Model-based sensitivity analysis for outcome reporting bias in the meta analysis of benefit and harm outcomes

John Copas[1], Anthony Marson[2], Paula Williamson[3], and Jamie Kirkham[3]

[1]Department of Statistics, University of Warwick, UK
[2]MRC North West Hub for Trials Methodology Research, Department of Molecular and Clinical Pharmacology, University of Liverpool, UK
[3]MRC North West Hub for Trials Methodology Research, Department of Biostatistics, University of Liverpool, UK

August 7, 2017

Abstract

Outcome reporting bias (ORB) occurs when outcomes in research studies are selectively reported, the selection being influenced by the study results. For benefit outcomes, Copas et al. (*Biostatistics* 2014) shows how risk assessments using the ORBIT risk classification scale can be used to calculate bias-adjusted treatment effect estimates. This paper presents a new and simpler version of the benefits method, and shows how it can be extended to cover the partial reporting and non-reporting of harm outcomes. Our motivating example is a Cochrane systematic review of 12 studies of Topiramate add-on therapy for drug-resistant partial epilepsy (Pulman et al. 2014). Bias adjustments for partially reported or unreported outcomes suggest that the review has overestimated the benefits and underestimated the harms of the test treatment.

Outcome reporting bias; Meta analysis; Selective reporting; ORBIT classification.

1 Introduction

Systematic reviews of clinical trials aim to include all relevant studies conducted on a particular topic and to provide an unbiased summary of their results, producing the best evidence on the benefits and harms of medical treatments. Recommendations made by the National Institute for Health and Clinical Excellence (NICE), who issue guidance on the use of treatments and procedures within the NHS, are based largely on systematic reviews [1]. Outcome reporting bias (ORB) has been defined as the selection, on the basis of the results, of a subset of the original outcomes recorded for inclusion in publication of trials [2]. A recent updated review of empirical research provides strong evidence that outcomes that are statistically significant have higher odds of being fully reported (range of odds ratios: 2.2 to 4.7) [3]. Much of the work to date on ORB has been limited in relation to evidence surrounding reporting bias for benefit outcomes with little attention being paid to harm outcomes. Important harm outcomes may also be subject to outcome reporting bias when trialists prefer to focus on the positive benefits of an experimental intervention [4]. In other words, for harm outcomes, bias may occur if specific harms had been reported, but the data were presented or suppressed in a way that would mask the harm profile of particular interventions.

The ORBIT (Outcome Reporting Bias in Trials) study described the prevalence and impact of outcome reporting bias in systematic reviews for benefit outcomes [5], while the more recent ORBIT II study has assessed whether selective reporting is likely to have occurred in harm outcomes [4]. Across the two ORBIT studies, a nine point classification system (reproduced here as Table 1) was developed to assess the risk of ORB within study reports for benefit outcomes while a complementary 13 point classification system (reproduced here as Table 2) was developed to assess the risk in harm outcomes.

In the presence of outcome reporting bias within a review, sensitivity analyses have been proposed to assess the impact of outcome reporting bias on an individual review [6]. However, these methods are limited to benefit outcomes, where bias can arise if outcomes which are measured but fail to

show a significant treatment effect are less likely to be reported. In this article we show how the model-based sensitivity approach in [6] can be adapted to adjust for ORB in harm outcomes where the resultant bias may not be related to the significance of the outcome data. We apply our method to a Cochrane systematic review of Topiramate add-on therapy for drug-resistant partial epilepsy.

2 Motivating example

Missing outcome data

The Cochrane systematic review *Topiramate add-on for drug-resistant partial epilepsy* [7] evaluates the efficacy and tolerability of topiramate when used as an add-on treatment for people with drug resistant partial epilepsy. The review considered two benefit outcomes and 12 harm outcomes. There were 11 studies included in the review, whilst one study was excluded due to there being no relevant outcome data. All 12 studies were considered in this assessment. The outcome matrix in Table 3 shows which outcomes were reported for each trial, and which outcomes were missing or partially reported. Outcomes are considered to be fully reported if sufficient information is provided such that the outcome could be included in the review meta analysis. Partially reported outcomes are those that are inadequately reported for inclusion in the review meta analysis (for example, an effect size was presented with no measure of precision or exact p-value). From the outcome matrix it is clear that the primary benefit outcome (50% reduction in seizure frequency) and harm outcome (treatment withdrawal) were reported in all included studies. In the excluded study these outcomes may have been measured but not reported. The other main benefit outcome (seizure freedom) was only reported in five of the eligible studies. Of the remaining eleven harm outcomes, the number of studies reporting data ranged from three (speech difficulty) to nine (fatigue).

Outcome reporting bias assessment

Three of the authors (JJK, AM, PRW - two statisticians and one clinician with expertise in neurology) gave a classification independently for each eligible study that did not report or partially reported on the review benefit and harm outcomes in accordance with the classification system presented in Tables 1 and 2. The agreed classifications are included in the outcome matrix (Table 3). Justifications for each of the classifications can be found in Sup-

plementary Table 2. In summary, all trials either not reporting or partially reporting on a review benefit outcome were awarded High Risk (classifications A, D, E, and G in the ORBIT terminology), except for the outcome seizure freedom for the Rosenfeld study, which was classified as Low Risk (classifications B, C, F and H). The majority of trials not reporting on the review harm outcomes were also classified as High Risk (classifications P1, P2, R1, R2, R3, S1 and S2 in the ORBIT terminology) apart from five of the harm outcomes in both the Yen and Zhang studies which were classified as Low Risk (classifications P3, T1, T2 and U). The trial report for the Previtera study specifically mentioned that the outcome weight loss was not measured (classification V), and hence is classified as No Risk in the final column of Table 2. As explained in Section 3, the method developed in this paper makes no distinction between No Risk and Low risk, and so for the purposes of this analysis No Risk is simply coded as Low Risk.

3 Sensitivity analysis for benefits and harms

Our former paper [6] showed how ORBIT risk classifications for benefit outcomes can form the basis for a general likelihood-based sensitivity analysis. In this section we present a simpler version which is easier to use in practice, and then show how the method can be extended to include harm outcomes.

3.1 Unadjusted estimates

The standard fixed effects model for outcome y_i is

$$y_i \sim N(\theta, \sigma_i^2) \ , \tag{1}$$

where θ is the (assumed constant) treatment effect and σ_i is the within-study standard deviation. If we ignore the studies where the outcome of interest is unreported, then the standard (unadjusted) fixed effects estimate of the treatment effect is

$$\hat{\theta}_U = \frac{\sum_{Rep} w_i y_i}{\sum_{Rep} w_i} \ , \tag{2}$$

where the suffix *Rep* on the summation signs means adding over all those studies for which the value of y_i is reported, and the weights w_i are

$$w_i = \sigma_i^{-2}.$$

The corresponding (unadjusted) $100(1-\alpha)\%$ confidence limits are

$$\{\hat{\theta}_U - z_\alpha s_\theta,\ \hat{\theta}_U + z_\alpha s_\theta\}\ , \tag{3}$$

where s_θ is the standard deviation of $\hat{\theta}_U$ given by

$$s_\theta = (\sum_{Rep} w_i)^{-\frac{1}{2}}\ ,$$

and $z_\alpha = \Phi^{-1}(1-\alpha/2)$ is the appropriate standard normal percentage point (e.g. $z_{0.95} = 1.96$). The usual (unadjusted) p-value for the null hypothesis $H_0 : \theta = 0$ is

$$p_U = P\left\{\chi^2 \geq \left(\frac{\hat{\theta}_U}{s_\theta}\right)^2\right\}\ , \tag{4}$$

where χ^2 denotes the chi-squared distribution on one degree of freedom.

These are standard calculations, but before showing how these can be extended to bias-adjusted estimates it is worth pointing out that (2), (3) and (4) can also be calculated through a likelihood-based approach. From the fixed effects model (1), the (unadjusted) log-likelihood function for θ (the sum of the log probabilities of what has been observed) simplifies to (omitting irrelevant additive constants)

$$-\frac{1}{2}\sum_{Rep}\frac{(y_i-\theta)^2}{\sigma_i^2}\ ,$$

which in turn simplifies to

$$L_U(\theta) = -\frac{(\theta-\hat{\theta}_U)^2}{2s_\theta^2}\ . \tag{5}$$

The unadjusted estimate $\hat{\theta}_U$ in (2) is the maximum likelihood estimate (MLE) in the sense that it is the value of θ which maximizes $L_U(\theta)$. The unadjusted confidence interval (3) can also be obtained directly from $L_U(\theta)$ by using a standard asymptotic result in the theory of likelihood functions, that the distribution of twice the log likelihood ratio for any single value of θ is (approximately) χ^2 on one degree of freedom. This means that the confidence limits in (3) are the two values of θ which satisfy the equation

$$L_U(\theta) = L_U(\hat{\theta}_U) - \frac{1}{2}z_\alpha^2\ . \tag{6}$$

Equivalently, they are the two values of θ which maximize

$$-\{L_U(\hat{\theta}_U) - L_U(\theta) + \frac{1}{2}z_\alpha^2\}^2 \ . \tag{7}$$

The unadjusted p-value (4) is

$$p_U = \mathrm{P}[\chi^2 \geq 2\{L_U(\hat{\theta}_U) - L_U(0)\}] \ . \tag{8}$$

3.2 ORB-adjusted estimates for benefit outcomes

For each study with at least one unreported or partially reported benefit outcome, we assume that either (i) the outcome was measured but failed to show a significant treatment effect, or (ii) the outcome was not measured or was unreported for reasons unconnected with the study results. Case (i) will lead to a bias in the overall treatment effect estimate, but not in case (ii). If case (i) is known to be true, the bias will be in the direction away from the null, resulting in over-estimation of the treatment effect and exaggeration of significance (p-value too small). Assuming that the ORBIT risk assessment correctly identifies case (i) as High Risk (Table 1, last column), explicit bias adjustments can be obtained by adding extra terms onto the likelihood function (5) which take account of those studies with missing or partially reported outcomes and High Risk ORBIT assessments.

From (1), the probability that the outcome y_i is not significant (at the α level) is

$$\mathrm{P}(-z_\alpha\sigma_i < y_i < z_\alpha\sigma_i) = \Phi\left\{z_\alpha - \frac{\theta}{\sigma_i}\right\} - \Phi\left\{-z_\alpha - \frac{\theta}{\sigma_i}\right\} \ , \tag{9}$$

where Φ is the cumulative distribution function of the standard normal distribution. So if it is assumed that a High Risk ORBIT assessment correctly identifies those studies which fall under case (i), then we can get a modified likelihood function by adding the log of (9) onto $L_U(\theta)$ for each study in which the benefit outcome is assigned a High Risk ORBIT classification. The resulting log likelihood function for a benefit outcome is therefore

$$L_B(\theta) = -\frac{(\theta - \hat{\theta}_U)^2}{2s_\theta^2} + \sum_{High} \log\left[\Phi\left\{z_\alpha - \frac{\theta}{\sigma_i}\right\} - \Phi\left\{-z_\alpha - \frac{\theta}{\sigma_i}\right\}\right] \ . \tag{10}$$

The bias-adjusted estimate $\hat{\theta}_B$, the corresponding confidence interval and p-value, can be obtained in exactly the same way as in Section 3.1 by using the

log likelihood $L_B(\theta)$ instead of $L_U(\theta)$. So $\hat{\theta}_B$ is the value of θ which maximizes (10), maximizing the function analogous to (7) gives bias-adjusted confidence limits, and the analogue of (8) gives the corresponding bias-adjusted p-value. The second term in (10) adds more weight to values of θ close to zero, and so, as expected, the bias-adjusted MLE from $L_B(\theta)$ is always closer to the null than the MLE from $L_U(\theta)$.

Comparing these adjusted estimates and confidence intervals for benefit outcomes with the corresponding unadjusted results in Section 3.1 gives a relatively simple sensitivity analysis for assessing the robustness of conclusions to outcome reporting bias. This has a different focus than the more elaborate method proposed in [6], which is designed to show the sensitivity of bias corrections for benefit outcomes on the reliability (ρ_1) and specificity (ρ_2) of the ORB assessments. The adjustments proposed here take exactly the same form as the unadjusted results and so the sensitivity analysis for different benefit outcomes can be easily summarized in a table, whereas the results of applying the method in [6] lead to a separate contour plot for each benefit outcome showing how the bias corrections depends on the values of ρ_1 and ρ_2. The two methods are formally equivalent when $\rho_1 = \rho_2 = 1$, described in section 3.4 of [6] as the 'special case of correct risk assessments'. A further advantage of the current approach is that the method extends directly to harm outcomes as shown in the next section.

3.3 ORB adjusted estimates for harm outcomes

As discussed in Section 1, ORBIT risk classifications for harm outcomes focus on the possibility that researchers may concentrate on the positive benefits of the experimental treatment and so be more likely to suppress those harms outcomes which cast the new treatment in an unfavourable light. If we assume that a positive value of the treatment effect for a harm outcome indicates an increase in the incidence of a particular side effect, then the corresponding ORB-adjustment method is very similar to the benefits case except that we now identify a High Risk ORBIT assessment with those unreported studies for which $y_i > 0$ (treatment is actually harmful). These are the studies shown as High Risk in the final column of Table 2.

From (1),

$$\mathrm{P}(y_i > 0) = \Phi\left(\frac{\theta}{\sigma_i}\right) \ , \tag{11}$$

and so the overall likelihood for a harm outcome is

$$L_H(\theta) = -\frac{(\theta - \hat{\theta}_U)^2}{2s_\theta^2} + \sum_{High} \log\left[\Phi\left(\frac{\theta}{\sigma_i}\right)\right] \ . \tag{12}$$

The ORB adjustment for a harm outcome follows from $L_H(\theta)$ in exactly the same way as before: the adjusted estimate $\hat{\theta}_H$ maximizes $L_H(\theta)$, and by replacing $L_U(\theta)$ with $L_H(\theta)$ in (7) and (8) we get the corresponding confidence limits and p-value. The second term in (12) adds more weight to positive values of θ, and so the bias-adjusted MLE from $L_H(\theta)$ is usually larger and further away from the null than the unadjusted MLE from $L_U(\theta)$.

3.4 Calculating the bias adjustments for benefits and harms

To fit the likelihood-based models of sections 3.2 and 3.3 we need the values of y_i and σ_i^2 for each of the studies reporting the outcome of interest, and estimates of σ_i^2 for each of the studies not reporting the outcome and classified as High Risk. Data for the reporting studies can be obtained directly from the individual study results, but the values of σ_i^2 for unreported outcomes will not usually be available. We follow [6] by suggesting how these can be imputed from data on the study sample sizes.

The quantities y_i and σ_i^2 for the studies reporting each outcome depend on the statistical properties of the outcomes being measured. The most common case, as in the example of Section 2, is when the studies are comparing the incidence of a binary response, observing a_i responses out of n_{1i} patients in the treatment group and c_i responses out of n_{2i} patients in the control group. The raw data for the ith study are the four frequencies (a_i, b_i, c_i, d_i), where $b_i = n_{1i} - a_i$ and $d_i = n_{2i} - c_i$. Taking θ to be the log relative risk, y_i and σ_i^2 are

$$y_i = \log\left\{\frac{a_i n_{2i}}{c_i n_{1i}}\right\} \ , \tag{13}$$

and

$$\sigma_i^2 = \frac{n_{1i} - a_i}{n_{1i} a_i} + \frac{n_{2i} - c_i}{n_{2i} c_i} \ . \tag{14}$$

When a study observes one or more zero cells, we follow [8] by adding 0.5 to all of the cells for that study, except when $a_i = c_i = 0$ or $b_i = d_i = 0$ in which case the study is omitted altogether (this occurred only twice amongst the

12×14 outcome calculations in the epilepsy example). We have used the zero-cell convention often used in Cochrane reviews, but a number of other conventions have also been suggested.

If the total sample sizes $n_i = n_{1i} + n_{2i}$ are available (or estimated from the study report) for all of the studies being considered, then it is possible to impute the values of σ_i^2 even for studies where the outcome of interest is unreported. For most statistical problems, variances of estimates are, at least approximately, inversely proportional to the sample size. This suggests that

$$\sigma_i^{-2} \approx k n_i$$

for some proportionality factor k. Following [6], the value of k can be estimated from the data on the reporting studies by

$$\hat{k} = \frac{\sum_{Rep} \sigma_i^{-2}}{\sum_{Rep} n_i} \ . \tag{15}$$

The value of σ_i^2 for a High Risk study with sample size n_i can then be imputed as

$$\sigma_i^2 \approx \frac{1}{\hat{k} n_i} \ . \tag{16}$$

3.5 ORB adjustments for the Topiramate example

Details of the Topiramate example have been set out in Section 2 above. The meta analysis considered 12 studies and 14 outcomes (two benefit outcomes and 12 harm outcomes). Table 3 shows which of these studies reports data on these outcomes, and, for the studies with missing data, which of these are rated as High Risk by the ORBIT assessment. The data take the form discussed in Section 3.4, each reporting study measuring the incidence of the appropriate binary response for the treatment and control groups. The Cochrane review in [7] reports a detailed (unadjusted) meta analysis of these studies. The raw data file which contains the sample sizes for each included study and the number of events for each intervention and each review outcome considered, can be found in Supplementary Table 1.

To implement the methods of Sections 3.1-3.3, the log relative risks y_i and standard errors σ_i for the reporting studies are calculated as in (13) and (14). Values of σ_i for the missing studies are imputed using the formula in (16). Table 4 shows the resulting unadjusted and ORB-adjusted relative risk

estimates and confidence intervals for these data. Note that the figures shown in the table are relative risks and not log relative risks, so a typical estimate $\hat{\theta}$ is displayed as $\exp(\hat{\theta})$. For the value of α, we have followed the Cochrane review by calculating 95% confidence intervals for benefit outcomes and 99% confidence intervals for the harm outcomes. As expected, the ORB-adjusted benefit treatment effects and confidence limits are smaller (closer to the null) than the corresponding unadjusted values, whilst the harm estimates are all larger (further from the null). Evidently, ignoring outcome reporting bias tends to exaggerate the benefits of the treatment and underestimate the severity of its side-effects.

There was little shift in the bias adjusted estimates for the review primary benefit outcome (50% reduction in seizure frequency) because there was only one study that was suspected of High Risk ORB for this outcome. When considering seizure freedom (benefit outcome), where nearly half the included studies did not report any useable data on this outcome, the bias adjusted result suggested that the original unadjusted estimate was overestimating treatment effect by more than 20%. Of the 12 harm outcomes, two with a non-statistically significant result became significant after adjustment, and for five of the harm outcomes (42%) the harmful effect of treatment was underestimated by 10% or more.

3.6 Simulating the example

This section reports the results of a simulation study designed to assess the statistical properties of these methods, with a particular focus on study characteristics similar to those seen in the example of Section 3.5. This is essentially a parametric bootstrap — we fit a model to the data of interest, take the fitted model as if it was the true model, and then simulate a large number of times to assess the sampling properties of the resulting estimates and confidence intervals. The results are summarized in Table 5.

The parameters for the simulation are provided by the analysis of section 3.5. For each of the 14 outcomes, the true value of the log relative risk θ is taken to be the adjusted maximum likelihood estimate taken from Table 4, and the study variances σ_i^2 are taken to be the estimates calculated directly from the study data or imputed using (16). This allows us to simulate values of y_i from (1). Making the ORB assumptions set out at the start of Section 3.2, we can then simulate the incidence of High risk and Low risk missing studies in such a way that the expected numbers of Reported (R), High

risk (H) and Low risk (L) studies match the numbers observed in the data (the last three columns of Table 5). For each outcome, the Low risk studies are randomly chosen from the complete set of 12 studies, and the High risk studies are chosen at random from amongst those remaining studies which have a non-significant value of y_i (for a benefit outcome) or a positive value of y_i (for a harm outcome). The probabilities used in selecting these missing studies are taken as the corresponding data proportions. For each simulation, we can then calculate the unadjusted and adjusted estimates and confidence intervals as in Sections 3.1-3.3. By including all 14 outcomes the simulations cover a wide range of different rates of missing outcomes.

Table 5 shows the simulation estimates of three performance measures: bias $E(\hat{\theta}-\theta)$, mean squared error $E(\hat{\theta}-\theta)^2$, and CI risk defined as $P(\theta \notin CI)$, or one minus the coverage of the confidence interval CI. The simulation sample size of 10,000 ensures that the simulation standard error in these figures is at most two units in the third decimal place. The bias of the unadjusted estimates reflects outcome reporting bias: as expected the bias is positive for the benefit outcomes and negative for the harm outcomes. The bias of the adjusted estimates reflects asymmetry in the likelihood functions (10) and (12), particularly noticeable if the number of high risk studies is relatively large compared to the number of reported studies, as is the case for some of the harm outcomes lower down in the table. However, for most of the outcomes the adjusted estimates are, on average, closer to the true value of θ than the unadjusted estimates, and in all cases the bias of the adjusted estimate is more than offset by the improvement in variance. Mean squared error (m.s.e. = variance + bias^2) is usually taken as the best overall measure of estimation accuracy: as seen in the table the adjusted estimates have lower m.s.e. than the unadjusted estimates for all 14 of the outcomes. The quality of a confidence interval CI is usually measured by its coverage, the actual probability $P(\theta \in CI)$ that the true value of the parameter belongs to CI, and by how much this differs from its nominal value, taken here to be 95% for benefit outcomes and 99% for harm outcomes. Simulation estimates of $P(\theta \notin CI) = 1 - P(\theta \in CI)$ are shown in the pair of columns headed 'CI risk'. In all but one of the outcomes, CI risk for the adjusted estimate is smaller than for the unadjusted estimate, and the single difference the other way round may just be the result of simulation error.

4 Mantel-Haenszel estimates

4.1 ORB adjustments for Mantel-Haenszel estimates

Section 3 is based on maximum likelihood methods for estimating the treatment effect parameter θ and formulating its ORB adjustments. Amongst alternative approaches, Mantel-Haenszel methods [9] are sometimes used in Cochrane reviews, including the published review of the Topiramate data in [7]. This section suggests how the maximum likelihood approach can be extended to cover Mantel-Haenszel (and other) estimates. The main change is a reformulation of the definitions of $\hat{\theta}_U$ and σ^2_θ given in Section 3.1. Section 4.2 illustrates Mantel-Haenszel ORB adjustments by a second re-analysis of the Topiramate data.

If the missing outcomes are ignored, the Mantel-Haenszel estimate of the unadjusted log relative risk is

$$\tilde{\theta}_U = \log\left(\frac{\sum_{Rep}\tilde{w}_i \exp(y_i)}{\sum_{Rep}\tilde{w}_i}\right) \ , \tag{17}$$

where the weights are now

$$\tilde{w}_i = \frac{c_i n_{1i}}{n_i} \ .$$

The term $\exp(y_i)$ in (17) is the relative risk for the ith study (formula (13) without the log), and so the main difference between (17) and (2) is the position of the log transformation: $\tilde{\theta}_U$ is the log of a weighted average of the relative risks, whereas $\hat{\theta}_U$ is a weighted average of the logs of the relative risks. This difference can be noticeable if study sample sizes are small especially if some of the observed frequencies are relatively close to zero, but for large sample sizes there is usually little difference between them.

Following the notation in [8], the variance of $\tilde{\theta}_U$ is

$$\tilde{s}^2_\theta = \frac{P}{RS} \ ,$$

where

$$P = \frac{n_{1i}n_{2i}(a_i + c_i)) - a_i c_i n_i}{n_i^2} \ , \ R = \frac{a_i n_{2i}}{n_i} \ , \ S = \frac{c_i n_{1i}}{n_i} \ .$$

The corresponding $100(1-\alpha)\%$ Mantel-Haenszel confidence interval for θ is

$$\{\tilde{\theta}_U - z_\alpha \tilde{s}_\theta, \ \tilde{\theta}_U + z_\alpha \tilde{s}_\theta\} \ . \tag{18}$$

As before, (17) and (18), as well as its associated p-value

$$\tilde{p}_U = P\left\{\chi^2 \geq \left(\frac{\tilde{\theta}_U}{\tilde{s}_\theta}\right)^2\right\} ,$$

can be calculated using the same likelihood approach as set out in Section 3.1, but now applied to the Maentel-Haenszel version of the unadjusted (pseudo-) likelihood

$$\tilde{L}_U(\theta) = -\frac{(\theta - \tilde{\theta}_U)^2}{2\tilde{s}_\theta^2} .$$

The contribution of a High Risk study to the likelihood, the log of either (9) for a benefit outcome or (11) for a harm outcome, depends only on the model (1) and so is exactly the same as before. So the Mantel-Haenszel version of the benefits likelihood is

$$\tilde{L}_B(\theta) = -\frac{(\theta - \tilde{\theta}_U)^2}{2\tilde{s}_\theta^2} + \sum_{High} \log\left[\Phi\left\{z_\alpha - \frac{\theta}{\sigma_i}\right\} - \Phi\left\{-z_\alpha - \frac{\theta}{\sigma_i}\right\}\right] ,$$

and similarly

$$\tilde{L}_H(\theta) = -\frac{(\theta - \tilde{\theta}_U)^2}{2\tilde{s}_\theta^2} + \sum_{High} \log\left[\Phi\left(\frac{\theta}{\sigma_i}\right)\right]$$

for a harm outcome. Maximizing these likelihoods gives the corresponding ORB adjustment to the Mantel-Haenzsel estimate. The associated confidence intervals and p-values also follow from the likelihoods as before.

4.2 Mantel-Haenzsel ORB adjustments for the Topiramate meta analysis

Table 4, giving the results of applying the maximum likelihood methods of Section 3 to the Topiramate example, is not exactly comparable to the Cochrane review in [7] which was based on Mantel-Haenzsel methods. Table 6 shows the corresponding results using the theory of Section 4.1. As expected, the unadjusted estimates in Table 6 now agree exactly with the published review estimates in [7]. The Mantel-Haenzsel unadjusted relative risks are all larger then the maximum likelihood estimates, although the differences

are usually small with the possible exception of three of the harm outcomes where the raw data show a relatively high incidence of zero cells in the control group. The sizes of the bias adjustments shown in Table 6 are very similar to the corresponding bias adjustments in Table 4, reflecting the fact that the two versions of the likelihoods differ only in their unadjusted estimation components, with exactly the same likelihood contributions from the High Risk studies.

5 Random effects model

All of the methods developed so far have been based on the fixed effects model (1). The model assumes homogeneity across the studies, in the sense that all the treatment effects y_i are estimating the same common parameter θ. If the problem of ORB is ignored, then this would mean that the variation between the observed outcomes y_i would be adequately explained by the within-study variances σ_i^2. The usual homogeneity test statistic for this is

$$Q = \sum_{Rep} w_i (y_i - \hat{\theta}_U)^2 \ . \tag{19}$$

Under the fixed effects model, the value of Q would then have a χ^2 distribution on $(n_{Rep} - 1)$ degrees of freedom, where n_{Rep} is the number of studies reporting the outcome of interest. This can be tested for each of the outcomes being considered. In the Topimarate meta analysis, for example, all of the 14 values of Q turn out to be less than the 95th percentile of the appropriate χ^2 distribution, suggesting that the fixed effects model is probably a reasonable assumption in this case. However, some caution is needed when interpreting the Q test in the presence of ORB. Arguably, if θ is reasonably small then for benefit outcomes the presence of ORB is likely to mask some of the values of y_i close to the null, which may increase the observed scatter amongst the reporting studies. The Q test (19) would then tend to be conservative. It is not clear whether this would also hold for harm outcomes.

If the scatter of the observed y_is is substantial, or if heterogeneity is otherwise anticipated from the context of the studies, then the random effects model

$$y_i \sim N(\theta, \sigma_i^2 + \tau^2)$$

is usually the preferred alternative. There are now two unknown parameters, θ as before, and τ^2, the random effects variance. The likelihood functions in

Section 3 generalize to

$$L_U(\theta, \tau^2) = -\frac{1}{2} \sum_{Rep} \left\{ \log(\sigma_i^2 + \tau^2) + \frac{(y_i - \theta)^2}{\sigma_i^2 + \tau^2} \right\} ,$$

$$L_B(\theta, \tau^2) = L_U(\theta, \tau^2) + \sum_{High} \log \left[\Phi \left\{ \frac{z_\alpha \sigma_i - \theta}{\sqrt{(\sigma_i^2 + \tau^2)}} \right\} - \Phi \left\{ \frac{-z_\alpha \sigma_i - \theta}{\sqrt{(\sigma_i^2 + \tau^2)}} \right\} \right] ,$$

and

$$L_H(\theta, \tau^2) = L_U(\theta, \tau^2) + \sum_{High} \log \left\{ \Phi \left(\frac{\theta}{\sqrt{(\sigma_i^2 + \tau^2)}} \right) \right\} .$$

To estimate θ by maximum likelihood we now have to maximize these likelihood functions over both parameters. Equivalently, we can do this in two stages. First, maximize out τ^2 for each individual value of θ to give the profile likelihood function for θ. For example, the unadjusted fixed effects likelihood $L_U(\theta)$ becomes

$$L_U^{(P)}(\theta) = \max_{\tau^2} L_U(\theta, \tau^2) ,$$

and similarly for the bias adjusted profile likelihoods $L_B^{(P)}(\theta)$ and $L_H^{(P)}(\theta)$. The second stage is to find the values of θ which maximize these profile likelihoods. The second stage also gives the asymptotic confidence limits for θ. After replacing $L_U(\theta)$ in (6), (7) and (8) by the appropriate profile likelihood, we can follow exactly the same procedure as set out before in Section 3. This likelihood approach to random effects models extends the method of Hardy and Thompson in [10].

6 Discussion

The ORBIT methodology is gaining wide acceptance, particularly amongst the Cochrane community of systematic reviewers, as the accepted approach for studying ORB by re-examining studies (i) that might have been eligible for inclusion in a review or meta analysis, but were subsequently not included because they did not fully report on particular review outcomes of interest, and (ii) that were included in relation to some, but not all, of the review outcomes of interest. If outcome data are unavailable from trial authors then review authors should assess the potential impact of ORB on their reviews

and meta-analyses by using sensitivity analysis approaches. For benefit outcomes a number of approaches for adjusting review meta-analyses for ORB have been proposed ([11], [12]), although the method suggested in this article is the currently recommended approach by the ORBIT team, as this model-based approach reflects directly the empirical finding that significant outcomes are more likely to be reported than those that are found not to be significant. No previous methodology existed for adjusting for ORB in a meta analysis of a harm outcome, although evidence suggests that the incomplete reporting of harm outcomes may be worse than benefit outcomes [13]. The novel ORB bias adjustment approach presented in this paper maps the concept identified in the ORBIT II study that biased reporting or non-reporting of harms data results from study authors (intentionally or otherwise) masking the harms profile of particular interventions in favour of the benefit results. This method complements the theoretical framework already recommended for benefit outcome adjustment. As the epilepsy example shows, ORB can effect the statistical conclusions to review meta analyses and also has the potential to effect the clinical interpretations of treatment indications in both benefit and harm outcomes.

ORB as a missing data problem.

Like all missing data problems, statistical analysis is impossible unless we make assumptions about the missing data mechanism. The usual approach in section 3.1 ignores ORB altogether, tacitly assuming that the missing outcomes are missing at random [14]. This assumption is equivalent to a Low Risk assessment under the ORBIT methodology, contradicting the High Risk assessments found for many of the missing outcomes in the epilepsy example. Our method makes the more plausible assumption that, for High Risk benefit outcomes, the actual (or potential) values of y_i, had they been reported, would not be significantly different from the null and so lie somewhere in the interval (9). For a High Risk harm outcome, our assumption is that the unobserved y_i would be somewhere in the interval $(0, \infty)$. These assumptions give an explicit selection model for the general missing data methodology introduced in [15] and reviewed in [14]. The interval structure of these assumptions means that ORB is an example of censored data, although quite different from the more usual censored life-time applications. If we think of (1) as a parametric survival model, the harms likelihood (12) is similar to the standard likelihood for right-censored data, the first term corresponding to the fully observed outcomes and the second term corresponding to data censored at the value $y = 0$. The second term in the benefits likelihood (10)

corresponds to interval censoring using (9).

Seen in this light, the sensitivity analysis suggested in this paper is a comparison between two missing data assumptions, that data are missing at random (the unadjusted approach) or that the missing studies rated as High risk correctly identify those unobserved estimates which are non-significant or positive (the adjusted approach). The question is not whether the adjustment assumptions are known to be correct, but whether the contrast between the resulting estimates gives a useful indication of sensitivity to ORB. We are comparing a statistically naive assumption (missing at random) with a more plausible assumption guided by the systematic re-appraisal of the individual research studies offered by the ORBIT methodology.

Statistical assumptions.

As well as these missing data assumptions, the example also relies on other important assumptions. We have taken the study treatment effects y_i to be log relative risk, although, arguably, log odds ratios could also have been used. The Cochrane guide [8] gives a useful summary of the treatment effect measures most often used in clinical trails, and how to calculate their variances. Note that the use of log transformations on relative risks and odds ratios is a technical device to improve normal approximations and so is not an assumption in the same sense. For example, the theory of Section 3 is based on log relative risks, but the numerical results in Tables 4 and 6 are transformed back to actual relative risks. Although normal approximations are almost always used in practice, the theory of Section 3 can also be extended to non-normal distributions. The first term in the likelihoods $L_B(\theta)$ and $L_H(\theta)$ remain as before, with $\hat{\theta}_U$ and s_θ^2 now being the appropriate maximum likelihood estimate and variance using the reported data. The quantities within the square brackets in the second terms of these likelihoods, however, need to be replaced by the appropriate probability calculations corresponding to (9) and (11). Another important assumption is the fixed effects model (1), as used in the original published Cochrane review of the epilepsy studies. Section 5 shows how these ORB adjustments can also be used for random effects models. The imputation of the missing variances using (16) is a further assumption. However, trying alternative version of this formula suggests that the adjusted estimates are reasonably robust: for example, the statistical conclusions in terms of the significance or non-significance of the adjusted estimates indicated in Table 4 continue to hold even if the imputed missing outcome variances are scaled up or down by a factor of two.

Application of the ORB sensitivity analysis.

In this paper the principal aim is to present the new methodology with an application. There is clearly a separate discussion with regards to a) when the adjustment should be applied, and to which outcomes, and b) whether the adjusted estimates make a difference clinically. Both these decisions are largely at the discretion of the clinical expert and may differ depending on the application. Our recommendation is that the adjusted estimates are always presented alongside the unadjusted estimated, as illustrated in Tables 4 and 6, such that the robustness of the conclusions to outcome reporting bias can be assessed.

Implementing the ORB adjustments suggested in this paper involve the maximization of likelihood functions and associated functions such as (7). These are relatively easy to calculate using a general maximization routine such as `nlminb` in `R` (or `Splus`), as we have used to obtain the results in Tables 4, 5 and 6. Similar facilities are routinely available in other software packages. The method extends easily to random effects models as described in Section 5, since the profile likelihood step described in that section can also be achieved using the same maximization routine. A web-based application is being developed which will enable users to easily implement these methods (http://www.outcome-reporting-bias.org/).

The correct application of the statistical adjustment method for outcome reporting bias presented in this paper relies on a formal assessment of the risk bias of missing study outcome data using ORBIT methodologies which were published in the BMJ. To date, the risk of bias assessments have only been validated for use in randomised controlled trials for benefit outcomes, and applied in only a handful of non-randomised reports for harm outcomes ([3], [4], [5]). The application of the adjustment method to non-randomised studies is therefore beyond the scope of this current research.

Supplementary material

Supplementary Table 1 gives the raw data for the Topiramate example, presented in the form of an input file for the web application referred to at the end of Section 6. The first three rows of the table indicate, respectively, outcome type (benefit or harm), percentage for the confidence intervals (95 or 99), and outcome title. Subsequent rows correspond to the 12 studies in the order given in Table 3. In the body of Supplementary Table 1, the first two columns give the sample sizes (n_{1i}, n_{2i}). The subsequent columns are

arranged in pairs, giving the observed study frequencies (a_i, c_i) for each outcome. A blank in these columns indicates that the outcome was not reported (or only partially reported), in which case the High/Low ORBIT assessment is indicated.

Supplementary Table 2 gives the detailed classification tables for each of the studies in the Topiramate review, providing justification for the ORBIT assessments.

Funding

The ORBIT and ORBIT II studies were funded by the Medical Research Council (grant number G0500952 and MR/J004855/1). Financial support for the ORBIT website was provided by the MRC Network of Hubs for Trials Methodology Research (MR/L004933/1-N67). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of this manuscript. *Competing interests*: AM is co-ordinating Editor of the Cochrane Epilepsy Group.

Acknowledgement

We are grateful to an anonymous referee for helpful comments on an earlier version of this paper.

References

1. Alderson P and Tan T. Use of Cochrane Reviews in NICE clinical guidelines [editorial]. *Cochrane Database Syst Rev 2011 Aug9*; **12**, ED000032.

2. Hutton JL and Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Applied Statistics* 2000; **49**, 359-370.

3. Dwan K, Gamble C, Williamson PR and Kirkham J J for the Reporting Bias Group. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias - An Updated Review. *PLoS ONE* 2013; **8(7)**, e66844. doi:10.1371/journal.pone.0066844

4. Saini P, Loke YK, Gamble C, Altman DG, Williamson PR and Kirkham JJ. Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews. *British Medical Journal* 2014; **349**, g6501

5. Kirkham JJ, Dwan K, Dodd S, Altman DG, Smyth R, Jacoby A, Gamble C and Williamson PR. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *British Medical Journal* 2010; **340**, c365.

6. Copas JB, Dwan K, Kirkham JJ and Williamson PR. A model-based correction for outcome reporting bias in meta-analysis. *Biostatistics* 2014; **15**, 370-383.

7. Pulman J, Jette N, Dykeman J, Hemming K, Hutton J L and Marson AG. Topiramate add-on for drug-resistant partial epilepsy. *Cochrane Database of Systematic Reviews* **2014**, Issue 2. Art. No.: CD001417. DOI: 10.1002/14651858.CD001417.pub3.

8. Deeks JJ and Higgins JPT. Statistical algorithms in Review Manager 5. http://community.cochrane.org/sites/default/files/uploads/handbook/Statistical_Methods_in_RevMan5-1.pdf.

9. Mantel N and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**, 719-748.

10. Hardy RJ and Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**, 619–629.

11. Williamson PR and Gamble C. Application and investigation of a bound for outcome reporting bias. *Trials* 2007; **8**, 9.

12. Kirkham JJ, Riley RD and Williamson PR. A multivariate meta-analysis solution for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine* 2012; **31**, 2179–2195.

13. Chan, AW, Krleza-Jeric K, Schmid I and Altman DG.. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Journal* 2004; **171**, 735-40.

14. Little RJA and Rubin DB. *Statistical Analysis with Missing Data, 2nd ed.* (2002); Wiley: New York.

15. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**, 581-592.

Table 1: ORBIT 1 table (*on separate word file*)

Table 2: ORBIT 2 table (*on separate word file*)

Table 3: ORBIT matrix (*on separate word file*)

Table 4: Unadjusted and ORB-adjusted maximum likelihood estimates and confidence intervals (benefits 95%, harms 99%)

	Unadjusted RR			Adjusted RR			n_{Rep}	n_{High}
	estimate	CIlower	CIupper	estimate	CIlower	CIupper		
Benefits								
50% seizure reduction	2.72	2.17	3.40	2.63	2.12	3.28	11	1
Seizure freedom	3.22	1.27	8.14	2.53	1.12	5.57	6	5
Harms								
Treatment withdrawal	2.26	1.33	3.85	2.29	1.35	3.89	11	1
Dizziness	1.45	1.00	2.10	1.55	1.10	2.20	7	5
Headache	0.96	0.65	1.40	1.12	0.81	1.55	8	4
Nausea/vomiting	1.45	0.68	3.10	1.87	1.06	3.57	4	7
Paraesthesias	3.73	1.45	9.65	4.23	1.80	10.36	6	6
Weight loss	2.92	1.26	6.75	3.09	1.43	6.92	8	3
Fatigue	2.12	1.39	3.22	2.15	1.43	3.24	9	2
Somnolence	2.15	1.40	3.30	2.21	1.46	3.36	8	3
Concentration impairment	6.34	1.74	23.08	6.82	2.12	23.76	6	4
Speech difficulty	2.72	0.61	12.21	3.98	1.39	14.48	3	8
Thinking abnormality	5.45	2.14	13.84	5.78	2.43	14.27	4	6
Ataxia	2.26	1.08	4.72	2.58	1.34	5.12	4	6

Table 5: Simulation study of the accuracy of the unadjusted and adjusted estimates in the example

		Bias		m.s.e.		CI risk		Mean no.		
	θ	$E(\hat{\theta}-\theta)$		$E(\hat{\theta}-\theta)^2$		$P(\theta \notin CI)$		R	H	L
Benefits		unadj	adj	unadj	adj	unadj	adj			
50% seizure reduction	0.968	0.032	-0.015	0.014	0.013	0.055	0.058	11	1	0
Seizure freedom	0.903	0.096	-0.077	0.204	0.135	0.068	0.047	6	5	1
Harms										
Treatment withdrawal	0.829	-0.012	0.004	0.045	0.042	0.011	0.010	11	1	0
Dizziness	0.441	-0.071	0.013	0.039	0.023	0.024	0.008	7	5	0
Headache	0.109	-0.100	0.012	0.030	0.016	0.029	0.010	8	4	0
Nausea/vomiting	0.628	-0.132	0.109	0.148	0.086	0.035	0.017	4	7	1
Paraesthesias	1.442	-0.115	0.026	0.202	0.134	0.019	0.010	6	6	0
Weight loss	1.127	-0.026	0.026	0.084	0.073	0.013	0.010	8	3	1
Fatigue	0.764	-0.005	0.011	0.027	0.025	0.011	0.010	9	2	1
Somnolence	0.794	-0.015	0.016	0.034	0.030	0.011	0.009	8	3	1
Concentration impairment	1.921	-0.030	0.071	0.215	0.182	0.012	0.009	6	4	2
Speech difficulty	1.381	-0.181	0.127	0.422	0.241	0.042	0.022	3	8	1
Thinking abnormality	1.754	-0.034	0.065	0.208	0.162	0.016	0.013	4	6	2
Ataxia	0.948	-0.045	0.113	0.130	0.101	0.017	0.013	4	6	2

Table 6: Unadjusted and ORB-adjusted Mantel-Haenszel estimates and confidence intervals

	Unadjusted			Adjusted		
	estimate	CIlower	CIupper	estimate	CIlower	CIupper
Benefits						
50% seizure reduction	2.97	2.38	3.72	2.87	2.31	3.57
Seizure freedom	3.41	1.37	8.51	2.66	1.19	5.78
Harms						
Treatment withdrawal	2.44	1.45	4.10	2.47	1.48	4.13
Dizziness	1.54	1.07	2.22	1.64	1.16	2.32
Headache	0.99	0.67	1.44	1.14	0.83	1.58
Nausea/vomiting	1.50	0.71	3.15	1.90	1.08	3.59
Paraesthesias	3.91	1.51	10.12	4.40	1.87	10.83
Weight loss	3.47	1.55	7.79	3.60	1.69	7.92
Fatigue	2.19	1.42	3.40	2.22	1.46	3.42
Somnolence	2.29	1.49	3.51	2.35	1.55	3.57
Concentration impairment	7.81	2.08	29.29	8.25	2.45	29.89
Speech difficulty	3.37	0.80	14.13	4.48	1.55	16.01
Thinking abnormality	5.70	2.26	14.38	6.02	2.54	14.79
Ataxia	2.29	1.10	4.77	2.61	1.36	5.16