute response, there was 42.1% relapse.

The authors interpret these results as showing harm with antidepressants, which they speculatively relate to animal data on monoaminergic effects of these agents. They conclude that the biological effects of antidepressants will actually increase the risk of relapse in long-term treatment, compared to no treatment (placebo).

This interpretation ignores the problems of the enriched design, and, as a result, this kind of analysis highlights the importance of always comparing treatment results to the natural history of an illness.

This analysis enriches the results for placebo response. The patients treated acutely who respond to placebo, stay on placebo; the patients treated acutely who respond to antidepressant, come off antidepressant. One should ask the question why these placebo acute responders responded to placebo? Did they actually respond to placebo, in some way that the inert pill, with its concomitant psychosocial warmth factors, had a direct effect producing response? Or is placebo a stand-in for natural recovery, spontaneous remission, as part of the natural history of recurrent, episodic depression?

The latter is a possibility, at least for part, if not all, of the placebo “response.” Over a century of *natural history* research, especially before the treatment era in past decades, has established the fact that there is an episodic course to recurrent unipolar depression, in which there are periods of acute symptomatology, and periods of natural remission [12–14]. During periods of natural remission, patients will stay well, often for years, without any treatment. The recovery of some patients with placebo, in those 7 studies, may well reflect natural cycling out of acute episodes in unipolar depression. Once patients have cycled out of acute episodes, they are in natural remission, which, in the case of recurrent unipolar depression, based on a century of research, usually involves over a year of remission before the next depressive episode[16]. In the 7 placebo maintenance response studies, no study exceeded 12 months of follow-up, and, in reading the appendix attached to the meta-analysis, it appears that the mean duration of follow-up was less than two months in six of the seven studies (range 1.4–1.9 months).

In other words, the lack of relapse really means that a patient improved spontaneously from acute depression in a two month study (the usual duration of acute depression studies), and then that patient remained well for another two months. This is not robust evidence of long-term stability on placebo. It represents the fact that when spontaneous remission occurs from acute depression, it lasts at least two months (and indeed usually up to a year), without any treatment.

In contrast, in the antidepressant discontinuation studies analyzed, all patients responded in acute treatment (usually two months in duration), and then 42% relapsed in maintenance treatment after the antidepressant was discontinued. One might question whether serotonin withdrawal syndrome, which can mimic depressive episodes, occurred in some cases. But separate from that issue, a century of natural history research has led to a clear consensus that the mean duration of a typical depressive episode in unipolar depression is 6–12 months [12–14]. If a patient is treated to recovery at two months, and then the treatment is stopped, it is proven that such a patient will relapse into the mood episode rapidly, because the 6–12 month period of the biological persistence of a mood episode has not yet elapsed. This has proven with antidepressants in depression, and with neuroleptics in mania, repeatedly [19].

In sum, this creative analysis of the maintenance MDD literature suffers from a complete lack of awareness of the impact of the enriched design; the analysis is enriched for placebo response, and thus biased against antidepressant effect. The most conceptually parsimonious, and empirically well-supported, interpretation, based on extensive clinical literature in human beings (as opposed to speculative biological extrapolations from animal studies), would be to view these results as mainly reflective of the natural history of depression - not specific harm from antidepressants nor special benefit from placebo.

Maintenance data in STAR*D—Though STAR*D is mainly reported in terms of its acute data, one analysis so far also provides maintenance data [1], and it is perhaps underappreciated that the STAR*D maintenance data may be the best evidence we have to date on long-term efficacy with antidepressants in unipolar depression. Further, STAR*D was designed to be, and is, generalizable to the real world of complex, comorbid, recurrently depressed patients, as opposed to the cleaner populations studied in most RCTs (designed for FDA registration by the pharmaceutical industry).

As noted previously, STAR*D is a double-blind randomized study; all the following maintenance data after the first phase of treatment (i.e., with the dozen or so antidepressant treatments given besides citalopram) involve randomized, not observational, data.

The basic results are as follows: Of subjects who acutely responded or remitted to antidepressants in STAR*D, only about one-half stayed well at one year (sustained remission). In other words, preselecting those patients who have acute benefit with antidepressants, as noted above, one-half will maintain benefit. Since one-half get acute benefit, and one-half of that group have sustained maintenance benefit, only one-quarter of the overall sample has long-term maintenance remission with antidepressants in unipolar depression [20]. Based on STAR*D, there appears to be much less long-term benefit with antidepressants in unipolar depression than has often been assumed.

Objections to our critique of enriched maintenance designs—The above critique of enriched maintenance designs is neither widely known nor generally accepted. It is novel, rarely stated, and, when stated, strongly opposed by most researchers involved with maintenance studies in psychopharmacology.

There has not been much published discussion of this topic, but one objection that could be raised is that the enriched design is not biased because those who respond acutely to a drug treatment are both “true drug responders” and “placebo drug responders” – meaning, some of them would have responded to placebo had they been given placebo. Thus the design is not solely biased towards the study drug. This objection would only make sense if all patients were equally likely to respond to drugs or placebo; if 50% of patients “really” responded to drug (true drug response) and 50% would have responded to placebo had it been given (placebo drug response), then a maintenance randomization of those acutely responsive subjects to drug versus placebo would be valid. Ironically, this would only be the case if the critiques of Kirsch and colleagues [2] is correct, i.e., if antidepressants are not more effective than placebo for acute depression.

If antidepressants *are* more effective than placebo for acute depression in most patients, as we believe we showed earlier in this article, then the percentage of true drug responders should be higher than the percentage of those who would have responded to placebo anyway (placebo drug responders). In a hypothetical group of acutely depressed patients treated with antidepressant X, and later randomized to a maintenance study of X versus placebo, the reality is that there would not have been a 50–50 split between true drug responders and placebo drug responders before maintenance randomization; the split would be 60–40, or 70–30, or even higher in favor of drug X.

In other words, since antidepressants are better than placebo acutely, enrichment for acute efficacy before maintenance RCTs *is indeed biased* in favor of antidepressants as opposed to later treatment with placebo. Enrichment entails bias.

Interestingly, many psychiatric researchers appear to fully understand this critique as applied to the maintenance meta-analysis by Andrews and colleagues [8]; they appreciate that such an analysis entails “apples and oranges”, picking out placebo responders and comparing how they later did when continued on placebo, versus drug responders and how they later did when switched to placebo. Placebo responders are just different than drug responders, it is said. We agree. All placebo responders, by definition, respond to placebo, while only some probably would respond to drug. Thus such analyses are biased in favor of placebo response.

But while this enriched method – a species of selection bias that is unique to maintenance clinical trial design -[10] is rejected by many in our field in relation to the claim that placebos are as good or better than antidepressants, the exact same method is used to assert that antidepressants are more effective than placebo. The reason for such selectivity about accepting or rejecting the same research methodology is not entirely pellucid.

Conclusions

Numerous reviews and meta-analyses of the antidepressant literature in MDD, both acute and maintenance, appear to wish to make larger claims than their research methods allow. Specifically, based on the available FDA database analyses, it is false to claim that antidepressants are, in a general sense, ineffective in acute depressive episodes. The claim that they lack such benefits is disproven by standard valid methods of pooling effect size differences and by using appropriate meta-analytic models. Correction of those effect size difference for a floor effect, so that relative (instead of absolute) effect size differences are calculated, also shows that antidepressant benefit is seen not only in severe depression, but also in moderate depression. These analyses confirm lack of benefit of antidepressants over placebo in mild depression.

One can turn the attention-getting conclusions of the review by Kirsch and associates around: Instead of concluding that antidepressants are ineffective acutely except for the most extreme depressive episodes, correction for the statistical floor effect proves that antidepressants are effective acutely except for the mildest depressive episodes.

The claim that antidepressants are completely ineffective, or even harmful, in maintenance treatment studies involves unawareness of the enriched design effect, which, in that analysis, was used to ensure placebo efficacy. The same problem exists for the standard interpretation of those studies, though: they do not prove antidepressant efficacy either, since they are biased in favor of antidepressants.

In sum, in an objectively and statistically valid an assessment as possible, we would conclude that antidepressants are effective for acute depressive episodes that are moderate to severe, but not in mild depression. For maintenance efficacy, the research designs used have been biases in their favor, and it would seem more objective to conclude that antidepressant long-term efficacy is not proven, but neither do those data support the conclusion that they are harmful.

Acknowledgments

This work was partly supported by grant 5R01MH078060 from the National Institute of Mental Health (SNG), and a scholarship from the National Commission for Scientific and Technological Research (CONICYT) of the government of Chile (PAV).

REFERENCES

1. Rush AJ, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry. 2006; 163(11):1905–17. [PubMed: 17074942]
2. Kirsch I, et al. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. PLoS Med. 2008; 5(2):e45. [PubMed: 18303940]
3. Ioannidis JP. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? Philosophy, Ethics, and Humanities in Medicine. 2008; 3(14)
4. Geddes JR, et al. SSRIs versus other antidepressants for depressive disorder. Cochrane Database Syst Rev. 2000; (2):CD001851. [PubMed: 10796826]
5. Kornstein SG, K.J. Ahmed S, Thase M, Friedman ES, Dunlop BW, Yan B, Pedersen R, Ninan PT, Li T, Keller M. Assessing the efficacy of 2 years of maintenance treatment with venlafaxine extended release 75–225 mg/day in patients with recurrent major depression: a secondary analysis of data from the PREVENT study. International clinical psychopharmacology. 2008; 23(6):357–363. [PubMed: 18854724]
6. Andrews PW, K.S. Halberstadt LJ, Gardner CO, Neale MC. Blue again: perturbational effects of antidepressants suggest monoaminergic homeostasis in major depression. Front Psychol. 2011; 2(159):1–24. [PubMed: 21713130]
7. Turner EH, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008; 358(3):252–60. [PubMed: 18199864]
8. Horder J, M.P. Waldmann R. Placebo, Prozac and PLoS: significant lessons for psychopharmacology. J Psychopharmacol. 2010 In press.
9. First, MB.; Blacker, D. Handbook of psychiatric measures. Rush, AJ., editor. American Psychiatric Press; Washington D. C.: 2000.
10. Schmitt AB, et al. Differential effects of venlafaxine in the treatment of major depressive disorder according to baseline severity. Eur Arch Psychiatry Clin Neurosci. 2009; 259(6):329–39. [PubMed: 19255709]
11. Montgomery SA, Lecrubier Y. Is severe depression a separate indication? ECNP Consensus Meeting September 20, 1996, Amsterdam. European College of Neuropsychopharmacology. Eur Neuropsychopharmacol. 1999; 9(3):259–64. [PubMed: 10208298]

12. Feinberg M, et al. The Carroll rating scale for depression. III. Comparison with other rating instruments. Br J Psychiatry. 1981; 138:205–9. [PubMed: 7272611]

13. Fountoulakis KN, Möller HJ. Antidepressant drugs and the response in the placebo group: the real problem lies in our understanding of the issue. Journal of Psychopharmacology. 2011 In press.

14. Davis JM, G.W. Qu J, Prasad P, Leucht S. Should we treat depression with drugs or psychological interventions? A reply to Ioannidis. Philos Ethics Humanit Med. 2011; 6(8)

15. Trivedi MH, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry. 2006; 163(1):28–40. [PubMed: 16390886]

16. Kraepelin, E. Manic-Depressive Insanity and Paranoia. Robertson, GM., editor. E & S Livingstone; Edinburgh: 1921.

17. Frank E, et al. Three-year outcomes for maintenance therapies in recurrent depression. Arch Gen Psychiatry. 1990; 47:1093–1099. [PubMed: 2244793]

18. Goodwin, F.; Jamison, K. Manic Depressive Illness. 2nd edition. Oxford University Press; New York: 2007.

19. Goodwin FK, Whitham EA, Ghaemi SN. Maintenance Treatment Study Designs in Bipolar Disorder: Do They Demonstrate that Atypical Neuroleptics (Antipsychotics) are Mood Stabilizers? CNS Drugs. 2011; 25(10):819–827. [PubMed: 21936585]

20. Suppes T, et al. Risk of reoccurrence following discontinuation of lithium treatment in bipolar disorder. Arch Gen Psychiatry. 1991; 48:1082–8. [PubMed: 1845226]

21. Cavanagh J, Smyth R, Goodwin GM. Relapse into mania or depression following lithium discontinuation: a 7-year follow-up. Acta Psychiatr Scand. 2004; 109(2):91–5. [PubMed: 14725588]

22. Briscoe, BE-MR. The evidence for the long-term use of antidepressants as prophylaxis against future depressive episodes. American Psychiatric Association Annual Meeting Oral presentation; 2010.

23. Kocsis JH, T.M. Trivedi MH, Shelton RC, Kornstein SG, Nemeroff CB, Friedman ES, Gelenberg AJ, Dunner DL, Hirschfeld RM, Rothschild AJ, Ferguson JM, Schatzberg AF, Zajecka JM, Pedersen RD, Yan B, Ahmed S, Musgnung J, Ninan PT, Keller MB. Prevention of recurrent episodes of depression with venlafaxine ER in a 1-year maintenance phase from the PREVENT Study. J Clin Psychiatry. 2007; 68(7):1014–23. [PubMed: 17685736]

Table 1

Summary of analysis of reviews of antidepressant efficacy in RCTs of MDD

Study	n	Trials reviewed	Effect sizes [95% CI]	Comments
Rush et al 2006 STAR*D RCT	3671	1	About 60% acute remission, About 30% maintenance remission	No pbo group. Good acute efficacy is shown, but maintenance efficacy is about one-half less than acute efficacy.
Kocsis et al. 2007 and Kornstein et al 2008 Maintenance RCT of Venlafaxine vs pbo	1stmaintenance study (year 0) n=1096 2nd maintenance study (year 1) n=114	2	92% 2 year efficacy reported; in fact, this reflects 11% of original sample	“Double enrichment” design. 2nd maintenance study sample was only about 10% of the initial sample
Turner et al. 2008 MA of FDA database of RCT	12564	74	0.37 [0.33,0.41] for published studies vs 0.15 [0.08, 0.22] for unpublished studies. ES was 0.31[0.27, 0.35] when all studies are combined.	31% of studies were unpublished, accounting for 27.5% of the sample
Kirsch et al. 2008 MA of FDA database of RCT	5133	35	Overall standardized effect size 0.61. Absolute ES HDRS 9.6 drug, and 7.8 pbo.	NICE criterion for clinical significance was absolute ES of 3 HDRS points or standardized ES of d=0.5 for AD-Pbo difference. Overall nonstandardized effect size of 0.32 increases to 0.40 when corrected for baseline severity (authors do not discuss)
Horder et al 2010 Reanalysis of Kirsch et al MA	5133	35	Absolute HDRS difference between AD-pbo = 2.70 (including negative unpublished studies)	Reanalysis was based on a) random effects rather than fixed effects model as in Kirsch and colleagues, & b) pooling ES differences study by study, rather than summing all studies and then ES difference: These changes produce much larger effect size near the NICE threshold
Davis et al. 2010 Narrative summary of MAs and RCTs s	Not reported	Not reported	Mean acute difference AD-pbo =23.6% Mean Maintenance difference AD-pbo= −36%	Uncritical about bias toward ADs in maintenance studies using the enriched design
Fountalakis and Moller 2011 Reanalysis of Kirsch et al MA	5133	35	Mean AD ES was 10.05, not 9.60 as in Kirsch and colleagues. AD-pbo difference was 2.18, not 1.80 as in Kirsch and associates. Venlafaxine and paroxetine absolute HDRS ES were 3.12&3.22, respectively, exceeding NICE threshold, but not nefazodone or fluoxetine do	Reanalysis was based on weighting the mean difference by sample size.
Andrews et al. 2011 MA of some AD maintenance RCT	3,454	46 trials	Risk Difference AD-pbo for relapse −0.20, meaning 20% increased rate of relapse with AD than with placebo.	MA used an “enriched” design in favor of the pbo arm.
Briscoe and El-Mallakh 2011	449	5	5 RCTs examined for AD efficacy after 6 months. 4/5 studies showed no benefit with AD over pbo.	Only analysis to correct for enriched design, which is biased in favor of ADs. Removes relapses due to AD withdrawal.
Ghaemi and Vöhringer 2011 Reanalysis of Kirsch et al MA to correct for statistical floor effect	5133	35	Relative effect size for mild depression was 5% (HDRS <24), 12% for moderate (24<HDRS>28), and 16% for severe depression (HDRS >28}.	NICE criterion is met by 11.5% relative difference between AD-pbo. This analysis disproves the claim by Kirsch et al that only severe depression has clinically meaningful ES. Moderate depression also met NICE criterion.

RCT=Randomized Controled Trial, FDA=Food and Drug Administration (USA), HDRS=Hamilton Depression Rating Scale, NICE=National Institute for Health and Clinical Excellence (UK), AD=Antidepressant, Pbo=Placebo, CI=Confidence Interval, ES=Effect Size,

Table 2

Relative effect size difference (drug/placebo) by depression severity in Kirsch and colleagues' meta-analysis (n trials= 35)

	Drug			Placebo			
Depression Severity*	Mean baseline HDRS score	Mean final change in HDRS score	Relative effect size measure (%)**	Mean Baseline HDRS score	Mean Final change in HDRS score	Relative effect size measure (%)	Relative effect size difference% (drug-placebo)
Mild (23%)	22.6	8.8	39	23.2	8	34	5
Moderate (54%)	25.6	10.5	41	25.4	7.4	29	12
Severe (23%)	28.75	12.0	42	28.2	7.2	26	16

Moderate= At least one arms is rated > 24< 28 on HDRS

Severe= At least one arm (drug or placebo) is rated ≥ 28 on HDRS

*Mild= At least one arm (drug or placebo) is rated ≤24 on HDRS

**Relative effect size = absolute mean HDRS change/mean baseline HDRS score