Review

Placebo, Prozac and PLoS: significant lessons for psychopharmacology

Jamie Horder[1], Paul Matthews[2] and Robert Waldmann[3]

Journal of Psychopharmacology
25(10) 1277–1288

DOI: 10.1177/0269881110372544
jop.sagepub.com

Abstract

Kirsch et al. (2008, Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 5: e45), conducted a meta-analysis of data from 35 placebo controlled trials of four newer antidepressants. They concluded that while these drugs are statistically significantly superior to placebo in acute depression, the benefits are unlikely to be clinically significant. This paper has attracted much attention and debate in both academic journals and the popular media. In this critique, we argue that Kirsch et al.'s is a flawed analysis which relies upon unusual statistical techniques biased against antidepressants. We present results showing that re-analysing the same data using more appropriate methods leads to substantially different conclusions. However, we also believe that psychopharmacology has lessons to learn from the Kirsch et al. paper. We discuss issues surrounding the interpretation of clinical trials of antidepressants, including the difficulties of extrapolating from randomized controlled trials to the clinic, and the question of failed trials. We call for more research to establish the effectiveness of antidepressants in clinically relevant populations under naturalistic conditions, for example, in relapse prevention, in patients with co-morbidities, and in primary care settings.

Keywords

Antidepressants, meta-analysis, placebo

Introduction

On 26 February 2008, *PLoS Medicine* published a paper by Kirsch et al. entitled 'Initial severity and antidepressant benefits' (Kirsch et al., 2008). This paper presented a meta-analysis of the data held by the US Food and Drug Administration (FDA) from 35 randomized placebo controlled trials of four newer antidepressants in the acute treatment of major depression.

Kirsch et al. concluded that, while antidepressants are statistically superior to placebo, the magnitude of the drug–placebo difference is small, being on average just 1.8 points on the Hamilton depression scale (HAMD). They also reported that the drug–placebo difference is correlated with the pre-treatment severity of depression.

Adopting UK National Institute for Clinical Excellence (NICE) criteria for 'clinical significance', namely, that a treatment effect is clinically significant if the benefit over placebo is at least three HAMD points or a Cohen's *d* effect size of 0.5 (National Institute for Clinical Excellence, 2004), they concluded that the effect of antidepressants 'reaches conventional criteria for clinical significance only for patients at the upper end of the very severely depressed category' and offered the recommendation that 'there seems little evidence to support the prescription of antidepressant medication to any but the most severely depressed patients, unless alternative treatments have failed to provide benefit'.

This paper attracted much comment, including radio and front-page newspaper coverage. Kirsch et al. (2008) was still ranked amongst the most read items on the *PLoS Medicine* website as of March 2010, and has been accessed over 200,000 times to date. Several academic responses have since been published (Kelly, 2008; Khan and Khan, 2008; McAllister-Williams, 2008a, 2008b; Moller, 2008; Nutt and Malizia, 2008; Parker, 2009; Turner and Rosenthal, 2008).

The lead author of the 2008 paper, Irving Kirsch, has previously published two other meta-analyses of antidepressant effects. The first appeared in 1998 and included 19 methodologically diverse studies of a variety of compounds, some of which are not in clinical use as antidepressants. It concluded that 'the inactive placebos produced improvement that was 75% of the effect of the active drug' (Kirsch and Sapirstein, 1998).

The second meta-analysis, in 2002 (Kirsch et al., 2002), utilized the FDA database of results submitted by pharmaceutical companies, and reported that for fluoxetine,

[1]Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, UK.
[2]Department of General Medicine, Milton Keynes Hospital, Milton Keynes, UK.
[3]Facoltà di Economia, Università di Roma 'Tor Vergata', Rome, Italy.

Corresponding author:
Jamie Horder, Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, Oxfordshire OX3 7JX, UK
Email: jamie.horder@googlemail.com

nefazodone and venlafaxine, 'Approximately 80% of the response to medication was duplicated in placebo control groups, and the mean difference between drug and placebo was approximately 2 points [on the Hamilton Depression Scale]'. The 2008 paper utilized largely the same dataset as in the 2002 analysis, although it also includes data on paroxetine, data on four paroxetine trials having become available in the intervening time.

The FDA requires manufacturers to submit all available clinical trial data regarding novel drugs prior to granting licensing approval, whether or not this data is published. Kirsch et al. therefore correctly noted that their analysis was free of publication bias, a well documented problem in the antidepressant literature (Melander et al., 2003) and elsewhere in medicine.

In this review, we will argue that Kirsch et al. (2008) is a seriously flawed analysis which draws misleading conclusions on the basis of unusual and potentially biased statistical techniques. However, we also believe that psychopharmacology has something to learn from this paper, since, like depression, ostensibly learned critiques on antidepressants have a tendency to recur every few years.

Statistical issues

Commentators on Kirsch et al. have largely accepted its statistical soundness. For instance, the otherwise critical McAllister-Williams wrote: 'The meta-analysis described in this paper has been robustly carried out using standard methodology... Undoubtedly the findings in this analysis are robust, as far as the studies included in the analysis are concerned' (McAllister-Williams, 2008a). This seems to us a generous evaluation.

In this section we examine, and criticize, the statistical methods used in Kirsch et al. (2008). For further details on the issues raised in this section, readers should consult the Statistical Appendix to this paper.

Which 'drug–placebo difference'?

Kirsch et al. analysed the results of 35 clinical trials, each comparing patients randomly assigned to receive an antidepressant (venlafaxine, nefazodone, paroxetine, and fluoxetine) or a pill placebo. The outcome data consists of a mean Hamilton depression rating scale score (HAMD or HRSD) before treatment, and a mean score after treatment, for each group, in each trial. Hence there are $35 \times 2 = 70$ before–after pairs of HAMD scores and, correspondingly, 70 pre–post change scores.

For any given trial, the benefit of antidepressants over placebo is the difference in the change (improvement) scores between the two groups. Determining the mean benefit of antidepressants would therefore seem to be a simple matter of calculating the appropriately weighted mean of the between-group differences in the 35 trials. This mean is an estimate of average benefit an individual patient can expect to experience if they are given an antidepressant, relative to a patient from the *same population* randomized to receive a placebo, which is a number of clear clinical relevance.

However, Kirsch et al. did not perform such an analysis. Instead, they pooled the data for all of the antidepressant-treated subjects across the 35 trials, and likewise pooled the data for all placebo-treated subjects. They then averaged the improvement scores seen in the placebo subjects and, separately, the improvement seen in the drug subjects. Finally, they compared the two averages. This is the basis of their famous result that '... weighted mean improvement was 9.60 points on the HRSD in the drug groups and 7.80 in the placebo groups, yielding a mean drug–placebo difference of 1.80', and this was also the basis for the two main graphs in the 2008 paper.

This is a somewhat curious approach. Standard meta-analytic practice is to consider the difference in improvement between the drug group and the placebo group *within each trial*, and only then to average these difference values across trials. Re-analysing the data in this appropriate way, we find an overall weighted mean difference between the two groups of 2.70 (95% confidence interval (CI) 1.95, 3.44) HAMD points; rather higher than Kirsch et al.'s 1.80. Why the discrepancy?

Firstly, the decision by Kirsch et al. to calculate the mean drug and mean placebo improvement scores and then take the difference between them introduces bias. This approach effectively treats each trial as being two entirely separate experiments, one measuring improvement on placebo, the other measuring improvement on drug. It also assumes that the placebo improvement is the same across all trials, yet it is because the placebo response in any given trial cannot be determined beforehand that randomized controlled trials (RCTs) are conducted.

Kirsch et al.'s method is effectively blind to an important element of the information provided by the 35 trials, namely the pairing between the drug group and the placebo group in each trial. Although this approach was used previously by Kirsch and Sapirstein in their original 1998 antidepressant meta-analysis, they at least acknowledged its novel nature in that paper (Kirsch and Sapirstein, 1998). In the 2008 paper, Kirsch et al. did not discuss or defend their choice of method.

Secondly, Kirsch et al.'s use of fixed-effect precision-weighting to weight each trial also introduces bias. A fundamental assumption of fixed-effects analysis is that the magnitude of the effect being estimated is identical across trials. In fact, the magnitude of the antidepressant effect could vary between trials for numerous reasons, e.g. due to the use of different antidepressants or different patient populations. Indeed, one of Kirsch et al.'s claims is that the effects of antidepressants vary with differing average initial severity,

Therefore, it is appropriate to use random-effects weighting, which does not assume that effects are constant across trials, in this case. Random-effects weighting places more weight on smaller studies, which is a weakness whenever publication bias leads to over-representation of small, positive studies. However, we do not believe this is the case here, since this data-set is known to be free of publication bias as it is based on results submitted to regulatory authorities, including unpublished data.

Indeed, it could be said that Kirsch et al.'s methodology is in fact more biased than the publication bias that

they avoided. As stated above, our estimated mean drug–placebo difference, using a precision weighted random effects model, including unpublished data, is 2.70 HAMD points. Excluding those trials which were not published in full gives a result of 3.37 points. Thus, publication bias would have inflated the apparent effect by around 0.67 points. However, Kirsch et al.'s methodology produced a figure of 1.80, thus underestimating the effect by 0.90 points.

The non-standard use of standardized scores

In addition to the analysis using HAMD scores, discussed above, Kirsch et al. also present an analysis using a standardized measure of effect size, Cohen's *d*. Cohen's *d* is the mean of the difference between two groups of scores, divided by the standard deviation of that difference (Cohen, 1988). In this case, *d* was calculated, for drug and placebo groups separately, using the change in HAMD scores from the beginning to the end of the trial. Kirsch et al.'s overall estimate of the drug *vs.* placebo difference was $d = 0.32$.

In meta-analysis, converting results to *d* scores is useful when the trials of interest did not all share a common outcome measure, as it allows the relative magnitude of treatment effects to be compared, because *d* is a unitless number. In this case, however, all of the trials used the same scale to rate depression, and it is therefore appropriate to use the HAMD scores only, as the *Cochrane handbook* of meta-analysis states (The Cochrane Collaboration, 2008: Section 9.2.3). One advantage to doing so is that the mean HAMD difference is directly interpretable, while the standardized mean difference (*d* score) is a statistical construct.

Kirsch et al. did perform an analysis using raw HAMD scores, in addition to the one using *d* scores, but they do not mention that some of their most interesting results are *only* apparent if the *d* scores are used.

The most notable example is the statement that:

> '... Although differences in improvement increased at higher levels of initial depression, there was a negative relation between severity and the placebo response, whereas there was no difference between those with relatively low and relatively high initial depression in their response to drug. Thus, the increased benefit for extremely depressed patients seems attributable to a decrease in responsiveness to placebo, rather than an increase in responsiveness to medication.'

(See Kirsch et al.'s Figure 3, reproduced below as Figure 1.)

In one form or another, this claim appears three times in the Discussion section, as well as in the Abstract, and in the Editor's Summary of the paper. But it only holds true when the (inappropriate) standardized mean difference (d) analysis is used (Figure 1). If the raw mean difference score is used instead, this effect vanishes: while the finding of increasing drug–placebo difference with higher baseline severity remains, this is seen to be driven by increasing improvement in the drug group, and essentially constant improvement in the placebo group (Figure 2).

Using raw HAMD scores is more usual, and there is no reason to prefer the standardized mean difference (d) in this case. Previous meta-analyses have also found higher

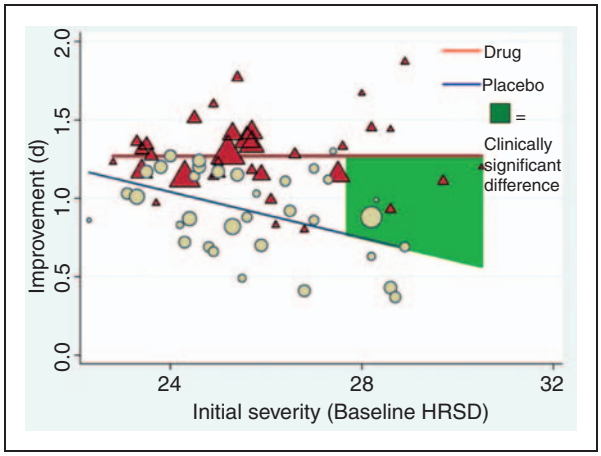

Figure 1. Figure 3 from Kirsch et al. (2008) showing the standardized mean difference (*d*) score for improvement on drug and placebo in relation to baseline HAMD score.

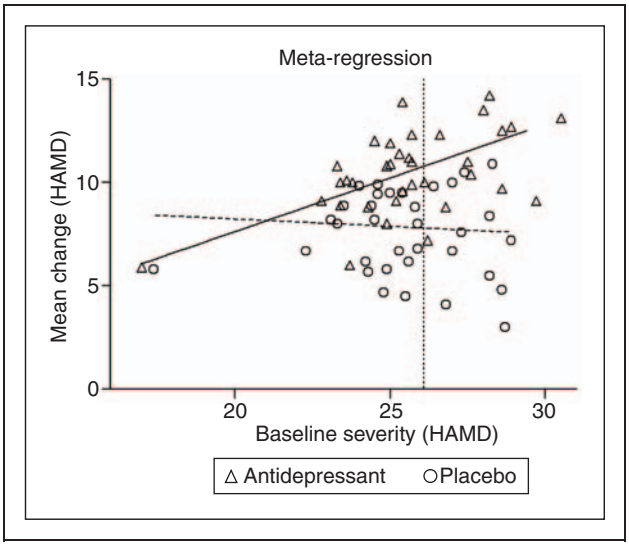

Figure 2. Trials from Kirsch et al. (2008) using mean HAMD difference rather than standardized effect size (*d*). Antidepressant and placebo groups for each trial are plotted separately, as in the original. A vertical broken line marks the threshold for a three-point change score difference between the regression lines for antidepressant and placebo groups. These fixed effects regression lines (weighted by sample size) produce a threshold around a baseline HAMD score of 26 points (precision weighted random and fixed effects regression lines are very similar). The placebo regression line slope is not statistically significantly different from zero; in other words, there is no decline in the magnitude of the placebo response in trials with higher mean baseline severity values. Note that the remarks in the section *Which 'drug–placebo difference'?* of this paper also apply to this analysis.

antidepressant improvements with increasing severity: see, for example, Khan et al. (2002), although unlike Kirsch et al. these authors rightly acknowledged that their data 'cannot be directly applied to clinical practice. Research

participants meet stringent exclusion and inclusion criteria and are not representative of the general population of patients with depression.'

Kirsch et al. also state that the NICE threshold for clinical significance, $d > 0.5$, is reached only with a baseline HAMD of about 28. However, the other criterion specified by NICE, a mean difference of greater than 3 HAMD points, is reached at a lower baseline severity of approximately 26, as can be seen in Kirsch et al.'s Figure 4. Both criteria are equally arbitrary, but it is curious that Kirsch et al. quoted only the one which led to conclusions less favourable to antidepressants, and although the difference between 28 and 26 is perhaps modest, it brings the 'clinical significance' threshold close to the average baseline severity of the studies, since the median weighted mean baseline score is 25.5.

Rating depression and rating antidepressants

We have criticized Kirsch et al.'s (2008) analysis for using non-standard, inappropriate and biased methods. However, even adopting standard meta-analytic statistics, we find an overall weighted average of just 2.7 points, on a scale with a maximum score of 52 (see Table 1). Although this might appear small, interpretation of the magnitude of the effect depends upon the assumptions we make about the nature of antidepressant clinical trials.

Better drugs or worse placebos?

Kirsch et al. contend that for the four antidepressants they consider, 'efficacy reaches clinical significance only in trials involving the most extremely depressed patients, and that this pattern is due to a decrease in the response to placebo rather than an increase in the response to medication'.

There are several problems with this statement. Firstly, as noted above, this finding is not robust because it is only apparent when standardized effect scores are used. Secondly, it makes no sense to say that the increasing efficacy of antidepressants in more severe depression is not due to an increase in the response to medication, but is on the contrary due to decreasing response to placebo. The pharmacological effects of any medication are defined as (improvement with medication – improvement with placebo).

Unfortunately, many commentators made misleading statements to the effect that even in severe depression, antidepressants were somehow not 'really' effective at all, because the apparent effect was driven by worse placebo effects. Not even the Editor's Summary in *PLoS Medicine* avoided such confusion: 'Additional analyses indicated that the apparent clinical effectiveness of the antidepressants among these most severely depressed patients reflected a decreased responsiveness to placebo rather than an increased responsiveness to antidepressants.' Nothing in the data suggests that the clinical effectiveness is merely 'apparent'.

What is 'very severe'?

Kirsch et al. state that:

> 'the differences between drug and placebo were not clinically significant in clinical trials involving either moderately or very severely depressed patients, but did reach the criterion for trials involving patients whose mean initial depression scores were at the upper end of the very severe depression category (mean HRSD baseline approx. 28).'

This is consistent with the terminology proposed by the American Psychiatric Association (American Psychiatric Association, 2000) and adopted by NICE (National Institute for Clinical Excellence, 2004): 'Moderate' 14–18, 'Severe' 19–22, and 'Very Severe' 23+. However, this scheme is arbitrary, and many commentators have argued that it overstates the severity of any given point on the Hamilton scale, e.g. Moncrieff and Kirsch (2005), who said that '"severe" depression [by these APA criteria, i.e. HAMD 19–22]... would generally be referred to as moderately depressed'.

Moreover, Kirsch et al. neither looked for, nor found, a correlation between depression severity and response to antidepressants at the level of individual *patients*. Rather, they found a correlation between mean depression severity amongst recruits to *trials* and reported effect size within such trials. It is entirely plausible that trials recruiting more severe patients tend to have other characteristics which lead to their finding larger drug–placebo differences. Furthermore, the included trials only comprise a relatively narrow range of HAMD scores, the great majority being between 23 and 28, i.e. within the 'very severe' range, as defined by the APA.

Therefore, Kirsch et al.'s claim that 'there seems little evidence to support the prescription of antidepressant medication to any but the most severely depressed patients, unless alternative treatments have failed to provide benefit', is a double extrapolation. It extrapolates from data on trials to data on patients, and it extrapolates from a regression line based on data from 'very severely' depressed patients to other patients. Thus, while this statement may be strictly accurate in

Table 1. A table summarizing the estimates of the overall effect of antidepressants over placebo produced using various statistical methods

Method	Estimated mean drug vs. placebo effect (HAMD)	95% confidence intervals	Notes
Kirsch et al. (2008) method	1.80	n/a	
Fixed-effect meta-analysis, HAMD change	2.40	1.92 to 2.88	
Random-effects meta-analysis, HAMD change	2.70	1.95 to 3.44	This is the standard meta-analytic approach

terms of the data-set analysed in Kirsch et al. (2008), it is equally true that there is little evidence in this data-set that antidepressants are *ineffective* in mild depression, because only one of the trials included was about patients with mild depression. Other results from beyond the FDA approval dataset are more relevant to this issue (Lima and Moncrieff, 2000).

Are all drugs equal?

Kirsch et al. pooled data on four different antidepressants: the SSRIs paroxetine and fluoxetine, the 5HT antagonist and weak SSRI nefazodone, and the SNRI venlafaxine. Combining such diverse drugs together seems curious: venlafaxine is widely regarded as more effective than SSRIs, at least in some patients (Nemeroff et al., 2008), whereas nefazodone is no longer widely used having been withdrawn from the market in Europe and the USA.

Kirsch et al. conclude that their data show the four drugs to be equally effective. However, using our analysis we find overall effects *vs.* placebo for nefazodone, fluoxetine, paroxetine and venlafaxine to be 1.65, 2.06, 3.38 and 3.54 Hamilton points, respectively.

In other words, venlafaxine and paroxetine had larger effect sizes than fluoxetine or nefazodone – venlafaxine's being more than twice nefazodone's – and the overall effect sizes of these two drugs were larger than the three-HAMD-point clinical effectiveness threshold defined by NICE. Kirsch et al. say that the differences between drugs are accounted for by differences in the baseline severity of patients in the relevant trials, but regressing against baseline severity we find estimated effects of 2.52, 2.96, 3.24 and 3.96 at a baseline HAMD of 26.

These figures are not statistically significantly different from one another, so it is not possible to conclude from this data that (for example) venlafaxine is more effective than other antidepressants, but it is equally impossible to draw the opposite conclusion, especially since these trials were not designed, or powered, to address this. This is something that can only be resolved by head-to-head trials. For a recent synthesis of the results of such trials, finding some drugs to be significantly more effective than others, see Cipriani et al. (2009), although see also Ioannidis (2009).

What is being measured?

Perhaps the most overlooked aspect of the data included in Kirsch et al. (2008) is that it is not about the effects of antidepressants upon depression, but rather about the effects of antidepressants on the HAMD.

For research purposes, it is often necessary to attempt to quantify the severity of depression, but there is no such thing as a 'measure' of depression in the same way as a measure of weight or temperature. As any clinician and any patient with experience of the illness will be able to attest, estimating the severity of depression can be a difficult task in itself.

The HAMD has been the most popular rating scale for use in antidepressant trials for almost 50 years. However, it has long been argued that it is far from ideal in this role. For example, the authors of one recent review noted numerous problems with the HAMD including its multidimensional structure, the fact that it describes a construct of depression which corresponds poorly to that found in DSM-IV (e.g. it does not measure feelings of worthless or hopelessness, or anhedonia except indirectly), and that several items have poor reliability (Bagby et al., 2004).

Furthermore, writing in the *British Medical Journal*, Moncrieff and Kirsch observed of the Hamilton scale that it has:

'a maximum score of 52 and contains seven items concerning sleep and anxiety, with each item on sleep scoring up to 6 points. Hence any drug with some sedative properties, including many antidepressants, could produce a difference of 2 points or more without exerting any specific antidepressant effect' (Moncrieff and Kirsch, 2005).

Although wrong in point of fact (the HAMD contains three sleep items, for early, middle, and late insomnia, with a maximum score of 2 each, or 6 in total, not 6 each), the argument is a valid one (see also Kirsch, 2008). A change in total HAMD score does not necessarily imply a change in mood.

However, this argument remains valid if it is turned on its head: a *lack* of change, or a small change, in total rating scale score does not imply a lack of change, or a small change, in the severity of depressed mood. For example, were a certain drug to improve mood and reduce suicidality, but also cause insomnia and reduce appetite, it might have no effect on total HAMD score, or might even increase it.

These considerations are theoretical. Fortunately, there is empirical data to draw on. Ten years ago Faries et al. directly tested, and rejected, the idea that the ability of antidepressants to lower HAMD scores is driven by 'non-specific' effects such as improving sleep or appetite. They found that the average drug–placebo effect size across eight trials of fluoxetine in depression was $d = 0.45$ using the Maier and Philipp subscale, as opposed to $d = 0.37$ with the full scale HAMD. The Maier and Philipp subscale of the HAMD consists of six Hamilton items: mood, feelings of guilt, impairment to work and activities, observed retardation, observed agitation, and psychic anxiety (Maier and Philipp, 1985). Thus the effect size of fluoxetine was increased by 22% by considering only those items which rate core depressive symptoms (Faries et al., 2000).

Faries et al. also examined each individual Hamilton scale item, and found that, across several trials of tricyclic antidepressants, the six items which showed the largest drug–placebo effect sizes were: feelings of guilt, early insomnia, mood, suicide, impairment to work and activities, and psychic anxiety. All are core symptoms of clinical depression. A recent analysis of two large paroxetine trials (Santen et al., 2008) came to a very similar conclusion.

The best evidence is therefore that antidepressants improve depression rating scale scores by treating depression. However, the size of the measured effect of antidepressants against placebo is determined by the rating scale used (Bech, 2009). By implication, even more sensitive rating scales could provide even higher effect sizes. It might be complained that it is circular logic to favour a rating scale merely because it discriminates between antidepressants and placebos, and then to use it to claim that antidepressants work well! But while the original Hamilton scale was simply the invention of Max Hamilton, HAMD subscales such as the Maier

and Philipp, the Bech–Rafaelson Melancholia Scale, and others, were developed following quantitative analysis of HAMD data and are superior on psychometric measures (Moller, 2001). They could therefore be called rather more 'evidence based'.

How small is 'small', or: what is 80% of an emotion?

We believe that it is problematic to describe any given HAMD difference as 'small' or 'large' in terms of clinically meaningful effects. Although NICE have called a difference of less than three HAMD points clinically insignificant, this is an arbitrary criterion, as Moncrieff and Kirsch noted (Moncrieff and Kirsch, 2005). Although our point estimate of the overall effect size was just below this threshold (2.70), the 95% CIs cannot exclude an effect of this magnitude (95% CI 1.95–3.44).

We also question the validity of statements such as 'The response to placebo in these trials was exceptionally large, duplicating more than 80% of the improvement observed in the drug groups'. The placebo groups may have shown around 80% of the HAMD change seen in the drug groups, but only under a number of assumptions can this be taken to mean that they experienced 80% of the clinical improvement. For example, this requires that every point of decrease on the HAMD corresponds to a certain constant degree of clinical improvement.

This seems implausible, especially considering the heterogeneous nature of the Hamilton items – a decrease on the Suicide item is surely more clinically significant than a decrease on Middle Insomnia. In mathematical terms, Kirsch et al. are treating the HAMD as an *interval scale,* in which a certain numerical difference in score corresponds to a certain difference in the phenomenon being measured, but the HAMD is in fact an *ordinal scale,* in which a higher score corresponds to a greater severity, but not necessarily to any given degree of greater severity.

A more fundamental point is that Kirsch et al.'s statement assumes that the HAMD change recorded in the placebo group represents real improvement. Yet in fact it includes many other factors, such as demand effects (in this case, a tendency for patients to act in ways which conform to the experimenter's expectation of improvement (Nichols and Maner, 2008)), regression to the mean, the impact of initial rating score inflation, and other such artefacts.

The design of a randomized controlled trial ensures that these effects are equally present in both the drug and the placebo group. Therefore the *difference* between the two groups does represent real improvement, but the improvement seen in the placebo group cannot be assumed to be such. *A fortiori*, the improvement in the placebo group cannot be read as a 'placebo effect' that would be seen in actual clinical practice.

In summary, existing placebo controlled trials of antidepressants are neither designed to measure, nor be capable of measuring, the absolute magnitude of either placebo or drug effects. Rather, appropriately powered clinical trials are useful as a means of confirming, or not confirming, the antidepressant efficacy of a given treatment, or of showing one treatment to be superior or equivalent to another.

Lessons for psychopharmacology

Further remarks on antidepressant trials

In this article, we have argued that the Kirsch et al. (2008) analysis is seriously flawed. A number of other commentators, e.g. (Bech, 2009; Broich, 2009; Hegerl and Mergl, 2010; Khan and Khan, 2008; Mathew and Charney, 2009; McAllister-Williams, 2008a; Moller, 2008; Nutt and Malizia, 2008; Parker, 2009; Turner and Rosenthal, 2008) have also criticized the Kirsch et al. (2008) paper. The main arguments advanced by one or more of these commentators include:

- The observation that the conclusion that antidepressants are of no clinical significance in most patients does not accord with psychiatrists' clinical experience.
- Criticism of Kirsch et al.'s attribution of the increasing drug–placebo difference with higher baseline scores to declining placebo effects, rather than to increasing pharmacological effects (see *Better drugs or worse placebos?* above). Note that, unlike previous commentators, we argue that this attribution is not only confused but also *statistically* flawed (see *The non-standard use of standardized scores* above).
- The argument that Kirsch et al.'s analysis included only short-term efficacy trials (typically six weeks' duration, sometimes as short as four weeks) and that data on longer-term relapse prevention with antidepressants has found large positive effects (Geddes et al., 2003). There is no reason to think that all of the benefits of antidepressants manifest themselves by six weeks in clinical practice.
- Criticism of the statement that 'only the most severely depressed patients' benefit from antidepressants. Because only two of the included trials were in in-patients, and patients with suicidality or co-morbid mental health problems are routinely excluded from antidepressant trials for ethical reasons, the trials analysed by Kirsch et al. did not sample patients who would be considered severely depressed in a clinical sense.
- The argument that many modern antidepressant trials, including those included in the analysis, utilize problematic methods of recruitment. Commonly criticized practices include finding volunteers through community advertisement, rather than by clinical referral; excluding patients with suicidal thoughts or actions; and excluding patients with co-morbid substance abuse, anxiety or personality disorders. The result, it's argued, is that many of the patients included in modern trials are not representative of patients in the clinic and may not even be clinically depressed in the usual sense. The fact that mean baseline HAMD scores are relatively high (as high as was the case 20 years ago (Walsh et al., 2002)) is ascribed to the conscious or unconscious 'inflation' of baseline HAMD scores by raters eager to recruit participants.
- The argument that the high frequency of contact with researchers experienced by patients in clinical trials could have a real or apparent therapeutic effect, at least for the duration of the study, which would not be seen in the clinic. Also, many antidepressant trials allow the use of concomitant medications such as hypnotics, which will have a beneficial effect.

- The argument that when the outcomes of antidepressant trials are expressed as categorical figures, such as the percentage of patients meeting some defined criteria for remission, as opposed to continuous numbers (such as HAMD change scores), drug–placebo differences appear larger. However, it should noted that Kirsch and Moncrieff have criticized this apparent superiority of categorical outcome data as an 'illusion' (Kirsch and Moncrieff, 2007).

As those familiar with the literature will know, concern has long been growing over the fact that trials of antidepressants often 'fail'. Many points have been made in an attempt to explain this phenomenon and such arguments largely overlap with those listed above; for a discussion see Montgomery (1999).

In the light of such arguments, we believe that it is impossible to draw any conclusions about the clinical significance of the pharmacological effects of antidepressants from short-term efficacy trials. Although the results of such trials are of interest to authorities concerned with licensing drugs as safe and effective, including the FDA, the UK's MHRA and the EU's EMEA, they are probably insufficient for agencies concerned with assessing clinical and cost effectiveness, such as NICE.

We therefore argue that the conclusions of Kirsch et al. (2008) regarding the (lack of) clinical effectiveness of antidepressants are not supported, even setting aside the statistical concerns raised in the first part of this paper.

Trials – what are they good for?

Commentators writing in the *British Journal of Psychiatry* seven years ago, following the publication of a very similar meta-analysis (Kirsch et al., 2002), said, 'we simply do not know how big the effect of antidepressants is in clinical practice because RCTs are not designed to tell us this' (Parker et al., 2003). This remains true.

Some may find this conclusion a comforting one, compared with the view that antidepressants have a very small effect in practice, but it is clearly an unsatisfactory state of affairs. Doctors, patients and policy-makers need to be able to make evidence-based decisions about the utility of antidepressants, some of the most prescribed medications in the world. At present we do not believe that this is possible in many cases, leaving psychopharmacology to rely upon clinical experience, rather than evidence, to judge how well antidepressants work in practice, as opposed to *if* they work, although it should be noted that this is not a problem with antidepressants alone and that the state of the evidence on psychological treatments for depression has been deemed even worse (Nutt and Sharpe, 2008).

So long as the evidence base on antidepressants remains so limited in scope, it seems likely that challenges such as those of Kirsch et al. will continue, with negative effects on public attitudes towards these drugs. Appeals to clinical experience, while they may be valid, are increasingly greeted with scepticism in an era of evidence-based medicine. Improving the quality and the quantity of research on existing antidepressants could therefore be more important to psychiatry than the development of new drugs.

Echoing others (Parker, 2009), we therefore call for more research on the clinical effectiveness of antidepressants in naturalistic settings, for example, in relapse prevention, in patients with psychiatric co-morbidities, and in primary care. We also recommend the development and use of evidence-based alternatives to the Hamilton depression scale.

The true lesson of the present controversies may be not that antidepressants do not work very well, but that antidepressant research does not work very well. Practically, the implication is not that we should be prescribing antidepressants less, but rather we should be studying them more – and better. If Kirsch et al. (2008) serves to prompt such research, then this much-maligned paper may turn out to be a valuable contribution to modern psychiatry.

Acknowledgements

JH wrote the first draft of the paper. PM and RW conducted the statistical analyses. All authors contributed to the final draft. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

American Psychiatric Association (2000) *Handbook of psychiatric measures*, Washington, D.C.

Bagby RM, Ryder AG, Schuller DR and Marshall MB (2004) The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *Am J Psychiatry* 161: 2163–2177.

Bech P (2009) Is the antidepressive effect of second-generation antidepressants a myth? *Psychol Med* 40(2): 181–186.

Broich K (2009) Committee for Medicinal Products for Human Use (CHMP) assessment on efficacy of antidepressants. *Eur Neuropsychopharmacol* 19: 305–308.

Cipriani A, Furukawa TA, Salanti G, et al. (2009) Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 373(9665): 746–758.

Cohen J (1988) *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale, NJ: Lawrence Earlbaum Associates.

Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M and Potter WZ (2000) The responsiveness of the Hamilton Depression Rating Scale. *J Psychiatr Res* 34: 3–10.

Geddes JR, Carney SM, Davies C, et al. (2003) Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. *Lancet* 361: 653–661.

Hegerl U and Mergl R (2010) The clinical significance of antidepressant treatment effects cannot be derived from placebo-verum response differences. *J Psychopharmacol* 24(4): 445–448.

Ioannidis JP (2009) Ranking antidepressants. *Lancet* 373: 1759–1760; author reply 1761–1752.

Kelly BD (2008) Do new-generation antidepressants work? *Ir Med J* 101: 155.

Khan A and Khan S (2008) Placebo response in depression: a perspective for clinical practice. *Psychopharmacol Bull* 41: 91–98.

Khan A, Leventhal RM, Khan SR and 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.

Kirsch I (2008) Antidepressant drugs 'work', but they are not clinically effective. *Br J Hosp Med (Lond)* 69(6): 359.
Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ and Johnson BT (2008) Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 5: e45.
Kirsch I and Moncrieff J (2007) Clinical trials and the response rate illusion. *Contemp Clin Trials* 28: 348–351.
Kirsch I, Moore TJ, Scoboria A and Nicholls SS (2002) The Emperor's New Drugs: an analysis of antidepressant medication data submitted to the U.S. Food and Drug Administration. *Prev Treat* 5(1).
Kirsch I and Sapirstein G (1998) Listening to Prozac but Hearing Placebo: a Meta-Analysis of Antidepressant Medication. *Prev Treat* 1(2): Article 2a.
Lima MS and Moncrieff J (2000) Drugs versus placebo for dysthymia. *Cochrane Database Syst Rev* 4: CD001130.
Maier W and Philipp M (1985) Comparative analysis of observer depression scales. *Acta Psychiatr Scand* 72: 239–245.
Mathew SJ and Charney DS (2009) Publication bias and the efficacy of antidepressants. *Am J Psychiatry* 166: 140–145.
McAllister-Williams RH (2008a) Do antidepressants work? A commentary on 'Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration' by Kirsch et al. *Evid Based Ment Health* 11: 66–68.
McAllister-Williams RH (2008b) Misinterpretation of randomized trial evidence: Do antidepressants work? *Br J Hosp Med (Lond)* 69: 246–247.
Melander H, Ahlqvist-Rastad J, Meijer G and Beermann B (2003) Evidence b(i)ased medicine – selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *Br Med J* 326: 1171–1173.
Moller HJ (2001) Methodological aspects in the assessment of severity of depression by the Hamilton Depression Scale. *Eur Arch Psychiatry Clin Neurosci* 251(suppl 2): II13–20.
Moller HJ (2008) Isn't the efficacy of antidepressants clinically relevant? A critical comment on the results of the metaanalysis by Kirsch et al. 2008. *Eur Arch Psychiatry Clin Neurosci* 258: 451–455.
Moncrieff J and Kirsch I (2005) Efficacy of antidepressants in adults. *Br Med J* 331: 155–157.
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.
National Institute for Clinical Excellence (2004) *Depression: Management of depression in primary and secondary care*. London: Clinical practice guideline No 23.
Nemeroff CB, Entsuah R, Benattia I, Demitrack M, Sloan DM and Thase ME (2008) Comprehensive analysis of remission (COMPARE) with venlafaxine versus SSRIs. *Biol Psychiatry* 63: 424–434.
Nichols AL and Maner JK (2008) The good-subject effect: investigating participant demand characteristics. *J Gen Psychol* 135: 151–165.
Nutt DJ and Malizia A (2008) Why does the world have such a 'down' on antidepressants? *J Psychopharmacol* 22: 223–226.
Nutt DJ and Sharpe M (2008) Uncritical positive regard? Issues in the efficacy and safety of psychotherapy. *J Psychopharmacol* 22: 3–6.
Parker G (2009) Antidepressants on trial: how valid is the evidence? *Br J Psychiatry* 19: 1–3.
Parker G, Anderson IM and Haddad P (2003) Clinical trials of antidepressant medications are producing meaningless results. *Br J Psychiatry* 183: 102–104.
Santen G, Gomeni R, Danhof M and Pasqua OD (2008) Sensitivity of the individual items of the Hamilton depression rating scale to response and its consequences for the assessment of efficacy. *J Psychiatr Res* 42(12): 1000–1009.
The Cochrane Collaboration (2008) *The Cochrane handbook for systematic reviews of interventions. Version 5.0.0*. Higgins JPT and Green S (eds.) Available from www.cochrane-handbook.org.
Turner EH and Rosenthal R (2008) Efficacy of antidepressants. *Br Med J* 336: 516–517.
Walsh BT, Seidman SN, Sysko R and Gould M (2002) Placebo response in studies of major depression: variable, substantial, and growing. *JAMA* 287: 1840–1847.

Statistical Appendix

Introduction

Raw data can be found in Table SA1. The data consists of 35 RCTs, each with a drug and a placebo group. The data is derived from Table 1 in Kirsch et. al. (2008). For each group, the figures available consist of mean baseline and post-treatment HAMD scores, mean change in HAMD score, and the Cohen's effect size (d) of the change score.

The standard deviation of the change scores (SDc) are not explicitly presented in Kirsch et al.'s Table 1, but can be inferred since by definition d = change score/SDc.

Reproducing Kirsch et al.'s results

We believe that Kirsch et al. derived their results, most notably an overall drug–placebo difference of 1.8 HAMD points, by calculating the precision-weighted (i.e. weighting by $N/(SDc^2)$) mean of the change scores for all of the antidepressant-treated subjects and all of the placebo-treated subjects, and then calculating the difference between these two scores.

Table SA1 shows that we were able to derive the 1.8 figure (1.78) from the data by performing this analysis, thereby reproducing Kirsch et al.'s results (G46).

Critique

We believe that this approach is flawed for a number of reasons. See our paper for details. We also argue that there are other problems with additional analyses presented in Kirsch et al. (2008). We provide the results of our own reanalyses of the same data in our paper.

In the meta-analysis, precision weighted analyses (fixed and random effects) were performed in RevMan 5.0[1]. Sample size weighted analyses were performed in Open Office Calc[2], weighting by the harmonic mean of the placebo and drug sample sizes. We also performed the same analyses using STATA 9[3].

In the meta-regression, precision and sample size weighted analyses were performed in SPSS[4] using macros by DB Wilson[5].

Notes

1. Review Manager (RevMan). Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2008.
2. OpenOffice.org 3.0.1.
3. STATA 9, StataCorp.
4. SPSS Statistics 17.0, SPSS Inc, 2008.
5. http://mason.gmu.edu/~dwilsonb/ma.html, DB Wilson, 2006.

Table SA1.

ANTIDEPRESSANTS

sd	change	d	n	change*N	n/var	change*n/var	n*var
7.633587786	10	1.31	74	740	1.269914	12.69914	4312.103024
8.32173913	9.57	1.15	101	966.57	1.458455488	13.95741902	6994.385558
7.606837607	8.9	1.17	153	1361.7	2.644132054	23.53277528	8853.188692
8.085106383	11.4	1.41	156	1778.4	2.386454294	27.20557895	10197.55545
8.661971831	12.3	1.42	175	2152.5	2.332407958	28.68861789	13130.2073
7.941176471	10.8	1.36	57	615.6	0.903868313	9.761777778	3594.550173
7.947019868	12	1.51	86	1032	1.361726389	16.34071667	5431.340731
7.462686567	10	1.34	80	800	1.43648	14.3648	4455.335264
8.680555556	12.5	1.44	22	275	0.29196288	3.649536	1657.744985
8.674698795	7.2	0.83	18	129.6	0.239201389	1.72225	1354.507185
9.565217391	11	1.15	181	1991	1.978285124	21.76113636	16560.30246
5.774509804	5.89	1.02	299	1761.11	8.96687142	52.81487267	9970.144079
7.805309735	8.82	1.13	297	2619.54	4.875017354	42.99765306	18094.08944
8.083832335	13.5	1.67	24	324	0.367262551	4.958044444	1568.360285
9.609375	12.3	1.28	51	627.3	0.552306167	6.793365854	4709.344482
8.861788618	10.9	1.23	36	392.4	0.458415958	4.996733945	2827.12671
10.43010753	9.7	0.93	33	320.1	0.30334467	2.942443299	3589.97572
6.79144385	12.7	1.87	36	457.2	0.780509641	9.912472441	1660.453545
6.75	10.8	1.6	40	432	0.877914952	9.481481481	1822.5
7.01754386	8	1.14	40	320	0.81225	6.498	1969.836873
8.389830508	9.9	1.18	41	405.9	0.582475258	5.766505051	2885.959494
7.819548872	10.4	1.33	37	384.8	0.60511557	6.293201923	2262.377749
10.1010101	10	0.99	40	400	0.39204	3.9204	4081.216202
8.198198198	9.1	1.11	39	354.9	0.580266876	5.280428571	2621.207694
10.91666667	13.1	1.2	13	170.3	0.109084552	1.429007634	1549.256944
7.109375	9.1	1.28	403	3667.3	7.973375196	72.55771429	20368.91479
6.18556701	6	0.97	19	114	0.496586111	2.979516667	726.9635455
11	8.8	0.8	20	176	0.165289256	1.454545455	2420
7.398373984	9.1	1.23	19	172.9	0.347121121	3.158802198	1039.982814
8.175182482	11.2	1.37	231	2587.2	3.456344866	38.7110625	15438.56359
9.793103448	14.2	1.45	46	653.2	0.479641936	6.810915493	4411.624257
7.853107345	13.9	1.77	64	889.6	1.03775995	14.42486331	3946.962878
10.25862069	11.9	1.16	65	773.5	0.617639997	7.349915966	6840.5544
7.952755906	10.1	1.27	69	696.9	1.090972454	11.01882178	4363.996528
8.208955224	11	1.34	227	2497	3.368604959	37.05465455	15296.83671
			n wheight sum	33039.52	p weight sum	533.2891705	
			sum n	3292	sum n/var	55.5990987	
			n wheight avg	10.0363062	p weight avg	9.591687329	
			var nweight	0.019470548	var pweight	0.017985903	
			sum n*var	211007.4696			
			treatment-placebon (sample size) weighted			treatment - placebo precision weighted	
			2.18459517			1.782339093	
			var treat-placebo n w			var treat-placebo p w	
			0.057976815			0.053726042	

Table SA1. Continued

PLACEBOS							
sd	Change	d	np	change*np	np/var	chnge*np/var	np*var
7.748031496	9.84	1.27	70	688.8	1.16604493	11.47388211	4202.239444
8.695652174	8	0.92	52	416	0.6877	5.5016	3931.94707
7.606837607	8.9	1.17	77	685.3	1.330706981	11.84329213	4455.526335
8.11965812	9.5	1.17	75	712.5	1.137590028	10.80710526	4944.663599
8.828828829	9.8	1.11	47	460.6	0.60296439	5.90905102	3663.566269
7.961165049	8.2	1.03	57	467.4	0.899335217	7.37454878	3612.668489
7.920792079	8	1.01	90	720	1.434515625	11.476125	5646.505245
7.416666667	8.9	1.2	78	694.2	1.418002777	12.62022472	4290.541667
8.73015873	5.5	0.63	24	132	0.314895868	1.731927273	1829.176115
8.54368932	8.8	1.03	24	211.2	0.328791322	2.893363636	1751.871053
9.545454545	8.4	0.88	163	1369.2	1.78893424	15.02704762	14851.8595
5.542857143	5.82	1.05	56	325.92	1.822722925	10.60824742	1720.502857
7.902777778	5.69	0.72	48	273.12	0.768566937	4.37314587	2997.787037
8.076923077	10.5	1.3	24	252	0.367891156	3.862857143	1565.680473
9.714285714	6.8	0.7	53	360.4	0.561634948	3.819117647	5001.469388
8.787878788	5.8	0.66	34	197.2	0.440261593	2.553517241	2625.711662
10.43478261	7.2	0.69	33	237.6	0.303072917	2.182125	3593.194707
6.785714286	7.6	1.12	38	288.8	0.825263158	6.272	1749.744898
6.811594203	4.7	0.69	38	178.6	0.819004074	3.849319149	1763.116992
7.045454545	6.2	0.88	40	248	0.805827263	4.996129032	1985.53719
8.403361345	10	1.19	42	420	0.594762	5.94762	2965.892239
7.790697674	6.7	0.86	37	247.9	0.609605703	4.084358209	2245.713899
10	4.1	0.41	42	172.2	0.42	1.722	4200
8.108108108	3	0.37	37	111	0.562811111	1.688433333	2432.432432
11.01010101	10.9	0.99	12	130.8	0.098991667	1.079009174	1454.667891
7.192982456	8.2	1.14	51	418.2	0.985716835	8.082878049	2638.688827
7.469879518	6.2	0.83	22	136.4	0.394271592	2.444483871	1227.5802
9.183673469	4.5	0.49	21	94.5	0.248992593	1.120466667	1771.137026
7.790697674	6.7	0.86	10	67	0.164758298	1.103880597	606.9497025
8.170731707	6.7	0.82	92	616.4	1.378053018	9.232955224	6141.99881
11.1627907	4.8	0.43	47	225.6	0.37718316	1.810479167	5856.57112
7.875	9.45	1.2	78	737.1	1.257747543	11.88571429	4837.21875
10.20689655	8.88	0.87	75	666	0.719902757	6.392736486	7813.555291
7.975806452	9.89	1.24	79	781.31	1.241874983	12.28214358	5025.465596
8.252173913	9.49	1.15	75	711.75	1.101347878	10.45179136	5107.378072
			n wheight sum	14455	p weight sum	218.5035761	
			sum n	1841	sum p	27.97974549	
			n wheight avg	7.851711027	p weight avg	7.809348235	
			var nweight	0.038506267	var pweight	0.035740139	
			sum n*var	130508.5599			

Table SA1. Continued

ANTIDEPRESSANT-PLACEBO DIFFERENCE antidepressant treated - placebo treated						
vardif	dif change	ni*	difchng*ni	1/vardifmean	dch*ni/vd	vardif*ni
118.303655	0.16	35.97222222	5.755555556	0.607882491	0.097261198	4255.645351
144.865709	1.57	34.32679739	53.8930719	0.467337919	0.733720532	4972.775837
115.727957	0	51.22173913	0	0.885209427	0	5927.787211
131.297793	1.9	50.64935065	96.23376623	0.77036676	1.463696844	6650.147967
152.977974	2.5	37.04954955	92.62387387	0.479107512	1.197768779	5667.765046
126.442433	2.6	28.5	74.1	0.450798034	1.172074887	3603.609331
125.894072	4	43.97727273	175.9090909	0.698586808	2.794347231	5536.477936
110.698635	1.1	39.49367089	43.44303797	0.713590793	0.784949873	4371.895468
151.567716	7	11.47826087	80.34782609	0.151498029	1.060486206	1739.733786
148.245026	−1.6	10.28571429	−16.45714286	0.138465405	−0.221544648	1524.805986
182.609086	2.6	85.76453488	222.9877907	0.939425516	2.442506342	15661.38335
64.0682288	0.07	47.16619718	3.301633803	1.514804132	0.106036289	3021.854712
123.376757	3.13	41.32173913	129.3370435	0.663900274	2.078007856	5098.142153
130.585032	3	12	36	0.183788293	0.551364878	1567.020379
186.707435	5.5	25.99038462	142.9471154	0.278465748	1.531561612	4852.598042
155.758111	5.1	17.48571429	89.17714286	0.224577703	1.145346285	2723.541828
217.671831	2.5	16.5	41.25	0.151604366	0.379010916	3591.585213
92.1696279	5.1	18.48648649	94.28108108	0.401131375	2.045770014	1703.892581
91.9603156	6.1	19.48717949	118.8717949	0.423718464	2.584682627	1792.047176
98.8843516	1.8	20	36	0.404512942	0.728123296	1977.687031
141.005738	−0.1	20.74698795	−2.074698795	0.294277255	−0.029427726	2925.444344
121.840315	3.7	18.5	68.45	0.303676169	1.123601824	2254.045824
202.030405	5.9	20.48780488	120.8780488	0.202769322	1.196338998	4139.159518
132.951871	6.1	18.98684211	115.8197368	0.285702856	1.74278742	2524.336178
240.395935	2.2	6.24	13.728	0.051896664	0.114172661	1500.070637
102.28221	0.9	45.27092511	40.7438326	0.87726414	0.789537726	4630.410247
94.0603393	−0.2	10.19512195	−2.03902439	0.219776734	−0.043955347	958.9566295
205.339858	4.3	10.24390244	44.04878049	0.099342514	0.427172811	2103.481476
115.430908	2.4	6.551724138	15.72413793	0.111727651	0.268146362	756.2714653
133.594465	4.5	65.79566563	296.0804954	0.985236753	4.43356539	8789.936766
220.512771	9.4	23.24731183	218.5247312	0.21114328	1.984746831	5126.329157
123.68692	4.45	35.15492958	156.4394366	0.568606302	2.530298046	4348.204961
209.420036	3.02	34.82142857	105.1607143	0.332431046	1.003941759	7292.304814
126.859815	0.21	36.83108108	7.734527027	0.58077154	0.121962023	4672.384134
135.485323	1.51	56.37417219	85.125	0.829987711	1.253281443	7637.872767
		ni weight sum	2804.3464	p weight sum	39.59134124	
		sum ni	1056.604711	sum p	16.50338193	
		ni weight avg	**2.654111202**	p weight avg	**2.398983518**	
		var nweight	0.13068599	var pweight	0.060593641	
		sum n*var	145899.6053			

Table SA1. Continued

Published?	Drug	TRIAL NAME	BASELINE HAMD (ANTDEP)	BASELINE HAMD (PLACEBO)
	Nefazadone	BMS 030A2-0004/0005	23.4	24
1	Nefazadone	BMS 03A0A-003	25.4	25.9
	Nefazadone	BMS 03A0A-004A	23.4	23.5
1	Nefazadone	BMS 03A0A-004B	25.3	25
1	Nefazadone	BMS 0A2-0007	25.7	26.4
	Nefazadone	BMS CN104-002	23.3	23.1
1	Nefazadone	BMS CN104-005	24.5	23.3
	Nefazadone	BMS CN104-006	23.8	23.5
1	Fluoxetine	ELC 19	28.6	28.2
	Fluoxetine	ELC 25	26.2	25.8
1	Fluoxetine	ELC 27	27.5	28.2
1	Fluoxetine	ELC 62 (mild)	17	17.4
	Fluoxetine	ELC 62 (moderate)	24.3	24.3
	Paroxetine	GSK 01-001	28	27.4
1	Paroxetine	GSK 02-001	26.6	25.9
1	Paroxetine	GSK 02-002	25	24.9
1	Paroxetine	GSK 02-003	28.6	28.9
1	Paroxetine	GSK 02-004	28.9	27.3
1	Paroxetine	GSK 03-001	24.9	24.8
1	Paroxetine	GSK 03-002	24.9	25.6
	Paroxetine	GSK 03-003	25.7	27
1	Paroxetine	GSK 03-004	27.6	27
1	Paroxetine	GSK 03-005	26.1	26.8
1	Paroxetine	GSK 03-006	29.7	28.7
	Paroxetine	GSK PAR 07	30.5	28.3
1	Paroxetine	GSK PAR 09	25.2	24.5
1	Paroxetine	GSK UK 06	23.7	24.2
	Paroxetine	GSK UK 09	26.8	25.5
	Paroxetine	GSK UK 12	22.8	22.3
1	Venlafaxine	W 203	25.6	25.3
1	Venlafaxine	W 206	28.2	28.6
1	Venlafaxine	W 301	25.4	24.6
1	Venlafaxine	W 302	25	24.4
	Venlafaxine	W 303	23.6	24.6
1	Venlafaxine	W 313	25.7	25.4