 DOI: 10.1111/j.1541-0420.2010.01498.x

Assessing the Sensitivity of Meta-analysis to Selection Bias: A Multiple Imputation Approach

James Carpenter,[1,*] Gerta Rücker,[2] and Guido Schwarzer[2]

[1]Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, U.K.
[2]Institute of Medical Biometry and Medical Informatics, University Medical Centre, Freiburg, Germany
[*]*email:* james.carpenter@lshtm.ac.uk

Summary. Evidence synthesis, both qualitatively and quantitatively through meta-analysis, is central to the development of evidence-based medicine. Unfortunately, meta-analysis is often complicated by the suspicion that the available studies represent a biased subset of the evidence, possibly due to publication bias or other systematically different effects in small studies. A number of statistical methods have been proposed to address this, among which the trim-and-fill method and the Copas selection model are two of the most widely discussed. However, both methods have drawbacks: the trim-and-fill method is based on strong assumptions about the symmetry of the funnel plot; the Copas selection model is less accessible to systematic reviewers, and sometimes encounters estimation problems. In this article, we adopt a logistic selection model, and show how treatment effects can be rapidly estimated via multiple imputation. Specifically, we impute studies under a missing at random assumption, and then reweight to obtain estimates under nonrandom selection. Our proposal is computationally straightforward. It allows users to increase selection while monitoring the extent of remaining funnel plot asymmetry, and also visualize the results using the funnel plot. We illustrate our approach using a small meta-analysis of benign prostatic hyperplasia.

Key words: Meta-analysis; Multiple imputation; Publication bias; Random effects model; Sensitivity analysis.

1. Introduction

Quantitative evidence synthesis through meta-analysis plays a key role in both evidence-based medicine and in social science research. The standard methods rely on the assumption that the available evidence is unbiased, so that valid inference can be made for the relevant population. Unfortunately this is often not the case, and empirical evidence has been found for the preferential submission, publication, and citation of studies with positive and/or statistically significant results (e.g., Chan et al., 2004; Rothstein, Sutton, and Borenstein, 2005; Nieminen et al., 2007).

Such biases distort the evidence base for meta-analysis, potentially resulting in unduly optimistic inference for intervention effects. It is therefore important to detect and where appropriate correct for them. Commonly used figures for visualizing one of the most common types of bias, publication bias, are funnel and radial plots. Such figures plot measures of precision against treatment effect. For instance, Figure 1 plots the standard error against treatment effect. In the presence of publication bias, or other systematic differences between small and large studies resulting in a systematic selection of evidence, the "funnel" is asymmetric, because smaller studies (with larger standard errors) tend to report larger effects. A number of statistical tests for such funnel plot asymmetry have been proposed; for a review see Sterne et al. (2008).

Beyond that, there have been a number of methods proposed for correcting meta-analysis intervention estimates for publication bias when funnel plot asymmetry is present (see, e.g., Moreno et al., 2009). Two of the most widely discussed methods in the literature are "Trim-and-Fill" (Duval and Tweedie, 2000b) and the Copas selection model (Copas and Shi, 2000). However, although the latter can work well, it encounters numerical problems in a nontrivial proportion of cases (Carpenter, Schwarzer, and Rücker, 2009) and the implicit "missing studies" cannot be readily visualized. The latter problem is avoided with "Trim-and-Fill," which essentially imputes the missing studies to create a symmetric funnel plot. This, however, is a strong assumption that may not always be appropriate (Schwarzer, Carpenter, and Rücker, 2010).

In this article, we adopt an approach based around a logistic selection model. We describe how the model can be approximately estimated using a multiple imputation approach. This essentially imputes missing studies under a missing at random assumption, and then reweights the imputed data to allow for nonrandom selection. The approach is rapid and computationally robust, thus allowing rapid exploration of various degrees of selection while avoiding the numerical integration or the use of Markov chain Monte Carlo, which is usually required to fit such models. Further, we show that given a particular selection mechanism, the extent of any remaining funnel plot asymmetry can be tested. Also, the implicit "missing studies" can be shown on the funnel plot. Thus we avoid some of the issues with the Copas selection model (especially when the number of studies is small) while retaining some of the advantages of "Trim-and-Fill."

The plan for the article is as follows. Section 2 describes a small meta-analysis in benign prostatic hyperplasia, which motivates and illustrates the development. The Copas

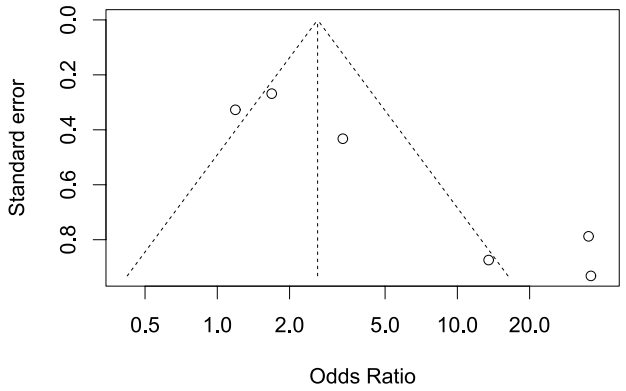

Figure 1. Funnel plot for the *serenoa repens* trials. Vertical broken line is odds ratio from fixed effect analysis.

Table 1
Summary data from studies contributing to the serenoa repens meta-analysis for men with benign prostatic hyperplasia (Wilt et al., 2002)

	Proportion of men with "very good" symptom reduction in	
Trial	*Serenoa repens* arm	Placebo arm
1	18/20	8/20
2	16/20	2/20
3	22/27	3/27
4	44/55	30/55
5	58/82	63/94
6	84/125	62/113

selection model and "Trim-and-Fill" are briefly reviewed in Section 3 before we describe our proposal in Section 4. We return to the example in Section 5 and conclude with a discussion in Section 6.

2. Motivating Example

Consider the data in Table 1, from a meta-analysis of *serenoa repens* versus placebo for men with benign prostatic hyperplasia (Wilt, Ishani, and MacDonald, 2002). The outcome is whether each participating man rates his symptom reduction "very good" or only "good." Figure 1 shows the "funnel plot" of odds ratios in favor of treatment against their standard errors, where the dotted lines show the usual fixed effects treatment estimate, ±2 standard errors.

The usual fixed and random effects meta-analysis treatment estimates are shown in Table 2. They both strongly favor *serenoa repens*, the random effects analysis markedly more so, as it puts more weight on the smaller studies when estimating the overall effect. The estimated between study heterogeneity is relatively large, $\hat{\tau}^2 = 1.44^2$.

Unfortunately Figure 1, together with the difference between the fixed and random effect analyses, suggests that the treatment benefit is systematically stronger in the smaller studies. This may be because smaller studies recruited patients who were more likely to respond to treatment. However, there is also a suspicion that one or more less favorable small studies remain unpublished.

Table 2
Odds ratios in favor of serenoa repens*, from various models*

Analysis	Odds ratio	95% CI	p-value
Fixed effects model	2.62	(1.90, 3.61)	<0.0001
Random effects model (fitted in `WinBUGS`)	5.58	(1.39, 22.5)	0.015
Sensitivity analysis, $\delta = 0.6$ (selection on log-odds ratio)	4.53	(1.11, 18.53)	0.04
Sensitivity analysis, $\delta = 0.6$ (selection on z-score)	4.09	(0.87, 19.26)	0.07
Trim-&-Fill (3 missing studies)	1.77	(0.62, 5.04)	0.28
Copas selection model (solution 1, 21 missing studies)	1.29	(0.42, 3.93)	0.66
Copas selection model (solution 2, 10 missing studies)	2.00	(1.14, 3.50)	0.02

There are a number of statistical tests to assess formally the extent of asymmetry in the funnel plot; applying one of the most reliable—the arcsin-Thompson test, which allows for heterogeneity (Rücker, Carpenter, and Schwarzer, 2008)—gives a p-value against the null hypothesis of no funnel plot asymmetry of 0.008. This suggests we should consider how robust the conclusions are to funnel plot asymmetry, and associated publication bias.

Thus we now briefly review two methods for estimating treatment effects in the presence of funnel plot asymmetry, make a new proposal, and describe an approximate estimation method via multiple imputation. We return to this example in Section 5.

3. Existing Methods for Correcting for Funnel Plot Asymmetry

3.1 *Trim-and-Fill*

Looking at Figure 1, the idea underlying the method is as follows. First, the number of studies in the asymmetric outlying part of the funnel plot (right side of Figure 1) is estimated using rank-based methods. These studies are then temporarily removed ("trimmed") from the right of Figure 1. Using the remaining studies, the estimated treatment effect, say $\hat{\mu}_{\text{symm}}$, is calculated. This is considered the true center of the funnel. A vertical line is then drawn at $\hat{\mu}_{\text{symm}}$ and each trimmed study reinstated in turn, together with its "missing" reflection about $\hat{\mu}_{\text{symm}}$. Finally, the resulting set of original and "reflected"/"filled" studies are used to estimate the treatment effect.

Three different methods have been proposed to estimate the number of missing studies (Duval and Tweedie, 2000a, 2000b). Two of these methods (L- and R-estimator) have been shown to perform better in simulations; we use the L-estimator here. Furthermore, we use a fixed effect model to determine $\hat{\mu}_{\text{symm}}$ and a random effects model for the meta-analysis of original and "filled" studies.

The trim-and-fill method is easy to implement and visualize, but the assumption that missing trials are "reflections" about $\hat{\mu}_{\text{symm}}$ is strong and often questionable.

3.2 *Copas Selection Model*

An alternative that avoids making the relatively strong assumptions underpinning the trim-and-fill method is the probit-selection model approach proposed by Copas and coauthors, e.g., Copas and Shi (2000) and references therein. We give a brief description.

Let $(z_j, \sqrt{v}_j)$ be the (log-odds ratio, standard error) from trial j, and $R_j > 0$ if trial j is published. The approach is based on the model:

$$\begin{aligned} z_j &= \mu + u_j + \sqrt{v}_j \epsilon_j \\ \text{probit}\,\Pr(R_j > 0) &= a + b/\sqrt{v}_j + \eta_j \\ u_j &\sim N(0, \tau^2) \text{ (heterogeneity)} \\ \begin{pmatrix} \epsilon_j \\ \eta_j \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \end{aligned}$$

The model is fitted over a user-defined range of (a, b), these giving different relationships between the chance of a trial's publication, $\sqrt{v}_j$ and, through the correlation ρ, z_j. The results show how $\hat{\mu}$, the random effects estimate of the overall log-odds ratio in favor of treatment, varies as the chances of small studies being published depends on their results.

This can work well, but many systematic reviewers are not trained statisticians, and so find interpreting the model difficult. Unlike the trim-and-fill method the results cannot be visualized on the funnel plot. Further, in around 20% of cases, there are awkward numerical problems in estimating the model (Carpenter et al., 2009). Nevertheless, our empirical evaluation indicated this selection model based approach is preferable to the "Trim-and-Fill" approach (Schwarzer et al., 2010).

4. New Proposal and Estimation via Multiple Imputation

Our proposal is based on a logistic selection model. We give details for binary data, which commonly arise in medical systematic reviews, but the approach could equally be used on continuous outcomes. Likewise, we describe the approach using selection on the log-odds ratio, but other selection mechanisms are possible, and we explore selection on the z-score in Section 5.

For study $j = 1, \ldots, J$, let r_{ij}, n_{ij} denote the number of events and patients, respectively, in arm $i = 1, 2$, where $i = 1$ corresponds to the control arm and $i = 2$ to the active arm. For simplicity, we assume that positive log-odds ratios favor the intervention, as in Table 2. The model is:

$$\begin{aligned} &\text{logit}\,\Pr(r_{ij} \text{ events given } n_{ij} \text{ at risk}) = \beta_{0j} + 1[i=2]\beta_{1j} \\ &\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right] \end{aligned} \quad (1)$$

$$\text{logit}\,\Pr(\text{observe trial } j) = \alpha + \delta \log(\text{odds ratio in trial } j). \quad (2)$$

Here, (1) is the usual random effects model (see Turner et al., 2000), with $1[i = 2]$ indicating the active intervention arm so that β_{1j} is the random effects log-odds ratio for the intervention effect. The second part (2) is the selection mechanism, where in this formulation δ is the log-odds relating observing trial j to the crude odds ratio in trial j.

If all trials are observed, then the log likelihood of models (1) and (2) are separate; the latter is degenerate and fitting the former gives the usual random effects estimate (Turner et al., 2000).

However, now suppose there are additional studies, $j = J + 1, \ldots, J + K$, whose events are missing but whose arm-specific denominators we denote using the above notation as n_{ij}. Now consider fitting the model to studies $j = 1, \ldots, J + K$. Because the last K have missing responses, the likelihood for models (1) and (2) does not separate, and must be jointly maximized. Usually, fitting this selection model would be done by numerical integration, or using Markov chain Monte Carlo with noninformative priors (e.g., Carpenter, Pocock, and Lamm, 2002). Notice that $\delta = 0$ corresponds to a particular missing at random mechanism for the trials; in this case, no matter how many studies we postulate to be missing, we get the same inferences for the parameters in (1). However, as δ increases from 0, the chance of observing a study increases with its log-odds ratio. Then, the more studies are missing, the more the random effects treatment estimate $\hat{\beta}_1$ is reduced.

Because there is often very little information in the model on δ, such models are often refitted over a range of δ values, allowing the sensitivity of inference to nonrandom selection to be explored. We return to this in Section 6.

4.1 *Fitting by Multiple Imputation*

Given (n_{1j}, n_{2j}) for the missing trials $j = J + 1, \ldots, J + K$, and assuming $\delta = 0$, so the missing data are missing at random, we can use multiple imputation (Rubin, 1987) $J + K$ studies.

We can fit model (1) to each of these M completed data sets, then use Rubin's rules to combine the results for final inference about the treatment effect β_1. However, if $\delta = 0$ there is no gain from doing this; as noted above inference for β_1 will be the same as for fitting (1) to the observed J studies.

However, as suggested by Carpenter, Kenward, and White (2007) we can reweight data imputed under a missing at random assumption (here events in each intervention arm in the missing studies) to allow estimation when data are missing not at random under model (2), i.e., when $\delta \neq 0$. Following imputation under a missing at random assumption, we can use this approach to rapidly estimate the treatment effect under nonrandom selection, thus exploring the effect of funnel plot asymmetry. Further, as we illustrate below, (i) as δ increases from 0 we can test for residual funnel plot asymmetry using any of the available tests and (ii) the results can be displayed using the funnel plot and accessibly presented for a nonstatistical audience.

To show how this proceeds, suppose there is one missing study, so that $K = 1$. Let $O_j = 1$ if study j is observed, and 0 otherwise. As above we have $m = 1, \ldots, M$ imputations; let the crude log-odds ratio from imputation m of study $j = J + 1$ be z_m. Suppose δ takes some nonzero value. We imputed data for study $J + 1$, denoted $R^m_{J+1} = (R^m_{1,J+1}, R^m_{2,J+1})$, where the first entry corresponds to control arm events and the second intervention arm events. This was imputed from

the distribution of the study data estimated from studies that are observed, $f(R \,|\, O_{J+1} = 1)$. However, as the study was unobserved and $\delta \neq 0$, we should have drawn from $f(R \,|\, O_{J+1} = 0)$. We therefore use importance sampling (see, e.g., Ripley, 1987) to reweight the imputed data. For imputation m we derive the weight in Web Appendix A as

$$w^m \propto e^{-\delta z^m_{J+1}}. \tag{3}$$

Thus, after normalization, the weights do not explicitly involve α, which is attractive as it does not need to feature in our sensitivity analyses.

Having calculated the weight w^m for each imputed "complete" meta-analysis data set of studies $j = 1, \ldots, J+1$, we can now apply Rubin's rules; but instead of weighting them all equally, we apply weights w^m.

Thus, if fitting (1) in turn to each imputed meta-analysis set gives $\hat{\beta}^m_1$, $m = 1, \ldots, M$ with variance $\hat{\sigma}^{2,m}_{\beta_1}$ the estimate of β_1 for the current value of δ is

$$\hat{\beta}_{1,\,MNAR} = \sum_{m=1}^{M} \hat{\beta}^m_1 w^m \Big/ \sum_{m=1}^{M} w^m. \tag{4}$$

The within imputation variance is $\hat{\sigma}^2_{W,\,MNAR} = \sum_{m=1}^{M} \hat{\sigma}^{2,m}_{\beta_1} w^m / \sum_{m=1}^{M} w^m$, and between imputation variance is $\hat{\sigma}^2_{B,\,MNAR} = \sum_{m=1}^{M} w^m (\hat{\beta}^m_1 - \hat{\beta}_{1,\,MNAR})^2 / \sum_{m=1}^{M} w^m$. Here, "MNAR" refers to "missing not at random"—that is, the estimates are those when $\delta \neq 0$, so that studies are selectively missing by (2).

Then, following Rubin's rules $\mathrm{Var}(\hat{\beta}_{1,\,MNAR}) \approx \hat{\sigma}^2_{W,\,MNAR} + \hat{\sigma}^2_{B,\,MNAR}$, where we omit the usual multiplier $(1 + 1/M)$ before the second term, because in applications we recommend choosing M large enough that it is irrelevant. Notice that if $\delta = 0$, then all the weights are equal, and (4), together with the variance formulae, revert to their usual forms.

The above argument has been developed when there was one missing study, i.e., $K = 1$. However, it extends to the case $K > 1$; in this case $w^m = \exp\{-\delta \sum_{j=J+1}^{J+K} z^m_j\}$. Note we could also apply this approach to other parameters in (1), such as the heterogeneity variance.

In summary, given the sizes of K missing trials, we fit a model to the observed data and impute the missing studies. Then, given δ, we can reweight to get an approximate estimate of the treatment effect under the implied nonrandom selection. In our experience we suggest using a large number of imputations, say $M = 1000$, as these can be generated in seconds and ensure that higher weights are distributed over a good number of imputations.

4.2 *Application to Publication Bias in Meta-analysis*

We now describe how to apply this in the context of publication bias in meta-analysis. Again, we first assume that there is $K = 1$ missing study. We proceed as follows:

(1) We need to choose a distribution from which to draw the missing study size from. This could draw upon information from other meta-analyses, or come from fitting a parametric model to the observed study sizes in the meta-analysis at hand. Here, we advocate sampling with replacement from the observed study sizes (keeping the pairing of intervention and control study size). We advocate choosing M large, say 1000. We then sample with replacement from observed study sizes $\{(n_{11}, n_{21}), \ldots, (n_{1J}, n_{2J})\}$ to obtain $m = 1, \ldots, M$ study sizes $(n^m_{1(J+1)}, n^m_{2(J+1)})$.

(2) Having fitted the usual random effects model (1) to the observed J studies, we impute the M "complete" studies as described in Section 5.

(3) We fit the usual random effects model to each of the imputed data sets, obtaining $\hat{\beta}^m_1$ and its variance.

(4) Using our current value of δ, we calculate the weights and in turn calculate $\hat{\beta}_{1,\,MNAR}$, and its standard error. Similarly we calculate, on each imputed data set, our chosen test statistic for funnel plot asymmetry and its standard error. Using the current weights we apply Rubin's rules to obtain the test statistic and p-value for remaining asymmetry. We can use the weights to calculate the expected number of events in the control and intervention arm. This gives the average position of the K "missing" studies on the funnel plot.

For each value of δ we then have (i) a confidence interval for the treatment effect and (ii) a p-value for residual funnel plot asymmetry. To guide interpretation, one may wish to increase δ till either (i) the intervention effect is no longer significant but funnel plot asymmetry remains or (ii) there is no longer evidence of funnel plot asymmetry, but the treatment effect remains significant.

Notice that the above steps work in the same way with $K > 1$ missing studies too. In this case sensitivity analyses involve two parameters: the number of missing studies, and the degree of selection δ. We could use these to create a three-dimensional plot exploring variation in treatment effect with number of missing studies and selection.

4.3 *Improving the Performance of the Method*

Our proposal is an application of importance sampling. For this to give reliable results, at a given level of selection, requires two conditions. First, the ratio of the density of the missing studies to the density of the observed studies must be bounded. In our setting, this bound gets larger as selection increases, and this can result in a few imputations taking almost all the weight. To address this, step 3 of the modified algorithm below has the effect of truncating the extreme tail of the distribution of the missing studies.

Second, we need a good model for the observed data. If this model has too heavy a tail in the direction of the missing studies, then as selection increases the method will be increasingly conservative. Here, as described above, we fit the usual random effects model to the observed data, and impute from this. However, we found the tail of this distribution was too heavy.

In Web Appendix B, we argue that the following algorithm gives an appropriate correction to the weights. Recall z_j is the log-odds ratio for observed study $j = 1, \ldots, J$ and let v_j be its variance. Let $\hat{\beta}_1$ be the random effects estimate, obtained by fitting (1) to the observed data, with estimated heterogeneity variance $\hat{\sigma}^2_1$. Form $\tilde{y}_j = z_j / \sqrt{\hat{\sigma}^2_1 + v_j}$ and $\tilde{x}_j = 1/\sqrt{\hat{\sigma}^2_1 + v_j}$. Regress $\tilde{y}_j$ on $\tilde{x}_j$ giving intercept $\tilde{\alpha}$ and slope $\tilde{\beta}$. As above, at each imputation $m = 1, \ldots, M$ suppose one study is imputed, with log-odds ratio z^m_{J+1}, and standard error v^m_{J+1}. Calculate the correction factor as

$$c^m = \exp\left\{\frac{-1}{2\left(\hat{\sigma}_1^2 + v^m\right)}\left[\left(z_{J+1}^m - \tilde{\beta} - \tilde{\alpha}\left(\hat{\sigma}_1^2 + v^m\right)\right)^2 - \left(z_{J+1}^M - \hat{\beta}_1\right)^2\right]\right\}. \quad (5)$$

If more than one study is imputed at each imputation m, the correction is calculated using the sum of the imputed log-odds ratios and the sum of their variances.

We use this correction as follows (assuming missing studies have lower odds ratios):

(1) Calculate the weights as described in Subsection 4.1;
(2) For each imputation for which the minimum of the log-odds ratios for the imputed studies is less than $\hat{\beta}_1 - 2\hat{\sigma}_1$ (i.e., the original random effects estimate minus twice the heterogeneity standard deviation), multiply the weight by (5);
(3) Normalize the adjusted weights, and remove any imputations with weight more than 10%;
(4) Renormalize the remaining weights, and apply them.

Web Appendix C presents the results of a simulation study evaluating this proposal.

5. Analysis of *Serenoa Repens* Data

Data are available from six studies (Table 1). Suppose one study is missing, so in the notation above, $K = 1$. Under (2) the model for observing studies $j = 1, \ldots, (J + K) = 6 + 1 = 7$ is:

$$\text{logit}\,\Pr(\text{observe trial } j) = \alpha + \delta \log(\text{odds ratio in favor of intervention in study } j).$$

Notice we are not constrained to have the log-odds ratio as the covariate; below we compare the results using the z-score.

To begin, we explore δ ranging between [0, 1.3]. Thus the odds ratio of observing a trial—when the trial itself reports an odds ratio of $e^1 = 2.7$ in favor of intervention—versus observing a trial that itself reports an odds ratio of 1 in favor of treatment, lies in $[1 = e^0,\ 4 \approx e^{1.3}]$.

We use `WinBUGS` (Spiegelhalter, Thomas, and Best, 1999) to fit the random effects meta-analysis model to the observed data and then impute $M = 1000$ data sets, each consisting of the six observed studies plus one imputed one. We then apply our proposal to explore the sensitivity of the results as we move δ away from zero. For selection on the log-odds ratio, we were able to fit a similar model in `WinBUGS` and compare the results.

The left panel of Figure 2 shows the results; for $\delta = 0$ the odds ratio is 5.58 ($p = 0.015$): this is the random effects meta-analysis estimate. Then, at each value of δ we want to explore, we can calculate not only the corresponding point estimate and its standard error, but also a test for remaining funnel plot asymmetry. We use the arcsin-Thompson test (Rücker et al., 2008). For $\delta > 0.6$ ($e^{0.6} = 1.8$) we find it just no longer significant at the 5% level. Thus with this level of selection and assuming one missing study, there is borderline evidence of asymmetry in the funnel plot. We note that the point estimates from the new method and `WinBUGS` are close, but the standard error from our method is slightly smaller.

The right panel of Figure 2 shows the results for selection on the z-score. For any given δ, this results in a larger correction.

Under our method, for each δ and for selection on the log-odds ratio and z-score, we can readily calculate the expected position of the missing study on the funnel plot. Figure 3 reveals why, as δ increases, selection on the z-score results in a greater reduction in the odds ratio. It also explains why when we select on the z-score (unlike for the log-odds ratio), as δ increases the test for funnel plot asymmetry initially declines in significance before becoming more significant.

Figure 3 also shows the trim-and-fill method results: its reflection assumption suggests three "missing trials" shown with "+"). We see that the method is likely to be conservative in this example; further, its implicit selection mechanism is quite different to the log-odds ratio or z-score. Lastly, Table 2 summarizes the above results and includes a Copas selection model analysis. Estimation was very tricky and two solutions were found corresponding to a probability of residual funnel plot asymmetry equal to 0.05.

6. Discussion

We have applied a multiple imputation based approach to sensitivity analysis for assessing the impact of funnel plot asymmetry in meta-analysis. Such asymmetry is often linked to publication bias and other small study effects.

The *serenoa repens* data show marked funnel plot asymmetry; our analysis shows only mild adjustment is sufficient for the results to be consistent with no benefit of the intervention. Further, our approach enabled us to explore and visualize the effect of selection on the log-odds ratio and the z-score. In this example, we find selection on the z-score, relative to the log-odds ratio, up-weights imputed studies with smaller variance (Figure 3). Further, the pattern of observed data led to an unexpected finding: while increasing selection on the log-odds ratio reduced funnel-plot asymmetry, increasing selection on the z-score increased it (after an initial decline). This unexpected difference confirms that caution is needed when using the extent of remaining funnel plot asymmetry as a guide to whether an analysis has sufficiently accounted for selection.

The above results illustrate the flexibility of our proposal, which is based on importance sampling. For this to work reliably, we need a model for the observed data that does not have too heavy a tail in the direction of the missing studies. We found the usual random effects model had too heavy a tail in this direction, and this led us to the proposal in Subsection 4.3. Using this, we have obtained results that are close to those from a similar model in `WinBUGS` (Figure 2) and perform acceptably in simulations (Web Appendix C). Nevertheless, our method is approximate and where selection is strong, it may struggle.

Our sensitivity analysis adjusts for publication bias on the premise it causes funnel plot asymmetry, results in the loss of K studies, and is described by our selection model. Of course, this selection model is only one of many options. If selection depended only on the study sample size, it would induce no bias; bias is induced when selection depends in some way on the response. Thus Hedges (1992) proposed a step-selection function based on p-values, but as Vevea and Woods (2005) point out, this requires meta-analysis with 100–200 studies. Jackson, Copas, and Sutton (2005) consider a meta-analysis of 77 studies where the outcome is death in the operating

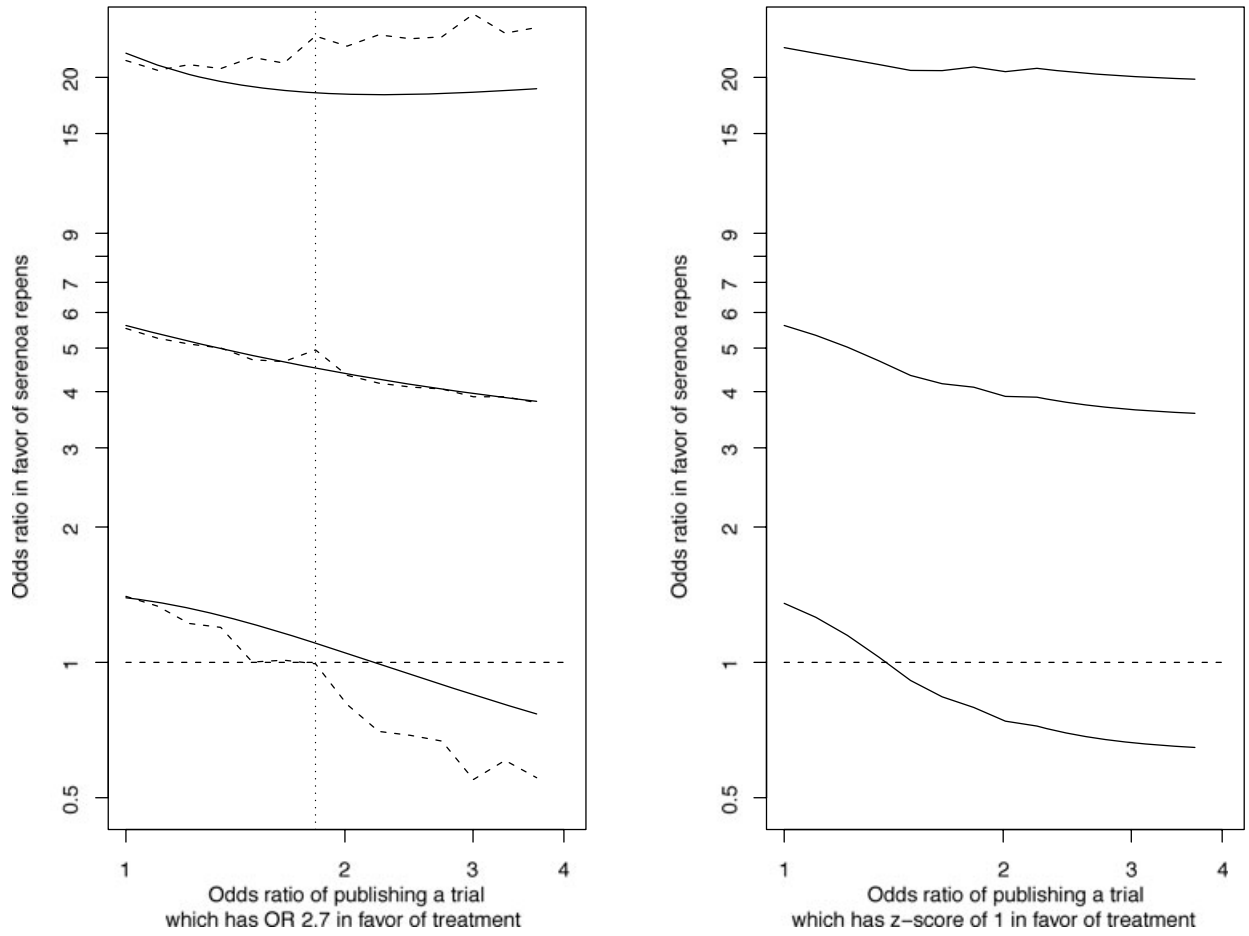

Figure 2. Change in estimated odds ratio as selection increases. Left panel: selection on the log-odds ratio; solid lines are estimated odds ratio and 95% confidence interval from new method; dashed lines are estimated odds ratio and 95% confidence interval obtained from `WinBUGS`. Vertical line corresponds to the value of δ at which the arcsin-Thompson test for publication bias is no longer significant at the 5% level. Right panel: selection on the z-score (results from new method only).

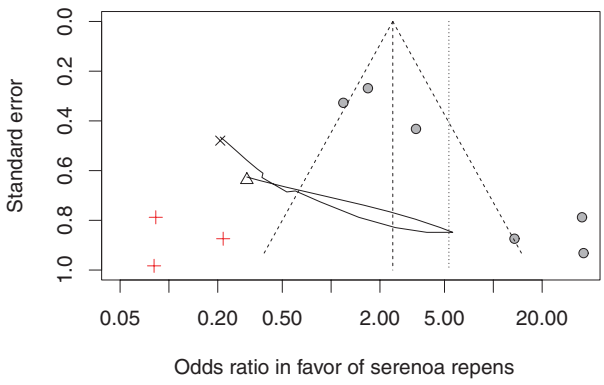

Figure 3. Funnel plot for the *serenoa repens* studies. Left vertical line is odds ratio from fixed effect analysis; right vertical line is odds ratio from random effects analysis. Line ending in Δ: selection on log-odds ratio—mean position of missing study as δ ranges from 0 to 1.2. Line ending in $\times$: selection on z-score—mean position of missing study as δ ranges from 0 to 1.2. Crosses: three additional studies proposed by the trim-and-fill method.

theater or death subsequently, and consider a selection model dependent on both these outcomes and time. Baker and Jackson (2006) propose a selection model relating to journal impact factors. All these approaches involve joint estimation of the selection model and the model of interest. Either they use a particular form of selection model (e.g., probit) so the likelihood can be calculated directly, or they require numerical/Monte Carlo integration. Unfortunately, as Copas and Shi (2000) note, there is often little information on the selection parameters in such models when—as is often the case—the number of studies in the meta-analysis is small.

Our approach is in the spirit of Vevea and Woods (2005), who propose prespecifying, rather than estimating, the weights in Hedges' selection model, and look at the sensitivity of the conclusions as they vary these weights. This is preferable for meta-analyses with small numbers of studies: indeed the reported median number of trials per meta-analysis in Cochrane reviews is around six (Mallett and Clarke, 2002). In addition, a key requirement for the practical use of these methods is that systematic reviewers, who are usually not trained statisticians, can understand the selection mechanism. Thus, although our method can be applied with more

sensitivity parameters, we favor focusing on a small number of parameters.

We conclude that local sensitivity analysis for funnel plot asymmetry using the method described here is very flexible, sidesteps computational issues that may arise with other approaches, and—perhaps most importantly in practice—readily allows the implicit modeling assumptions to be graphically presented.

7. Supplementary Materials

The Web Appendices, referenced in Sections 4 and 6 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

This work was funded by Deutsche Forschungsgemeinshaft grant FOR 534 Schw 821/2-2. JC is additionally funded by a U.K. ESRC fellowship, RES 063-27-0257. We thank the associate editor and referees for their stimulating comments.

References

Baker, R. and Jackson, D. (2006). Using journal impact factors to correct for the publication bias of medical studies. *Biometrics* **62,** 785–792.

Carpenter, J., Pocock, S., and Lamm, C. J. (2002). Coping with missing data in clinical trials: A model based approach applied to asthma trials. *Statistics in Medicine* **21,** 1043–1066.

Carpenter, J. R., Kenward, M. G., and White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random—a weighting approach. *Statistical Methods in Medical Research* **16,** 259–275.

Carpenter, J. R., Schwarzer, G., and Rücker, G. (2009). Empirical evaluation showed that the Copas selection model provided a useful summary in 80% of meta-analyses. *Journal of Clinical Epidemiology* **62,** 624–631.

Chan, A., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., and Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials. *Journal of the American Medical Association* **20,** 2457–2465.

Copas, J. B. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* **1,** 247–262.

Duval, S. and Tweedie, R. (2000a). A nonparametric "trim-and-fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* **95,** 89–98.

Duval, S. J. and Tweedie, R. L. (2000b). Trim-and-fill: A simple funnel-plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56,** 455–463.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science* **7,** 246–255.

Jackson, D., Copas, J., and Sutton, A. J. (2005). Modelling reporting bias: The operative mortality rate for ruptured abdominal aortic aneurysm repair. *Journal of the Royal Statistical Society, Series A* **168,** 737–752.

Mallett, S. and Clarke, M. (2002). The typical Cochrane review. How many trials? How many participants?*International Journal of Technology Assessment in Health Care* **18,** 820–831.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., and Cooper, N. J. (2009). Assessment of regression based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology* **9,** 2.

Nieminen, P., Rucker, G., Miettunen, J., Carpenter, J. R., and Schumacher, M. (2007). Statistically significant papers in psychiatry were cited more often than others. *Journal of Clinical Epidemiology* **60,** 939–946.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (eds). (2005). *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*. London: Wiley.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rücker, G., Carpenter, J. R., and Schwarzer, G. (2008). Arcsin tests for publication bias. *Statistics in Medicine* **27,** 746–763.

Schwarzer, G., Carpenter, J., and Rücker, G. (2010). Empirical evaluation suggests Copas selection model preferable to trim-and-fill method for selection bias in meta-analysis. *Journal of Clinical Epidemiology* **63,** 282–288.

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1999). *WinBUGS Version 1.2 User Manual*. Cambridge, U.K.: MRC Biostatistics Unit.

Sterne, J., Egger, M., Moher, D., Carpenter, J., Harbord, R., Higgins, J., Jones, D., Sutton, A., and Tetzlaff, J. (2008). Addressing reporting biases. In *Cochrane Handbook for Systematic Reviews of Interventions*, J. P. T. Higgins and S. Green (eds), 297–333. Chichester, U.K.: Wiley.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., and Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* **19,** 3417–3432.

Vevea, J. L. and Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a-priori weight functions. *Psychological Methods* **10,** 428–443.

Wilt, T., Ishani, A., and MacDonald, R. (2002). *Serenoa repens* for benign prostatic hyperplasia. *The Cochrane Database of Systematic Reviews* **3,** CD001423; doi:10.1002/14651858.CD001423.

Received December 2009. *Revised July* 2010.
Accepted July 2010.