STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:943–955
Published online 30 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2240

Bias resulting from the use of 'assay sensitivity' as an inclusion criterion for meta-analysis

Lois A. Gelfand[1], Daniel R. Strunk[1], Xin M. Tu[2], Ronald E. S. Noble[1] and Robert J. DeRubeis[1,*,†]

[1]*Department of Psychology, University of Pennsylvania, Philadelphia, PA, U.S.A.*
[2]*Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, U.S.A.*

SUMMARY

Assay sensitivity has been proposed as a criterion for including psychiatric clinical outcome studies in meta-analyses. The authors assess the performance of assay sensitivity as a method for determining study appropriateness for meta-analysis by calculating expected standard drug vs placebo effect sizes for various combinations of high quality and flawed studies. In the absence of flawed studies, expected effect sizes are close to unbiased only when sample sizes are very large. In the presence of flawed studies, expected effect sizes tend to be substantially biased except under simultaneous conditions of high power, a large proportion of flawed studies, and a population standard vs placebo effect size of flawed studies considerably lower than that of high quality studies. The authors conclude that this method is not robust and can lead to serious bias. Unless it can be shown that specific conditions hold, assay sensitivity should not be used to make quality judgments of studies. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: comparative outcome research; placebo controls; assay sensitivity; meta-analysis

INTRODUCTION

The summary of information from clinical research studies is crucial for health care decisions and public health policy. A meta-analysis combines the quantitative information reported in completed studies into a single estimate of the effect size of a treatment, usually in comparison to a control condition. (The effect sizes in this paper indicate comparisons between conditions, rather than change within treatments, unless otherwise noted.) The ability of meta-analysis to contribute to useful health care decisions rests on the ability of reviewers to assess the quality of individual studies, in order to remove, to the extent possible, studies that would systematically bias the effect size estimate calculated [1–4]. So far, most of the methods proposed

*Correspondence to: Robert J. DeRubeis, Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104-6241, U.S.A.
†E-mail: derubeis@psych.upenn.edu

Received 13 February 2004
Accepted 7 March 2005

for assessing study quality have been labour-intensive, and potentially open to criticisms of subjectivity. The methods of each study are scrutinized according to an individualized list of criteria [3], a previously developed checklist [1] or a quality scale [2, 4, 5]. (The CONSORT statements [6–8] for improving the reporting of randomized controlled trials, though professedly not quality checklists, contain methodological items that could be used in these ways.) Trials that do not meet a threshold can be excluded, analysed separately, or given a lower weight than other studies.

'Assay sensitivity' (AS) is a concept used by the United States Food and Drug Administration (FDA) to interpret the results of a single three-arm study including a standard drug, an investigational drug, and a pill placebo. A study in which the standard drug significantly outperforms the placebo demonstrates AS and is considered informative; a study lacking AS (a 'failed' trial) is considered uninformative. The FDA makes a dichotomous decision to approve or not to approve a new drug. Klein and others [9, 10] have suggested using AS to determine whether a given trial comparing a standard drug to a psychotherapy should be included in a meta-analysis. That is, one would include those studies with AS, and exclude studies either (a) without AS or (b) in which AS cannot be determined because of the absence of a placebo condition. The advantages of the AS approach are obvious: A quick look at the results (the significance of the standard vs placebo difference, or the recognition that there is no placebo condition) could substitute for (or serve as a decisive, objective addition to) the process of scrutinizing individual methods of a study. Klein [9] suggested this method (which we will call 'the AS method') in the context of comparisons between drugs and psychotherapies, but the same reasoning would apply to all studies involving the comparison of a standard drug and another treatment, including other drugs.

In this paper, we investigate the theoretical performance of the AS method by calculating the effect sizes one would expect in the long run if this method were adopted. We approach this discussion in three stages. First (Model 1), we discuss the circumstances under which the AS method would correctly classify all or nearly all studies. If the combinations of effect sizes and powers of studies required for this to occur are plausible in a psychiatric literature, the AS method could produce unbiased results. Second (Model 2), we examine how discarding good studies from meta-analyses biases the estimations of 'good study' effect sizes. During this discussion, we consider the validity of sample characteristics as a methodological issue, and briefly touch upon the difference between the bias expected from an infinite number of trials vs the bias possible in a small number of trials. Third (Model 3), we examine how the combination of discarding good and bad studies from meta-analysis affects the estimation of the good study effect size.

METHOD

We consider two populations of three-arm studies: a population of 'good' studies, which are well-conducted and expected to produce unbiased effect size estimates, and a population of 'bad' studies, which are poorly conducted and may be biased against the standard drug, biased for the standard drug, or unbiased, depending on the nature of the study flaws. For example, inadequate dosing can bias a study against the standard drug [10], and non-blind outcome evaluation can bias a study in favour of the standard drug [3]. We assume that a study flaw changes the location, but not the shape of the effect size distribution of a treatment condition.

We consider good and bad studies with balanced designs and equal known variances for the three conditions; standard (S), placebo (P), and other (O) treatment. In addition, we assume that the outcome measure X has a normal distribution in all three study arms.

Let $S =$ standard, $P =$ placebo, and $O =$ other treatment. X_S, X_P, and X_O are independent. Let $X_S, X_P, X_O \sim \mathrm{N}(\mu_i, \sigma^2)$ and $n = n_S = n_P = n_O$, $i \in \{S, P, O\}$.

The population effect size is

$$\delta_{ij} = (\mu_i - \mu_j)/\sigma \quad \text{where } i, j \in \{S, P, O\},\ i \neq j \tag{1}$$

The sample estimator of the population effect size is

$$d_{ij} = (\bar{X}_i - \bar{X}_j)/\sigma \tag{2}$$

Under these assumptions, each d_{ij} follows a normal distribution with mean δ_{ij} and standard error $(2/n)^{1/2}$, where n is the cell sample size [11]. We assume that the quality of a study only affects the mean response, so that the probability distributions of the standard vs placebo effect size estimators from both good and bad studies can be represented by normal distributions with standard errors of $(2/n)^{1/2}$. Effect size estimation involving the third arm of studies will be considered later.

All statistical tests are performed using a two-sided alpha of 0.05.

RESULTS

Model 1: *bad studies present*, *misclassification absent*

In order for the AS method to include all good studies and exclude all bad studies from meta-analysis, the effect of the standard drug condition must be significantly greater than the effect of the placebo condition in all of the good studies and in none of the bad studies. Thus, for all good studies to demonstrate AS, all standard vs placebo effect sizes must exceed the critical effect size that cuts off a right tail of area 0.025 on the null sampling distribution (all good studies must be powered to 100 per cent). Similarly, for no bad study to demonstrate AS, no standard vs placebo effect size from a bad study should exceed that critical effect size. A situation close to this is shown in Figure 1, in which the cell sample size is 60, and 99.9 per cent of the good studies demonstrate AS while 0.1 per cent of bad studies demonstrate AS. Achieving this degree of separation of good from bad studies requires a good effect size of 0.91 and a bad effect size of -0.19, assumptions that are unrealistic for standard vs placebo comparisons of treatments for psychiatric disorders. The effect size of 0.91 is considerably larger than the mean effect size estimates that are typically calculated in meta-analyses of treatments for mood and anxiety disorders [12, 13], and a negative effect size for bad studies would mean that in those studies, placebo tends to consistently outperform standard treatment. A substantial increase in sample size would be required to model more realistic effect sizes, but larger sample sizes are not commonly found in the literature. In short, the ideal case is unlikely to be a realistic case in the foreseeable future; AS cannot be expected to correctly classify all (or nearly all) good and bad studies.

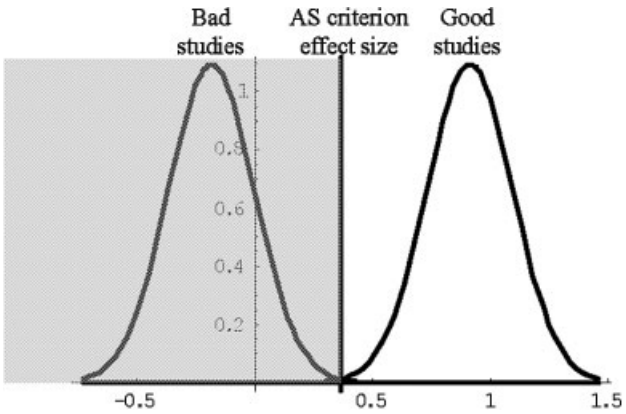

Figure 1. An example of probability distributions of good and bad study standard vs placebo effect size estimators in which the AS criterion correctly categorizes nearly all studies. Studies with effect size estimates greater than the criterion have AS, while those in the shaded area, with effect size estimates less than the criterion, do not. In this case, $n = 60$, the criterion effect size is 0.36, the population standard vs placebo effect size for the good studies is 0.91, that of the bad is -0.19, and 50 per cent of all studies are bad. In this example, 99.9 per cent of good studies demonstrate AS (good studies are powered to 99.9 per cent), while 0.1 per cent of bad studies demonstrate AS.

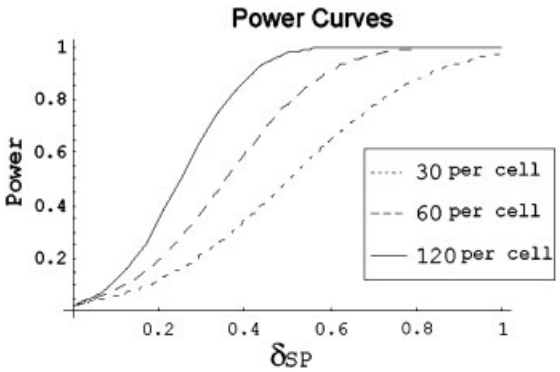

Figure 2. Power as a function of cell sample size and population effect size of a standard vs placebo comparison (δ_{SP}). Note: two-tailed alpha equals 0.05.

Model 2: bad studies absent

In the absence of bad studies, the influence of AS on the expected standard vs placebo effect size is determined by the power of a study. Good studies that lack AS exhibit type II error; using the AS method in the absence of bad studies involves excluding good studies due to chance. As power increases, the proportion of good studies that are incorrectly excluded from meta-analysis decreases, and the bias caused by the AS method decreases. Figure 2 depicts power curves for the standard vs placebo difference for three cell sample sizes that have been commonly used in studies of mood and anxiety disorders [14–16].

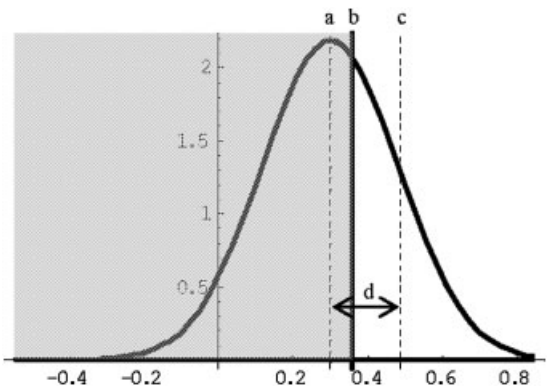

Figure 3. Left-truncated normal distribution showing that the estimate of the population effect size of a standard vs placebo comparison will be biased if trials without AS are excluded. In this case, n/cell is 60, and the population effect size of good studies is 0.30[a]. The effect size criterion for AS, d_{SP}^{CRIT}, is 0.36[b]. Power is 0.38; the 62 per cent of studies without AS (shaded region) are excluded and the biased effect size estimate calculated from the remaining 38 per cent (unshaded) is 0.48[c]. Thus, the bias is 0.18[d].

If all the studies under consideration are good, the biasing effect on the standard vs placebo comparison of excluding non-AS studies is depicted in Figure 3, in which the normal probability distribution of effect size estimator d_{SP} is left-truncated at the critical value d_{SP}^{CRIT}, leaving only the effect size estimators with AS, d_{SP}^{AS}, for meta-analysis. The mean of the standard vs placebo effect size estimators from studies with AS, $E(d_{SP}^{\mathrm{AS}})$, is calculated from the left-truncated distribution as in Reference [17]:

$$E(d_{SP}^{\mathrm{AS}}) = E(d_{SP} | d_{SP} \geqslant d_{SP}^{\mathrm{CRIT}}) = \delta_{SP} + \left(\frac{2}{n}\right) \frac{\phi(d_{SP}^{\mathrm{CRIT}})}{1 - \Phi(d_{SP}^{\mathrm{CRIT}})} \tag{3}$$

where $\phi(t)$ is the standard normal probability density function, $\Phi(t)$ is the standard normal cumulative distribution function, d_{SP}^{AS} is a sample estimator of the standard vs placebo effect size from a study with AS.

The critical value d_{SP}^{CRIT} is chosen so that $d_{SP}^{\mathrm{CRIT}} / \sqrt{2/n} = Z_{1-(\alpha/2)} = 1.96$

The mean of the left-truncated distribution is always larger than or equal to the untruncated one, thus $E(d_{SP}^{\mathrm{AS}}) \geqslant \delta_{SP}$.

For example, as shown in Figure 3, the population effect size of good studies is 0.30, but the estimate calculated from just the studies with AS is 0.48.

Figure 4 displays the biased effect size estimate expectations as a function of the population effect size for the three sample sizes used in Figure 2. Only when power is extremely high, and the percentage of excluded studies, therefore, extremely low, is the degree of bias negligible. A cell sample size of 120 and a population effect size exceeding 0.50, and a cell sample size of 60 and a population effect size exceeding 0.70, are associated with power (Figure 2) of higher than 0.96, and therefore with negligible bias (Figure 4). A cell sample size of 60 and an effect size of 0.50 is associated with a power of 0.78, the exclusion of $1 - 0.78 = 22$ per cent of studies, and bias of $0.57 - 0.50 = 0.07$. A cell sample size of 30 and an effect size of

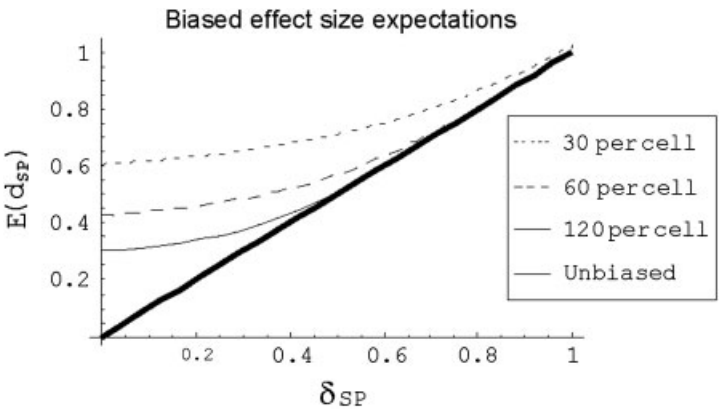

Figure 4. Biased effect size expectations as a function of good study population effect size and cell sample size, when only AS studies are included in meta-analysis. The thick diagonal line represents the unbiased population effect size. The degree of bias caused by excluding non-AS studies is shown by the vertical distance between a point on one of the curved lines and the corresponding point on the thick diagonal line.

0.50 is associated with 0.47 power, the exclusion of $1 - 0.47 = 53$ per cent of studies, and a biased effect size of 0.71, which is considerably (0.21) higher than the unbiased effect size of 0.50.

Drug responsive subpopulations. Each combination of effect size and sample size is associated with a specific biased effect size. It is widely recognized [18–20] that a population suffering from what is defined as a single psychiatric disorder can be composed of subpopulations heterogeneous in severity, type, time course of symptoms, or responsiveness to active or placebo treatment. A study demonstrating AS is sometimes described as having enrolled a drug-responsive sample [21, 22], which could be assumed to represent not the population that meets the inclusion/exclusion criteria of a trial, but a more drug-responsive subset of this population. Drug effects in this 'subpopulation' would overestimate the effects in the larger population; conversely, drug effects in the larger population would underestimate effects in the subpopulation.

There are several problems with using a sample outcome to define a subpopulation when performing a meta-analysis. One can conceive of a subpopulation of, for example, depressed outpatients, who are more responsive to a drug on average than are depressed outpatients as a whole. The subpopulation might be defined *a priori* by, for example, severity criteria, diagnostic subtype, or (lack of) comorbidity with other disorders. The subpopulation effect size would be greater than the effect size of the entire depressed outpatient population. A meta-analyst might be interested in estimating the effect size of the subpopulation and thus would reasonably exclude studies that did not explicitly sample from the subpopulation. The meta-analytic effect size would then provide an estimate of the effect of the treatment in the specified subpopulation.

However, combining effect size estimates from studies with sample characteristics that are defined *post hoc* from statistical analysis of outcome data, such as the assessment of AS, is another matter. In a population or subpopulation defined *a priori*, a meta-analysis averaging the

effect size estimates from an infinite number of studies sampling from the specified population, regardless of the sample size of those studies, will theoretically produce the population effect size. This is the basis of meta-analysis; a meta-analysis of a finite number of studies of possibly varying sample sizes provides an estimate of the one true population effect size. When sample characteristics are defined *post hoc*, what does a meta-analysis of studies with AS estimate? Taking the mean of an infinite number of studies with AS will, unlike the mean of an *a priori* subpopulation, result in different numerical values depending on the sample sizes of the studies. The values shown in Figure 4 can be interpreted as the results of such meta-analyses. Thus, a meta-analysis of studies with AS that all have the same sample size, for example, 30/cell, will estimate the biased effect size depicted in Figure 4; that is, if the population effect size is 0.50, the meta-analysis would estimate the biased effect size of 0.71. If the meta-analysis includes only studies of sample size 60/cell, it will estimate the biased effect size of 0.57. There is no one true effect size of a *post hoc* 'population' for a meta-analysis to estimate. Taking the mean effect size estimate of AS studies of various sample sizes would result in values that depend on the mix of the sample sizes as well as the true population effect size; there is no coherent interpretation of the result of such a meta-analysis. Studies cannot be excluded from meta-analyses on the basis of *post hoc* sampling characteristics such as those defined by AS. Subpopulations must be defined *a priori*, and anything that might be considered a sampling problem must be identifiable independently of outcome results.

Bias in small numbers of trials. Before we consider more complex assumptions, there is another aspect of bias that should be elucidated. The biased effect sizes shown in Figure 4 were calculated for an infinite number of trials. In a finite number, the proportion of studies discarded and the consequent degree of bias can be much higher. For example, for an infinite number of good studies powered to 0.78, 22 per cent would be discarded. For five such studies, to discard one study would be already to discard 20 per cent. Using the binomial distribution, we find that there is a 71 per cent chance that one or more of five studies would be discarded, and a 30 per cent chance that two or more would be discarded, leading to even more bias than shown in Figure 4. Unless very large sample sizes are employed, and medium-to-large [11] effect sizes represent reality, using the AS method can lead to substantially biased meta-analytic estimates under these circumstances.

Effect sizes involving the third treatment condition. The AS method uses as a criterion the standard vs placebo effect size, but not effect sizes involving the third treatment condition. In addition, the conditions are assumed to be independent. That is, the response of any individual in one of the conditions is assumed not to affect the response of any individuals in the other two groups. For example, the response of any individual in the placebo condition does not affect the response of any individual in the standard drug condition or the third condition. Given the lack of contribution of the third treatment condition to the determination of AS, along with the independence of the conditions, the estimate of the within-treatment mean effect of the third condition is unaffected by whether or not the AS method is used. The AS method increases the standard vs placebo effect size by selecting studies in which, on an average, the within-treatment mean effect of the standard drug is overestimated and the within-treatment mean effect of the placebo is underestimated. Under the model assumptions, the contributions to bias come equally from the overestimate of the within-treatment mean effect of the standard drug condition (from the right tail of the distribution of mean effects

of the standard drug condition) and the underestimate of the within-treatment mean effect of the placebo condition (from the left tail of the distribution of mean effects of the placebo condition). That is, the degree to which the mean effect of the standard drug is overestimated equals the degree to which the mean effect of the placebo is underestimated. Because the estimate of the mean effect of the third condition is unbiased, it follows that the bias of the standard vs other treatment equals the bias of the other treatment vs placebo. In other words, in the absence of bad studies, the bias in the effect size estimates involving the third treatment condition is equal to one-half of the bias in the standard vs placebo effect size (for a formal approach to this question, see Appendix A).

For example, consider again Figure 3, in which n/cell is 60, the population standard vs placebo effect size is 0.30, and the biased standard vs placebo effect size is 0.48 (a bias of 0.18). If the third treatment condition is equally effective as the standard (so that its effect size compared to placebo is also 0.30), the biased effect size of the third condition vs placebo is 0.39 (a bias of 0.18/2 or 0.09), and the biased effect size of the standard vs the third condition is 0.09 (the unbiased effect size is 0.0). In all cases, the third treatment condition's effect compared to placebo is overestimated, and its effect compared to the standard drug is underestimated.

Model 3: bad studies present, misclassification present

We have shown in Model 2 that when all studies are good, the AS method introduces bias into effect size estimates, and this bias cannot be dismissed by referring to a drug-responsive subpopulation. If all studies are good, there is no need to eliminate studies from meta-analyses, and no possible role for the AS method. However, if AS also excluded bad studies from meta-analyses, the disadvantage of excluding good studies could be outweighed by the benefits of excluding bad ones if certain assumptions hold. Although we do not know the population effect sizes for good and bad studies, we can examine the effects of the AS method in a number of circumstances, and show what kinds of conditions must be assumed for the AS method to produce unbiased long-run estimates of standard vs placebo effect sizes.

Considering the presence of bad studies requires taking into account, in addition to the sample size and population effect size for good studies, the population effect size for bad studies and the proportion of good and bad studies in the overall population of studies. As in the discussion of good studies alone, the two-sided alpha is 0.05. To minimize the number of figures in this paper, we consider only one sample size: 60/cell. When the sample size is 60/cell and the population effect size for good studies is 0.50, power is 0.78, close to the 0.80 typically considered adequate [23]. When good study effect sizes are overestimated, due to overly optimistic estimates resulting from publication bias [23] or other factors, actual power can be lower than nominal power. To take into account both the possibility of adequately and inadequately powered studies, we examine cases in which the sample size is 60/cell, and the population effect size for good studies is either 0.50 or 0.30 (in the latter case, power is 0.38). We vary the population effect size of bad studies as well as the proportion of bad and good studies in the population of studies.

The biased estimate, $E(d^{\mathrm{AS}}_{SP_{\mathrm{mixture}}})$, of the population effect size of the standard vs placebo difference is the mean of a left-truncated mixture probability distribution that is initially composed of the sum of the two normal distributions that represent the distributions of good and bad study effect sizes. This biased estimate can be calculated as a weighted average of the

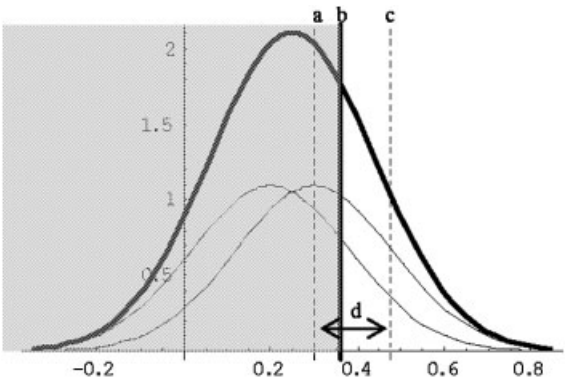

Figure 5. Left-truncated mixture distribution depicting the calculation of the long-run effect size estimate resulting from using the AS method in the presence of good and bad studies. In this case, the cell sample size is 60, the effect size of good studies is 0.30[a], the effect size of bad studies is 0.20, and the proportion of bad studies is 50 per cent. The effect size criterion for AS, d_{SP}^{CRIT}, is 0.36[b]. The shaded region shows non-AS studies, which would be excluded from meta-analysis. The biased effect size calculated from the AS studies (unshaded) is 0.47[c]. Thus, the bias is 0.17[d].

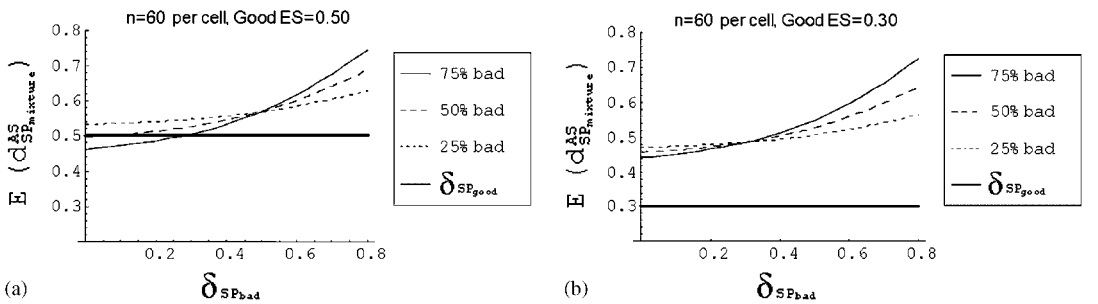

Figure 6. (a) and (b) Standard vs Placebo effect size expectations of mixtures of good and bad studies using the AS method, plotted against population effect sizes of bad studies. For (a) and (b), sample size is 60/cell. For A, the effect size of good studies is 0.50. For B, it is 0.30 (in each case represented by a thick horizontal line). The degree of bias expected from using the AS method is shown by the difference between a point on the curved line and the corresponding point on the thick horizontal line indicating the unbiased value. Values above the thick horizontal line reflect upward bias, in which the effectiveness of the standard drug is overestimated; values below the thick line reflect downward bias, in which the standard drug's effectiveness is underestimated.

means of the good and bad effect size distributions left-truncated at the critical value d_{SP}^{CRIT}, where the weights are the proportions of good and bad studies, respectively, represented in the left-truncated (AS) mixture distribution (see Appendix B). An example of this is depicted in Figure 5, and the biased estimates are shown in Figure 6.

In Figure 6, biased population standard vs placebo effect size estimates are shown based on using the AS method to exclude studies. As the bad study effect size increases, more and more bad studies are misclassified as good, and the expected effect size of AS studies increases, in

a concave manner. When the effect size of bad studies equals the effect size of good studies, the biased effect size is the same as when all studies are good; that is, the biased effect sizes include those shown in Figure 4. This is true regardless of the proportion of bad studies. For example, when the sample size is 60/cell, and both effect sizes equal 0.50, the biased effect size is 0.57 (a bias of 0.07); when both equal 0.30, the biased effect size is 0.48 (a bias of 0.18). The bias curves pivot about this point, and become steeper as the proportion of bad studies increases. Naturally, upward bias is present when the bad study effect size equals or surpasses that of the good studies. In some cases when the bad study effect size is lower than that of the good studies, this bias is reduced so that it is non-existent or negligible, but in other cases, the upward bias is not substantially counterbalanced. In general, the AS method seems to introduce minimal bias when the studies considered for meta-analysis are highly powered, the proportion of bad studies is substantial, and the effect size of the bad studies is considerably lower than that of the good studies, simultaneously.

Effect sizes involving the third treatment condition. A biased effect size estimate involving the third treatment condition in the presence of bad studies is a weighted average (similar to that described above) of the biased estimates from the good and bad studies, where the weights are the proportions of good and bad studies, respectively, represented in the left-truncated (AS) mixture distribution (see Appendix B).

Consider again Figure 5, in which n/cell is 60, the population standard vs placebo effect size for good studies is 0.30, and the population standard vs placebo effect size for bad studies is 0.20. Assume that the flaws in the bad studies do not affect the relative effectiveness of the standard and third treatment conditions, and that the third treatment condition is equally effective as the standard in both good and bad studies (so that its effect size compared to placebo is 0.30 in good studies and 0.20 in bad studies). The biased placebo vs standard effect size is 0.47 (a bias of 0.17), the biased effect size of the third treatment condition vs placebo is 0.37 (a bias of 0.07), and the biased effect size of the standard vs the third treatment condition is 0.10 (where the unbiased effect size is 0.0). Unlike the case in the absence of bad studies, in the presence of bad studies the bias in the comparison of the third treatment vs placebo is not necessarily equal to the bias in the comparison of the standard vs the third treatment.

DISCUSSION

At first glance AS, a metaphor that has been influential in the context of approving new drugs in the United States, appears to be an attractive, labour-saving alternative or additional criterion to traditional methods for assessing study quality and appropriateness for meta-analysis. However, there are several problems with the AS method. The AS method excludes good studies that exhibit type II error, biasing meta-analytic results in favour of the standard drug. The effect size estimates from these studies should be included in meta-analyses, because excluding them biases the results. The AS method does not live up to claims that it helps to select patients from subpopulations of interest; valid subpopulations must be defined *a priori*. The AS method, at best, can only help to eliminate bad studies that underestimate the effectiveness of a standard drug, not bad studies that exaggerate the drug's effectiveness. Even if bad studies tend consistently to underestimate the effectiveness of a standard drug, under

many circumstances using the AS method would still result in a substantially biased effect size estimate that would overestimate the drug's effectiveness.

Unless evidence is gathered to support the hypothesis that using the AS method reduces bias, meta-analysts should make quality judgments that are based on study methods, and that are independent of outcome. This means that, depending upon judgments made about each study's methodology, studies demonstrating AS can be excluded from meta-analyses, and studies not demonstrating AS, or not able to demonstrate AS because they do not contain a placebo condition, can be included. The placebo condition retains its importance in determining the efficacy of treatments, but the standard vs placebo comparison does not play a useful role in judging a study's quality.

APPENDIX A: EFFECT SIZE ESTIMATES INVOLVING THIRD TREATMENT ARM IN LEFT-TRUNCATED NORMAL DISTRIBUTION

Let δ_{ij} and d_{ij} be defined as in equations (1) and (2) in Method. Let $\theta_{ij} = d_{ij} - \delta_{ij}$. Then, it follows the model assumptions that $\theta_{ij} \sim \mathrm{N}(0, 2/n)$. Note that although X_S, X_P, X_O, are stochastically independent, θ_{ij} are not. For example, consider: $\theta_{SP} = [(\bar{X}_S - \bar{X}_P)/\sigma] - \delta_{SP}$ and $\theta_{SO} = [(\bar{X}_S - \bar{X}_O)/\sigma] - \delta_{SO}$. Both contain $\bar{X}_S$ and are not stochastically independent. However, θ_{ij} all have the same distribution.

Now, let $\theta_{SP}^{\mathrm{CRIT}} = d_{SP}^{\mathrm{CRIT}} - \delta_{SP}$. Then, we have

$$E(d_{SP}^{\mathrm{AS}}) = E(d_{SP} | d_{SP} \geqslant d_{SP}^{\mathrm{CRIT}}) = E(\theta_{SP} + \delta_{SP} | \theta_{SP} + \delta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}} + \delta_{SP}) = \delta_{SP} + E(\theta_{SP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}})$$

So, the bias for d_{SP}^{AS} is given by: $E(\theta_{SP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}})$.

Because θ_{SO} and θ_{OP} have the same distribution, it follows from the properties of conditional expectation [24] that

$$E(\theta_{SO} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = E(\theta_{OP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}})$$

And because any measure of the SP difference equals the sum of the SO and OP differences, $\theta_{SP} = \theta_{SO} + \theta_{OP}$; it follows that

$$\begin{aligned} E(\theta_{SP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) &= E(\theta_{SO} + \theta_{OP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = E(\theta_{SO} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) + E(\theta_{OP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) \\ &= 2E(\theta_{SO} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = 2E(\theta_{OP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) \end{aligned}$$

So, $E(\theta_{SO} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = E(\theta_{OP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = \frac{1}{2} E(\theta_{SP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}})$. Finally, since $d_{ij} = \theta_{ij} + \delta_{ij}$, it follows that

$$E(d_{SO} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = \delta_{SO} + E(\theta_{SO} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = \delta_{SO} + \tfrac{1}{2} E(\theta_{SP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}})$$

$$E(d_{OP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = \delta_{OP} + E(\theta_{OP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}}) = \delta_{OP} + \tfrac{1}{2} E(\theta_{SP} | \theta_{SP} \geqslant \theta_{SP}^{\mathrm{CRIT}})$$

Or, equivalently, from equation (3),

$$E(d_{SO} | d_{SP} \geqslant d_{SP}^{\mathrm{CRIT}}) = \delta_{SO} + \left(\frac{1}{n}\right) \frac{\phi(d_{SP}^{\mathrm{CRIT}})}{1 - \Phi(d_{SP}^{\mathrm{CRIT}})}$$

$$E(d_{OP}|d_{SP} \geqslant d_{SP}^{\mathrm{CRIT}}) = \delta_{OP} + \left(\frac{1}{n}\right) \frac{\phi(d_{SP}^{\mathrm{CRIT}})}{1 - \Phi(d_{SP}^{\mathrm{CRIT}})}$$

So, the bias for d_{SO}^{AS} and d_{OP}^{AS} is one-half of that for d_{SP}^{AS}.

APPENDIX B: EFFECT SIZE ESTIMATES FROM TRUNCATED MIXTURE DISTRIBUTION

A biased effect size estimate in the presence of both good and bad studies is a weighted average of the biased estimates from the good and bad studies, where the weights are the proportions of good and bad studies, respectively, represented in the left-truncated (AS) mixture distribution of standard vs placebo effect size estimators. If d^{CRIT} is the critical value, and $\Pr(\mathrm{AS}) = \Pr(d_{SP} \geqslant d^{\mathrm{CRIT}})$, the biased estimate $E(d_{ij_{\mathrm{mixture}}}^{\mathrm{AS}})$ is calculated as follows:

R $= \Pr(\mathrm{good} \cap \mathrm{AS}|\mathrm{AS})$ i.e. proportion of studies with AS that are good
$1 - R$ $= \Pr(\mathrm{bad} \cap \mathrm{AS}|\mathrm{AS})$ i.e. proportion of studies with AS that are bad
P $= \Pr(\mathrm{good})$ i.e. proportion of good studies in the mixture
$(1 - P)$ $= \Pr(\mathrm{bad})$ i.e. proportion of bad studies in the mixture
φ_{good} $= \Pr(\mathrm{AS}|\mathrm{good})$ i.e. power of good studies to detect standard vs placebo difference
φ_{bad} $= \Pr(\mathrm{AS}|\mathrm{bad})$ i.e. power of bad studies to detect standard vs placebo difference
$E(d_{ij_{\mathrm{good}}}^{\mathrm{AS}})$ = (biased) effect size estimate for good studies using good studies with AS
$E(d_{ij_{\mathrm{bad}}}^{\mathrm{AS}})$ = (biased) effect size estimate for bad studies using bad studies with AS

$$E(d_{ij_{\mathrm{mixture}}}^{\mathrm{AS}}) = E\lfloor (R) d_{ij_{\mathrm{good}}}^{\mathrm{AS}} + (1 - R) d_{ij_{\mathrm{bad}}}^{\mathrm{AS}} \rfloor = (R) E(d_{ij_{\mathrm{good}}}^{\mathrm{AS}}) + (1 - R) E(d_{ij_{\mathrm{bad}}}^{\mathrm{AS}})$$

$$R = \Pr(\mathrm{good} \cap \mathrm{AS}|\mathrm{AS}) = \frac{\Pr(\mathrm{good})\Pr(\mathrm{AS}|\mathrm{good})}{\Pr(\mathrm{AS})} = \frac{\Pr(\mathrm{good})\Pr(\mathrm{AS}|\mathrm{good})}{\Pr(\mathrm{AS} \cap \mathrm{good}) + \Pr(\mathrm{AS} \cap \mathrm{bad})}$$

$$= \frac{\Pr(\mathrm{good})\Pr(\mathrm{AS}|\mathrm{good})}{\Pr(\mathrm{good})\Pr(\mathrm{AS}|\mathrm{good}) + \Pr(\mathrm{bad})\Pr(\mathrm{AS}|\mathrm{bad})} = \frac{P\varphi_{\mathrm{good}}}{(1 - P)\varphi_{\mathrm{bad}} + P\varphi_{\mathrm{good}}}$$

$$1 - R = \Pr(\mathrm{bad} \cap \mathrm{AS}|\mathrm{AS}) = \frac{\Pr(\mathrm{bad})\Pr(\mathrm{AS}|\mathrm{bad})}{\Pr(\mathrm{AS})} = \frac{\Pr(\mathrm{bad})\Pr(\mathrm{AS}|\mathrm{bad})}{\Pr(\mathrm{AS} \cap \mathrm{good}) + \Pr(\mathrm{AS} \cap \mathrm{bad})}$$

$$= \frac{\Pr(\mathrm{bad})\Pr(\mathrm{AS}|\mathrm{bad})}{\Pr(\mathrm{good})\Pr(\mathrm{AS}|\mathrm{good}) + \Pr(\mathrm{bad})\Pr(\mathrm{AS}|\mathrm{bad})} = \frac{(1 - P)\varphi_{\mathrm{bad}}}{(1 - P)\varphi_{\mathrm{bad}} + P\varphi_{\mathrm{good}}}$$

Thus,

$$E(d_{ij_{\mathrm{mixture}}}^{\mathrm{AS}}) = \left[\frac{P\varphi_{\mathrm{good}}}{(1 - P)\varphi_{\mathrm{bad}} + P\varphi_{\mathrm{good}}}\right] E(d_{ij_{\mathrm{good}}}^{\mathrm{AS}}) + \left[\frac{(1 - P)\varphi_{\mathrm{bad}}}{(1 - P)\varphi_{\mathrm{bad}} + P\varphi_{\mathrm{good}}}\right] E(d_{ij_{\mathrm{bad}}}^{\mathrm{AS}})$$

REFERENCES

1. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association* 1996; **276**:637–639.
2. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *Journal of the American Medical Association* 1994; **272**:101–104.
3. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association* 1999; **282**:1054–1060.
4. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *The Lancet* 1998; **352**: 609–613.
5. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clinical Trials* 1996; **17**:1–12.
6. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine* 2001; **134**:663–694.
7. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Annals of Internal Medicine* 2001; **134**:657–662.
8. http://www.consort-statement.org/, accessed 3 December 2003.
9. Klein DF. Flawed meta-analyses comparing psychotherapy with pharmacotherapy. *American Journal of Psychiatry* 2000; **157**:1204–1211.
10. Quitkin FM, Rabkin JG, Gerald J, Davis JM, Klein DF. Validity of clinical trials of antidepressants. *American Journal of Psychiatry* 2000; **157**:327–337.
11. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates: Hillsdale, NJ, 1988.
12. Gould RA, Otto MW, Pollack MH. A meta-analysis of treatment outcome for panic disorder. *Clinical Psychology Review* 1995; **15**:819–844.
13. Joffe R, Sokolov S, Streiner D. Antidepressant treatment of depression: a meta-analysis. *Canadian Journal of Psychiatry* 1996; **41**:613–616.
14. Barlow DH, Gorman JM, Shear MK, Woods SW. Cognitive-behavioral therapy, imipramine, or their combination for panic disorder: a randomized controlled trial. *Journal of the American Medical Association* 2000; **283**: 2529–2536.
15. Elkin I, Shea MT, Watkins JT, Imber SD, Sotsky SM, Collins JF, Glass DR, Pilkonis PA, Leber WR, Docherty JP, Fiester SJ, Parloff MB. National Institute of Mental Health Treatment of Depression Collaborative Research Program: general effectiveness of treatments. *Archives of General Psychiatry* 1989; **46**:971–982.
16. Kozak MJ, Liebowitz MR, Foa EB. Cognitive behavior therapy and pharmacotherapy for obsessive-compulsive disorder: the NIMH-sponsored collaborative study. In *Obsessive-compulsive Disorder*: *Contemporary Issues in Treatment. Personality and Clinical Psychology Series*, Goodman WK, Rudorfer MV, Maser JD (eds). Lawrence Erlbaum Associates: Mahwah, NJ, 2000; 501–530.
17. DeGruttola V, Tu XM. Modeling progression of CD4-Lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**:1003–1014.
18. Laska EM, Klein DF, Lavori PW, Levine J, Robinson DS. Design issues for the clinical evaluation of psychotropic drugs. In *Clinical Evaluation of Psychotropic Drugs*: *Principles and Guidelines*, Prien RF, Robinson DF (eds). Raven Press: New York, NY, 1994; 29–67.
19. Leber P. The use of placebo control groups in the assessment of psychiatric drugs: an historical context. *Biological Psychiatry* 2000; **47**:699–706.
20. Klein DF. Cognitive therapy. Comment on: *British Journal of Psychiatry* 1994; **165**:126–130. Source: *British Journal of Psychiatry* 1994; 838.
21. Klein DF. Critiquing McNally's reply. *Behaviour Research and Therapy* 1996; **34**:859–863.
22. Klein DF. Control groups in pharmacotherapy and psychotherapy evaluations. *Prevention and Treatment*, 1, Article 1, posted 22 September 1997. Available at: http://journals.apa.org/prevention/volume1/97_a1.html. Accessed 3 December 2003.
23. Schatzberg AF, Kraemer HC. Use of placebo control groups in evaluating efficacy of treatment of unipolar major depression. *Biological Psychiatry* 2000; **47**:736–744.
24. Billingsley P. *Probability and Measures*. Wiley: New York, NY, 1986.

