Confidence Intervals and P-values for Meta Analysis with Publication Bias

Masayuki Henmi*and John B. Copas**

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.

**email:* M.Henmi@warwick.ac.uk

***email:* J.B.Copas@warwick.ac.uk

and

Shinto Eguchi

Institute of Statistical Mathematics and Department of Statistical Science,
The Graduate University for Advanced Studies, Minami-azabu, Tokyo 106-8569, Japan

email: eguchi@ism.ac.jp

Summary. We study publication bias in meta analysis by supposing there is a population (y, σ) of studies which give treatment effect estimates $y \sim N(\theta, \sigma^2)$. A selection function describes the probability that each study is selected for review. The overall estimate of θ depends on the studies selected, and hence on the (unknown) selection function. Our previous paper, Copas and Jackson (2004, A bound for publication bias based on the fraction of unpublished studies, *Biometrics* **60,** 146-153), studied the maximum bias over all possible selection functions which satisfy the weak condition that large studies (small σ) are as likely, or more likely, to be selected than small studies (large σ). This led to a worst-case sensitivity analysis, controlling for the overall fraction of studies selected. However, no account was taken of the effect of selection on the uncertainty in estimation. This paper extends the previous work by finding corresponding confidence intervals and P-values, and hence a new sensitivity analysis for publication bias. Two examples are discussed.

Key words: Publication bias; Selection model; Sensitivity analysis; Unpublished studies.

1. Introduction

The simplest set-up in meta analysis is to suppose that we have the results of n independent research studies, each giving an estimate y of some underlying treatment effect parameter θ. The standard fixed effects model is

$$y \sim N(\theta, \sigma^2). \tag{1}$$

We usually assume that the sample sizes in these studies are sufficiently large that we can take the within-study standard deviations σ as known, and equal to the standard errors reported in each study. Under this model, the maximum likelihood estimate of θ for observed study results (y_i, σ_i), $i = 1, 2, \cdots, n$, is the weighted average

$$\hat{\theta} = \frac{\sum w_i y_i}{\sum w_i}, \tag{2}$$

in which the ith study is given weight $w_i = 1/\sigma_i^2$. The corresponding standard normal deviate is

$$T(\theta) = (n\bar{w})^{\frac{1}{2}}(\hat{\theta} - \theta),$$

where $\bar{w} = \sum w_i/n$, leading to the confidence interval

$$\{\theta : |T(\theta)| \leq z_\alpha\} = [\hat{\theta} - z_\alpha(n\bar{w})^{-\frac{1}{2}}, \hat{\theta} + z_\alpha(n\bar{w})^{-\frac{1}{2}}], \tag{3}$$

where $z_\alpha = \Phi^{-1}(1 - \alpha/2)$ is the standard normal percentage point for coverage $1 - \alpha$. To evaluate the null hypothesis that $H_0 : \theta = \theta_0$ the corresponding two-sided P-value is

$$P_v = 2\Phi\{-|T(\theta_0)|\}. \tag{4}$$

Although widely used in practice, this simplistic model suffers from some very substantial problems, as increasingly recognized in the meta analysis literature. First, and most obvious, is heterogeneity: there may be systematic differences between the studies so that the variation between the ys is more than can be explained by the within-study variances alone. The usual approach is to add a random effect to each study so that model (1) still applies but with σ^2 replaced by $\sigma^2 + \tau^2$, where τ^2 is the random effects variance. This is the standard approach which we adopt here.

The second and more troublesome problem, which is the focus of this paper, is *publication bias*. This recognizes the tacit assumption in (1) that each y is randomly sampled, equivalent to assuming that the set of studies in the review is a random sample from some population of studies which have been, or could have been, carried out in our particular area of interest. In reality, the studies we have in the analysis are only those which have survived a lengthy process of *selection*, including the requirement that authors write up their results and that editors and referees accept them for publication, or if unpublished, that the studies are in a form which can be traced by the reviewer. Reviewers themselves have to assess the comparability and quality of each study they find, and are often highly selective in which studies they eventually choose for the meta analysis. None of these stages of selection can be plausibly described as random: each may induce a bias which needs to be taken into account in any inference about θ. Our aim is to suggest how we can modify (3) and (4) to allow for the extra uncertainty arising from these essentially unknown sources of bias. This extends the results of our earlier paper, Copas and Jackson (2004), which considered the size of the bias $\mathrm{E}(\hat{\theta}) - \theta$. For confidence intervals and P-values we need to examine the effect of study selection on the whole distribution of $\hat{\theta}$, not just on its expectation.

The concept of sampling studies from a population was made explicit in Copas and Jackson (2004), and we follow their approach again here. We describe the population of studies by a joint distribution of values of the pair (y, σ), and suppose that each population study (y, σ) has a probability $a(y, \sigma)$ of being selected. One extreme possibility is to suppose that $a(y, \sigma)$ is constant for all y and σ: this is pure random sampling and the standard inferences (3) and (4) remain valid. Another extreme is to suppose that only studies reporting 'significantly positive' results are selected: this puts $a(y, \sigma) = 1$ when $(y - \theta_0)/\sigma \geq z_\alpha$ and zero otherwise, where z_α is some fixed threshold (like 1.96). If the treatment effect actually is positive ($\theta > \theta_0$), this would imply that the probability that a study is selected decreases as σ increases. This means that small studies (large σ) are less likely to be selected than large studies, but the small studies that are selected into the meta analysis are more likely to be biased upwards. Equivalently, we can think of $1 - a(y, \sigma)$ as the probability of a study being *missing*: we then expect a tendency for the missing studies to be small in size (large σ) and more negative in outcome (smaller y). This model would result in the 'small study effect' frequently observed in practice, the funnel plot of the data (plot of σ^{-1} against y) showing a trend for the points near the bottom of the plot to be skewed towards larger estimates of the treatment effect when compared to the points near the top of the plot. We will note a hint of this pattern in both of the examples considered later.

Selection models of this kind have been widely discussed in the literature. If $a(y, \sigma)$ is known, or assumed to follow a sufficiently restricted parametric form, standard methods can be used to produce a 'bias corrected' inference (Hedges, 1984, Lane and Dunlap, 1978). Greenhouse and Iyengar (1994) extend this to include an extra parameter which measures the extent of publication bias: this parameter can be set to a range of fixed possibilities for a sensitivity analysis. Copas and Shi (2000b, 2001) pursue a similar idea using a Heckman-type selection model (Copas and Li, 1997). These and many other references are reviewed in Rothstein, Sutton and Borenstein (2005) and in Chapter 7 of Sutton *et al.* (2000). The latter text also serves as a good general introduction to the topic of meta analysis.

A central difficulty in all this work is the choice of the selection function $a(y, \sigma)$. It is clearly impossible to estimate it from the available data, and so any assumptions we make about it are essentially unverifiable. Two different selection functions give two different inferences, and we have no means of knowing which is correct. We follow Copas and Jackson (2004) by developing methods of inference which make the weakest possible assumptions about $a(y, \sigma)$, sufficiently weak that the assumptions are broadly acceptable, but not so weak that inference about θ is impossible. One effect of publication bias, noted above, is that small studies are more likely to be left out than larger studies. This is made explicit by assuming that the conditional probability of selection given σ, say $k(\sigma) = \mathrm{E}\{a(y, \sigma)|\sigma\}$, is a *non-increasing* function of σ. This means that, on average, large studies are more likely to be selected than small studies, and this is the only assumption we make about the selection process. Of course there can be no guarantee that this assumption is correct, and we can think of circumstances in which it might not be, but on the whole it seems reasonably plausible, and considerably weaker than the assumptions about selection which have sometimes been made in the literature.

Copas and Jackson (2004) use this assumption to derive an inequality for $|\mathrm{E}(\hat{\theta}) - \theta|$ which allows us to evaluate the worst-case bias for different values of the marginal selection probability $p = \mathrm{E}\{k(\sigma)\} = \mathrm{E}\{a(y, \sigma)\}$. If $p = 1$ (no selection) the bias is zero, but as p decreases from one the bias can take increasingly large positive or negative values. In practice, we want to use such a sensitivity analysis to find out how small p needs to be before the conclusion of a meta analysis is compromised. If a value of p only slightly less than one is sufficient (very few missing studies), then the conclusion is sensitive to publication bias and so should not be trusted. On the other hand, if an implausibly small value of p is needed to change the inference, then the conclusion is robust. To do this we need to see how selection affects the variance of $\hat{\theta}$ as well as the bias. We show how this can be done directly by deriving the analogous sensitivity analyses for (3) and (4).

In Section 2 we define our notation and assumptions more carefully, and briefly review the main result in Copas and Jackson (2004). Section 3 is the main section of the paper: using an extended definition of a confidence interval as discussed in Shao (2003) we show how the conventional confidence interval (3) can be widened to include all possible selection functions $a(y, \sigma)$ consistent with our assumptions. The corresponding result for P-values is given in Section 4.

Two examples are discussed in Section 5. By re-analyzing the same clinical trials example as in Copas and Jackson (2004) we compare the results with our previous work. For a more contentious example we re-analyze the data used in the meta analysis of Hackshaw, Law and Wald (1997) on the lung cancer risk of passive smoking. The possibility of publication bias in this example has been a matter of some dispute in the literature: our analysis shows that although study selection would imply that the relative risk has been exaggerated, it is unlikely to be sufficient to negate the main conclusion in Hackshaw *et al.* (1997) that passive smoking does pose a health risk, albeit at a more modest level than has been claimed.

Some concluding comments are given in Section 6. In order to make the presentation of the paper reasonably concise, we state the main results of Sections 3 and 4 as theorems, collecting the proofs together in Web Appendices associated with this paper (published as Supplementary Materials on the journal web site).

2. Preliminaries

As in (1) we assume that the outcome y of a typical study is normally distributed $N(\theta, \sigma^2)$. In a fixed effects model, θ denotes the common treatment effect over all studies whereas θ denotes the average treatment effect in a random effects model. The standard deviation σ varies across the population of studies, with distribution $f(\sigma)$, say. Each study in the population has a probability of being selected for inclusion in the meta analysis, defined by

$$a(y,\sigma) = P(\text{study selected} \mid y, \sigma).$$

As discussed in Section 1, our only assumption about the selection procedure is that the conditional probability

$$k(\sigma;\theta,a) = P(\text{study selected} \mid \sigma) = \mathrm{E}\{a(y,\sigma) \mid \sigma\} = \int_{-\infty}^{\infty} \sigma^{-1} a(y,\sigma)\phi\left(\frac{y-\theta}{\sigma}\right) dy$$

is a non-increasing function of σ, where ϕ is the density of the standard normal distribution.

Under this formulation, the joint distribution of (y,σ) for a selected study is

$$\frac{1}{\sigma p(\theta,a,f)} a(y,\sigma)\phi\left(\frac{y-\theta}{\sigma}\right) f(\sigma), \tag{5}$$

where $p(\theta,a,f)$ is the overall selection probability given by

$$p(\theta,a,f) = P(\text{study selected}) = \mathrm{E}\{a(y,\sigma)\} = \int_0^{\infty}\int_{-\infty}^{\infty} \sigma^{-1} a(y,\sigma)\phi\left(\frac{y-\theta}{\sigma}\right) f(\sigma) dy d\sigma.$$

The marginal distribution of σ for a selected study is then

$$f_o(\sigma;\theta,a,f) = \frac{1}{p(\theta,a,f)} k(\sigma;\theta,a) f(\sigma).$$

The suffix on $f_o(\sigma)$ is to emphasize that this is the distribution for *o*bserved studies, not to be confused with $f(\sigma)$ which is the distribution of σ over the assumed *population* of studies.

Our model is that the values of $(y_i,\sigma_i), i = 1,2,\ldots,n$ in the studies selected for the meta analysis are a random sample of size n from (5). For a fixed effects analysis, σ_i^2 is taken to be s_i^2, the observed within study variance of y_i. For a random effects analysis, σ_i^2 is taken to be

$$\sigma_i^2 = s_i^2 + \tau^2,$$

where τ^2 is the between study variance. In practice, τ^2 will have to be estimated (usually from the overall sample variance of the y_is) and the values of s_i^2 are themselves sample estimates. However, we follow most articles in this area by assuming that these variances are known.

If the usual model (1) is correct, then $\hat{\theta}$ in (2) is an unbiased estimate of θ, but it will suffer a bias if the data are in fact sampled from (5). Because of the simple form of $\hat{\theta}$ as a weighted average of y, the asymptotic bias is just $\mathrm{E}_o\{w(y-\theta)\}/\mathrm{E}_o(w)$, where $w = \sigma^{-2}$ and E_o denotes expectation over the distribution of *o*bserved values of (y, σ). Copas and Jackson (2004) show that if we fix $p(\theta, a, f) = p$ and $f_o(\sigma; \theta, a, f) = f_o(\sigma)$, and assume that $k(\sigma; \theta, a)$ is non-increasing, then

$$\left| \frac{\mathrm{E}_o\{w(y-\theta)\}}{\mathrm{E}_o(w)} \right| \leq \frac{\bar{\sigma}}{p} \phi\{\Phi^{-1}(p)\}, \tag{6}$$

where Φ is the standard normal distribution function and

$$\bar{\sigma} = \frac{\mathrm{E}_o(\sigma^{-1})}{\mathrm{E}_o(\sigma^{-2})} = \frac{\int_0^\infty \sigma^{-1} f_o(\sigma) d\sigma}{\int_0^\infty \sigma^{-2} f_o(\sigma) d\sigma}. \tag{7}$$

These bounds for the bias depend on p and on $f_o(\sigma)$ through the moments ratio (7), both of which are unknown. For a sensitivity analysis, Copas and Jackson (2004) suggested taking a range of possible values for p, and taking (7) equal to its value when $f_o(\sigma) = \hat{f}_o(\sigma)$, the empirical distribution of the values of σ actually observed in the meta analysis. This is equivalent to taking $\bar{\sigma} = \sum \sigma_i^{-1} / \sum \sigma_i^{-2}$. To aid interpretation they suggested taking $p = n/(n+m)$ with $m = 0, 1, \cdots$, so that m can be thought of as the number of missing (unpublished) studies.

It is worth emphasizing the logical steps which Copas and Jackson (2004) used in this argument, since we will follow essentially the same sequence of ideas in the more complicated settings of confidence intervals and P-values. The three essential steps are

Step 1: study the case when the functions $a = a(y, \sigma)$ and $f = f(\sigma)$ are given;

Step 2: study the results of *Step 1* when (a, f) are allowed to vary over all possibilities consistent with given values of p and $f_o(\sigma)$ and with the requirement that $k(\sigma)$ is non-increasing;

Step 3: for any integer m evaluate the results of *Step 2* for $f_o(\sigma) = \hat{f}_o(\sigma)$ and $p = n/(n+m)$, and repeat this for $m = 0, 1, 2, \cdots$.

3. Confidence intervals allowing for selection

Firstly, for *Step 1*, suppose that the selection function $a(y, \sigma)$ and the marginal distribution $f(\sigma)$ are both given. Then an asymptotic confidence interval for θ follows from the log-likelihood function under model (5), which is

$$l(\theta) \;=\; \sum_{i=1}^{n} \left\{ -\log p(\theta, a, f) + \log a(y_i, \sigma_i) - \log(\sqrt{2\pi}\sigma_i) - \frac{1}{2}\left(\frac{y_i - \theta}{\sigma_i}\right)^2 + \log f(\sigma_i) \right\}.$$

The corresponding standardized score statistic is

$$T(\theta, a, f) = \frac{\partial l/\partial\theta}{\sqrt{\mathrm{Var}_o\{\partial l/\partial\theta\}}} = \frac{\sqrt{n}\{\bar{w}(\hat{\theta} - \theta) - B_1(\theta, a, f)\}}{\sqrt{B_2(\theta, a, f) - \{B_1(\theta, a, f)\}^2}}, \tag{8}$$

where Var_o denotes variance with respect to the distribution (5), and

$$B_1(\theta, a, f) = \mathrm{E}_o\{w(y - \theta)\}\ , \quad B_2(\theta, a, f) = \mathrm{E}_o\{w^2(y - \theta)^2\}. \tag{9}$$

Since the statistic (8) converges in distribution to the standard normal distribution under model (5), we have the score-based asymptotic confidence interval for θ,

$$\{\theta : |T(\theta, a, f)| \leq z_\alpha\}. \tag{10}$$

Note that in the special case when $a(y, \sigma) = 1$ for all y and σ, so there is no selection, then $B_1 = 0$, $B_2 = \mathrm{E}_o(w)$ and so (10) reduces to the usual confidence interval (3) if $\mathrm{E}_o(w)$ is estimated by the sample mean $\bar{w}$ in the usual way.

Moving on to *Step 2*, we now need to expand the interval (10) to allow for all possible choices of (a, f) consistent with chosen fixed values of p and $f_o(\sigma)$, and with our monotonicity assumption on $k(\sigma)$. To do this, denote by S be the set of all trios (θ, a, f) which satisfy the following requirements:

$$(\theta, a, f) \in S \Leftrightarrow \begin{cases} p(\theta, a, f) = p, \;\; f_o(\sigma; \theta, a, f) = f_o(\sigma) \\ k(\sigma; \theta, a) \text{ is a non-increasing function of } \sigma \\ |T(\theta, a, f)| \leq z_\alpha \end{cases}$$

Since the distribution of $T(\theta, a, f)$ is asymptotically standard normal, the set S is a random set which includes the true values of (θ, a, f) with (asymptotic) probability $(1 - \alpha)$. Now define R to be the set of all values of θ such that there exists at least one pair (a, f) for which (θ, a, f) belongs to S. Then, as the event $(\theta, a, f) \in S$ necessarily implies that $\theta \in R$,

$$P(\theta \in R) \geq P((\theta, a, f) \in S) = P(|T(\theta, a, f)| \leq z_\alpha).$$

Thus

$$\liminf_{n \longrightarrow \infty} P(\theta \in R) \geq 1 - \alpha. \tag{11}$$

Using the rather general definition of confidence region discussed in Shao (2003, p.142), expression (11) establishes that R is a confidence region for θ with asymptotic significance level $1 - \alpha$.

We have given a formal definition of R as a confidence region, but for this to be useful we need firstly to confirm that it is an interval, and secondly to find its lower and upper limits. Both are established in the following theorem:

THEOREM 1. *The confidence region R is an interval with lower and upper limits*

$$\hat{\theta} + \frac{1}{\bar{w}} L(\alpha, p, f_o) \quad \textit{and} \quad \hat{\theta} + \frac{1}{\bar{w}} U(\alpha, p, f_o)$$

respectively, where

$$L(\alpha, p, f_o) = \min_{\lambda} C^*_-(\lambda, \alpha, p, f_o), \;\; U(\alpha, p, f_o) = \max_{\lambda} C^*_+(\lambda, \alpha, p, f_o)$$

with

$$C^*_\pm(\lambda, \alpha, p, f_o) \;\; = \;\; -B^*_1(\lambda, p, f_o) \pm n^{-\frac{1}{2}} z_\alpha \sqrt{B^*_2(\lambda, p, f_o) - \{B^*_1(\lambda, p, f_o)\}^2},$$

$$
\begin{aligned}
B_1^*(\lambda, p, f_o) &= p^{-1}\mathrm{E}_o[\sigma^{-1}\{\phi(\lambda\sigma + e) - \phi(\lambda\sigma - e)\}], \\
B_2^*(\lambda, p, f_o) &= \mathrm{E}_o(\sigma^{-2}[1 + p^{-1}\{(\lambda\sigma + e)\phi(\lambda\sigma + e) - (\lambda\sigma - e)\phi(\lambda\sigma - e)\}]),
\end{aligned}
$$

and where $e = e(\lambda, \sigma, p)$ is defined by

$$\Phi(\lambda\sigma - e) + \Phi(-\lambda\sigma - e) = p.$$

The proof of Theorem 1 is given in Web Appendix A to this paper.

Theorem 1 is the result of *Step 2*. To implement *Step 3*, we now take $f_o(\sigma) = \hat{f}_o(\sigma)$ and $p = \hat{p} = n/(n+m)$ for some fixed non-negative integer m. The resulting confidence interval is

$$\Big[\hat{\theta} + \frac{1}{\bar{w}}\hat{L}(m), \quad \hat{\theta} + \frac{1}{\bar{w}}\hat{U}(m)\Big], \tag{12}$$

where

$$\hat{L}(m) = L(\alpha, \hat{p}, \hat{f}_o) = \min_{\lambda} C_-^*(\lambda, \alpha, \hat{p}, \hat{f}_o), \quad \hat{U}(m) = U(\alpha, \hat{p}, \hat{f}_o) = \max_{\lambda} C_+^*(\lambda, \alpha, \hat{p}, \hat{f}_o) \tag{13}$$

and

$$C_\pm^*(\lambda, \alpha, \hat{p}, \hat{f}_o) \quad = \quad -B_1^*(\lambda, \hat{p}, \hat{f}_o) \pm n^{-\frac{1}{2}} z_\alpha \sqrt{B_2^*(\lambda, \hat{p}, \hat{f}_o) - \{B_1^*(\lambda, \hat{p}, \hat{f}_o)\}^2}.$$

The moments B_1^* and B_2^* needed here are

$$B_1^*(\lambda, \hat{p}, \hat{f}_o) = n^{-2}(n+m)\sum_{i=1}^{n} \sigma_i^{-1}\{\phi(\lambda\sigma_i + e_i) - \phi(\lambda\sigma_i - e_i)\}, \tag{14}$$

and

$$
\begin{aligned}
&B_2^*(\lambda, \hat{p}, \hat{f}_o) \\
= \; & n^{-1}\sum_{i=1}^{n} \sigma_i^{-2}[1 + n^{-1}(n+m)\{(\lambda\sigma_i + e_i)\phi(\lambda\sigma_i + e_i) - (\lambda\sigma_i - e_i)\phi(\lambda\sigma_i - e_i)\}],
\end{aligned} \tag{15}
$$

where $e_i = e(\lambda, \sigma_i, \hat{p})$ is defined by

$$\Phi(\lambda\sigma_i - e_i) + \Phi(-\lambda\sigma_i - e_i) = n(n+m)^{-1} \tag{16}$$

for $i = 1, \ldots, n$.

For the sensitivity analysis, interval (12) is calculated for $m = 1, \cdots$. When $m = 0$, the case of no selection, (14) and (15) are 0 and $\bar{w}$ respectively, so (12) is exactly the same as the conventional confidence interval (3). For $m \geq 1$, equation (16) is easy to solve numerically as the left hand side of (16) is a strictly decreasing function of e_i and so the solution for e_i is unique. The minimum and maximum required in (13) are also relatively straightforward to evaluate numerically as in both cases the solution for λ is again unique.

We demonstrate the results of this calculation in the examples in Section 5.

4. Bound for the P-value

In many applications of meta analysis we are interested in evaluating the evidence the data give about a null hypothesis $H_0 : \theta = \theta_0$ (for example that a relative risk equals one). We now study the effect of selection on the P-value (4). For this we follow the same three steps as before.

The solution to *Step 1* follows directly from (8): if a and f are given then the two-sided asymptotic P-value is

$$P_v(a, f) = 2\Phi\{-|T(\theta_0, a, f)|\}. \tag{17}$$

For *Step 2* we want to allow a and f to vary over all possibilities consistent with given values of p and f_o and with our monotonicity requirement. The typical effect of publication bias is that the evidence against H_0 is exaggerated (P-values too small), so for a worst case sensitivity analysis we want to evaluate the maximum value that (17) can take over these possibilities. This bound is given in the following theorem:

Theorem 2. *For given p and $f_o(\sigma)$, suppose that $p(\theta_0, a, f) = p$, $f_o(\sigma; \theta_0, a, f) = f_o(\sigma)$ and that $k(\sigma; \theta_0, a)$ is a non-increasing function of σ. Then*

$$P_v(a, f) \leq 2\Phi\{-T_{min}(\theta_0, p, f_o)\}, \tag{18}$$

where

$$T_{min}(\theta_0, p, f_o) = \min_{\lambda} \left| \frac{\sqrt{n}\{\bar{w}(\hat{\theta} - \theta_0) - B_1^*(\lambda, p, f_o)\}}{\sqrt{B_2^*(\lambda, p, f_o) - \{B_1^*(\lambda, p, f_o)\}^2}} \right| . \tag{19}$$

The bound is attained when

$$a(y, \sigma) = \begin{cases} 1 & \mathit{if}\ y \leq \theta_0 + \sigma(\lambda^*\sigma - e^*) \\ 1 & \mathit{if}\ y \geq \theta_0 + \sigma(\lambda^*\sigma + e^*) \\ 0 & \mathit{otherwise} \end{cases} , \tag{20}$$

where λ^* *is the value of* λ *at which (19) is attained and* $e^* = e(\lambda^*, \sigma, p)$.

The proof of Theorem 2 is given in Web Appendix B to this paper.

For *Step 3* we evaluate the bound in the theorem for $f_o = \hat{f}_o$ and $p = \hat{p} = n/(n+m)$ as before. The values of B_1^* and B_2^* needed in (19) are exactly the same as the previous formulae (14) and (15). The value of λ minimizing (19) is unique, again as before. For a sensitivity analysis we do this calculation for $m = 1, 2, \cdots$.

When $m = 0$, meaning there is no selection, the upper bound reduces to the conventional P-value (4), as expected. The bound increases, or the evidence against H_0 weakens, as m becomes larger. If (4) is less than some conventional significance threshold (like 0.05), then there will be a value of m for which the bound crosses above this threshold. As discussed in Section 1, we take this value of m (the number of unpublished studies needed to discredit the claimed significance) as an informal measure of the robustness of the evidence to publication bias.

The two methods proposed here for sensitivity analysis, using confidence intervals and P-values, seem at first glance to be rather different. One involves a definition of confidence interval which is more general than the usual one, whereas the other adopts a worst case strategy more directly by finding an upper bound. Now in simple problems the familiar relationship between significance tests and confidence intervals is that the null hypothesis $\theta = \theta_0$ is significant at the α level if and only if θ_0 lies outside the confidence interval with confidence coefficient $(1 - \alpha)$. In a straightforward manner from the definition of the confidence interval R in Section 3, we can show that this natural relationship continues to hold in our more general setting. If we strengthen the requirement for significance to mean that the *maximum* P-value in Theorem 2 has to be less than α, then we end up rejecting precisely those values of θ_0 which lie outside the confidence interval of Theorem 1. This consistency between Theorems 1 and 2 will be demonstrated in the examples in the next section.

5. Examples

Clinical Trials Example

The example in Copas and Jackson (2004), taken from the Cochrane database, reports the results of 14 randomized clinical trials concerning the use of prophylactic corticosteroids in cases of premature birth. Briefly, if a birth is anticipated to be premature, the treatment is administered to the mother in order to improve the chance of the infant's survival. The events are the deaths of the infants, and θ is the underlying log-odds ratio comparing the probability of death in the treated group with the probability for a parallel sample of controls. In 13 out of the 14 trials the estimate y of θ is negative *i.e.* the treatment appears to be effective in reducing risk.

The raw data, and corresponding values of y_i and s_i, are listed in Table 1 of our previous paper, illustrated here in Figure 1. This is the funnel plot, the crosses on the graph being the points $(y_i, 1/s_i)$. Notice the tendency for points near the bottom of the graph (smaller studies) to have smaller y (stronger treatment effect) than the points near the top of the graph (larger studies). The horizontal bars through each point indicate the individual study confidence intervals $y_i \pm 2s_i$.

The natural model here is fixed effects, since the data give no evidence of heterogeneity (the maximum likelihood estimate of τ is in fact zero). Thus we set $\sigma_i = s_i$, giving

$\hat{\theta} = -0.48$, and with $\alpha = 0.05$ the confidence interval (3) is $(-0.71, -0.25)$. The P-value (4) is 5.3×10^{-5}, indicating strongly significant evidence that the treatment is effective. The data suggest that the treatment reduces mortality by almost 40%.

However, the clear trend in Figure 1 suggests there may be some missing studies with larger values of y_i, which would mean that the treatment effect has been exaggerated, possibly substantially so. For a given number of unpublished studies (m), formula (12) gives the confidence interval that takes into account the possibility of such a selection mechanism. With $\alpha = 0.05$, Figure 2 (solid lines and the left hand vertical scale) plots the confidence limits against m. The upper limit increases from its conventional value of -0.25 to cross the null line $\theta_0 = 0$ at $m = 13$. This is confirmed by the dashed line in Figure 2, which shows (using the right hand vertical scale) the corresponding bound for the P-value (18). This increases and rises above 5% when m reaches 13. Both analyses show that if there are 13 or more unpublished studies then the significance of the result is overturned, in the sense that there exists a selection mechanism within our assumptions for which θ might reasonably be positive (treatment actually harmful). If this number of unpublished studies is judged to be unreasonably large, meaning that only half of the studies have been selected, then the result in favour of the treatment seems reasonably safe, although the claim of a 40% reduction in risk needs to be interpreted with considerable caution.

Figure 2 of Copas and Jackson (2004) plotted the maximum bias (6) against m, but did not consider the effect of selection on the uncertainty of $\hat{\theta}$. They argued informally that as the upper confidence limit is -0.25, a bias of $+0.25$ would be needed to upset the inference. This happens when $m = 9$, considerably smaller (more conservative) than the value $m = 13$ from our analysis here.

Epidemiological Example

The second example is the meta analysis published by Hackshaw *et al.* (1997) of the accumulated evidence on lung cancer and passive smoking (environmental tobacco smoke), a topic of much current debate (related papers include Givens *et al.*, 1997; Poswillo *et al.*, 1998; Copas and Shi, 2000a and 2000b). Hackshaw's paper reviewed 37 published (mostly case-control) studies of the risk of lung cancer in female non-smokers whose spouses/partners did or did not smoke. Each of these studies reported an estimate of the relative risk (odds ratio) and a 95% confidence interval. Most of the 37 studies found an increased risk in the exposed group, but a few came to the opposite conclusion. The data are listed in detail in Hackshaw *et al.* (1997) and shown here in Figure 3, constructed in the same way as Figure 1 above. There is some hint of a drift to the right (greater risk) as we read down the plot from the larger to the smaller studies, but less marked than the trend the other way round in Figure 1.

The standard method of DerSimonian and Laird (1986) gives $\tau^2 = 0.0176$ and so we set $\sigma_i^2 = s_i^2 + 0.0176$. This gives the usual random effects analysis: the overall (average) log relative risk is $\hat{\theta} = 0.21$, with 95% confidence interval $(0.12, 0.30)$. According to this, the added risk from exposure is 23% with confidence interval $(13\%, 35\%)$. The P-value (4) is 5.2×10^{-6}, leading to the claim in Hackshaw *et al.* (1997) that there is very strong evidence for the risk of passive smoking. As before, the possibility of there being other studies reporting lower levels of risk raises doubts about the validity of these figures.

Figure 4 is the analogue of Figure 2 for these data. The lower confidence limit (lower solid line) and the bound for the P-value (dashed line) cross $\theta = 0$ and $P = 0.05$ respectively at the same point, $m = 19$. According to our argument, there would have to be as many as 19 other studies excluded from the meta analysis before the conclusion could be seriously questioned. As mentioned in Section 1, the possibility of study selection here has been the subject of some contention, but to imagine that there are as many as 19 studies of a comparable size which have been excluded does seem rather extreme. In this sense, the significance of the evidence stands, although the actual size of the risk may well have been exaggerated.

Copas and Shi (2000b) also re-analyse these data, but they use a parametric model for selection rather than the worst case strategy adopted here. For each choice of their selection parameter, they report a likelihood-based confidence interval for θ and an estimate of the expected number of unpublished studies (corresponding roughly to our m). Table 1 of their paper shows that when m reaches 28 the lower confidence limit reaches zero. As expected, their value is greater than the $m = 19$ found here, because we allow for all possible selection mechanisms which satisfy our monotonicity assumption and not just the particular selection formula which they assume. This illustrates the difficulty with this and other parametric approaches — it is impossible to check the validity of a selection model from the available data, and yet we can find another model for which the critical m is smaller. Arguably, parametric methods in this context are too sensitive to modelling assumptions to be very useful.

6. Comments

1. We have suggested that m can be interpreted as the number of unpublished studies. This should not be taken too literally: it is p and not the size of the population of studies which we are controlling in the sensitivity analysis. Instead of plotting the confidence limits and P-values against p, we use the simple transformation $m = n(1/p - 1)$ and plot them against m instead. Alternatively, we could think of p as an unknown parameter which we are estimating by $\hat{p} = n/(n + m)$. A completely different approach would be to fix $N = n + m$ as the sensitivity parameter instead of p, and base inference on the distribution of possible choices of n studies selected out of N.

2. Similarly, the idea of a population of studies is a mathematical model for discussing selection, and not a literal description of any particular body of research. In practice no two

studies will be exactly the same, even if they appear to be addressing the same question. There will always be differences in research protocol and design, and it is a matter of judgement which studies are deemed sufficiently similar to be included. The $n + m$ studies in our model are those which either have been deemed comparable, or would have been had they been published or otherwise accessible to the reviewer.

3. Our methods are based on the asymptotic distribution of $\hat{\theta}$ in (2) under distribution (5), and take no account of the particular characteristics of the observed sample. Often the data give us little or no information about selection, but consider the studies with large σ in Figure 1. If model (1) is correct, which means that the population distribution of y is symmetrical, then the skewness of this plot gives us some evidence that $a(y, \sigma)$ for these small studies is more likely to be a decreasing function of y than an increasing function of y, and this is evidence which our analysis ignores. Further, extremal selection functions such as (20) can give zero probability to certain values of y, and this could be contradicted by observed values. One advantage of fully parametric approaches such as Copas and Shi (2000b) is that by basing inference on the observed likelihood they retain what information there is in the funnel plot.

4. We have already commented on the difference between the arguments used in Sections 3 and 4. The key point is that, although $a(y, \sigma)$ is a function only of the estimate and variance of each study, marginal selection probabilities derived from it, such as $k(\sigma)$ and p, depend also on θ. Thus it is not simply a matter of finding the maximum and minimum of the confidence limits within our constraints on p and $k(\sigma)$, since these constraints involve θ which is not fixed but varies within the confidence interval. The difficulty does not arise for P-values since we only need to consider what happens at one fixed value $\theta = \theta_0$. If we had restricted our attention to the special case where the selection function a depends on y only through the standardized value $(y - \theta)/\sigma$, then the first problem would reduce to

that of finding the bound for the confidence limits, because the values of B_1 and B_2 defined in (9) would be independent of θ. However, it seems sensible to imagine that the selection of an individual study depends on its values of y and σ but not on the unknown quantity θ. Our rather cumbersome notation (such as $p(\theta, a, f)$ instead of just p) is our attempt to make these subtle dependences clear.

5. We envisage σ as a random variable across a population of studies, and allow selection to depend on σ as well as on y. This differs from most of the literature on publication bias which considers the distribution of each observed y_i, essentially conditioning on the observed values $\sigma_1, \sigma_2, \cdots, \sigma_n$. However, the difference is less than it may seem, as we end up estimating the required moments (or functionals) of f_o by their values at $f_o = \hat{f}_o$, the empirical distribution of the observed σ_is. We could rework Section 3 by evaluating the likelihood and score statistics *conditionally* on the observed σ_is. This would give standardized score statistic $|T^{(c)}|$, say. Then we can show that $|T^{(c)}| \geq |T(\theta, a, \hat{f})|$ where T is the unconditional standardized score statistic in (8). Thus if $|T^{(c)}| \leq z_\alpha$ then necessarily $|T| \leq z_\alpha$ which shows that the defining confidence property (11) holds both conditionally and unconditionally.

6. It would be interesting to extend our method to cover the case when τ^2 is estimated. We would then lose the simplicity of our theory because $\bar{w}(\hat{\theta} - \theta)$ would no longer simply be a linear function of the y_is. More complicated asymptotic approximations would be needed.

7. Our final comment, which applies to much of the literature on meta analysis as well as to this paper, is to point out the approximation involved when treating s_i as fixed in the case of 2 × 2 tables, as we have done in both examples in Section 5. When y_i and s_i are calculated from the same set of four frequencies they are correlated and so the conditional distribution of y_i given s_i is no longer the same as the unconditional distribution of y_i. There is no problem if the study sample sizes are large (as in our epidemiological example),

but this can be important if any of the observed frequencies are small (as in our clinical trials example).

Supplementary Materials

Web Appendices referenced in Sections 3 and 4 are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

Acknowledgements

We thank the referee and the associate editor for their comments. This research is supported by a grant from the *EPSRC*.

References

Copas, J. B. and Jackson, D. (2004). A bound for publication bias based on the fraction of unpublished studies. *Biometrics* **60,** 146-153.

Copas, J. B. and Li, H. G. (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society B* **59,** 55-95.

Copas, J. B. and Shi, J. Q. (2000a). Reanalysis of epidemiological evidence on lung cancer and passive smoking. *British Medical Journal* **320,** 417-418.

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

Copas, J. B. and Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research* **10,** 251-265.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7,** 177-188.

Givens, G. H., Smith, D. D. and Tweedie, R. L. (1997). Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* **12,** 221-250.

Greenhouse, J. and Iyengar, S. (1994). Sensitivity analysis and diagnostics. In *The Handbook of Research Synthesis*, H. Cooper and L. V. Hedges (eds), New York: Sage.

Hackshaw, A. K., Law, M. R. and Wald, N. J. (1997). The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical Journal* **315,** 980-988.

Hedges, L. V. (1984). Estimation of effect size under non random sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Education Statistics* **9,** 61-85.

Lane, D. M. and Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology* **31,** 107-112.

Poswillo, D. *et al.* (1998). *Report of the Scientific Committee on Tobacco and Health.* London: The Stationery Office.

Rothstein, H. R., Sutton, A. J. and Borenstein, M. (eds) (2005). *Publication Bias in Meta-Analysis.* Chichester: Wiley.

Shao, J. (2003). *Mathematical Statistics,* second edition. New York: Springer-Verlag,

Sutton, A. J., Abrams, K. R., Jones, D. R. and Sheldon, T. A. (2000). *Methods for meta-analysis in medical research.* Chichester: Wiley.

Captions for figures

Figure 1: Funnel plot for corticosteroids data. The horizontal bars through each point indicate the individual study confidence intervals.

Figure 2: Confidence limits and P-values for corticosteroids data. The solid lines show the upper and lower confidence limits, whereas the dashed line shows the bound for the P-value against the number of unpublished studies.

Figure 3: Funnel plot for passive smoking data. The horizontal bars through each point indicate the individual study confidence intervals.

Figure 4: Confidence limits and P-values for passive smoking data. The solid lines show the upper and lower confidence limits, whereas the dashed line shows the bound for the P-value against the number of unpublished studies.

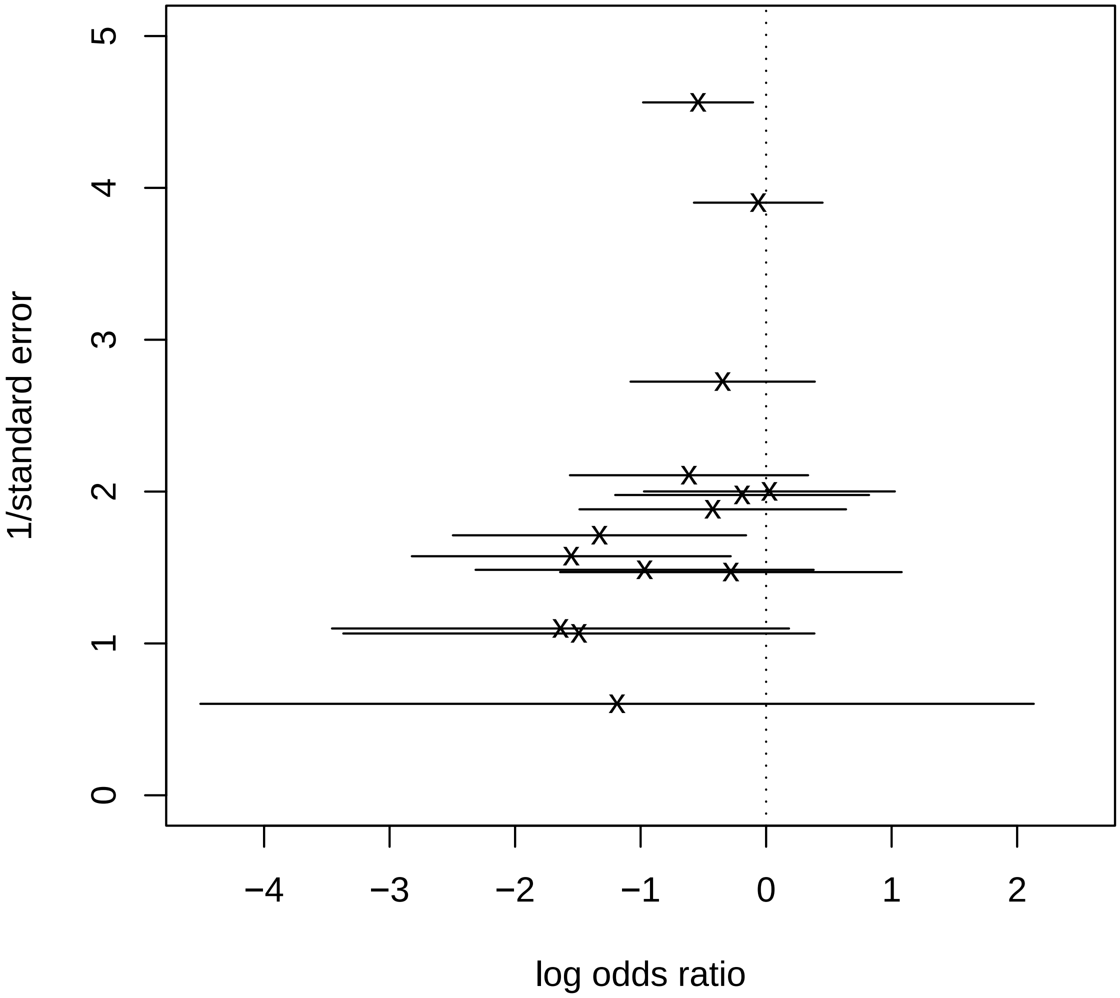

Figure 1: Funnel plot for corticosteroids data. The horizontal bars through each point indicate the individual study confidence intervals.

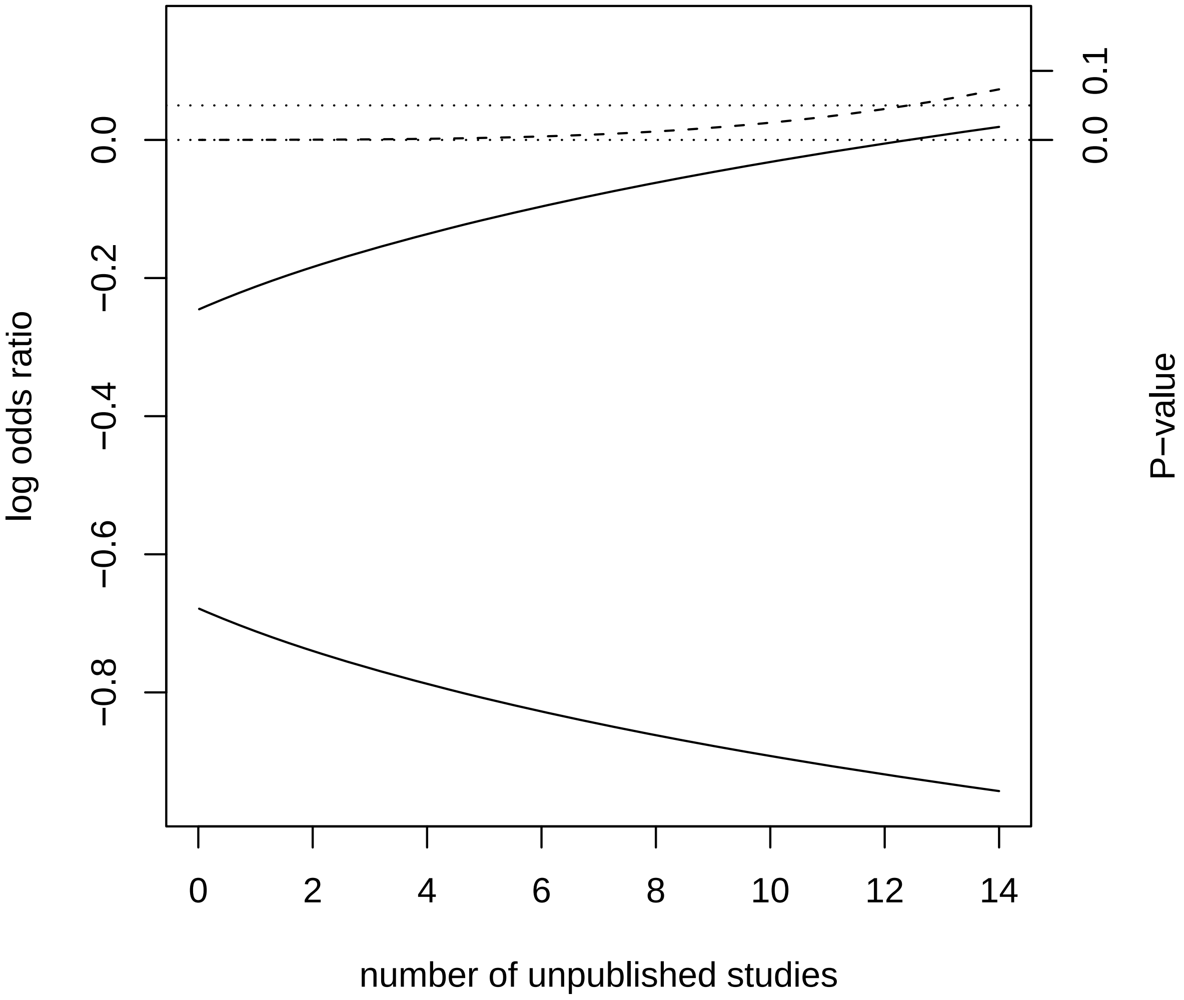

Figure 2: Confidence limits and P-values for corticosteroids data. The solid lines show the upper and lower confidence limits, whereas the dashed line shows the bound for the P-value against the number of unpublished studies.

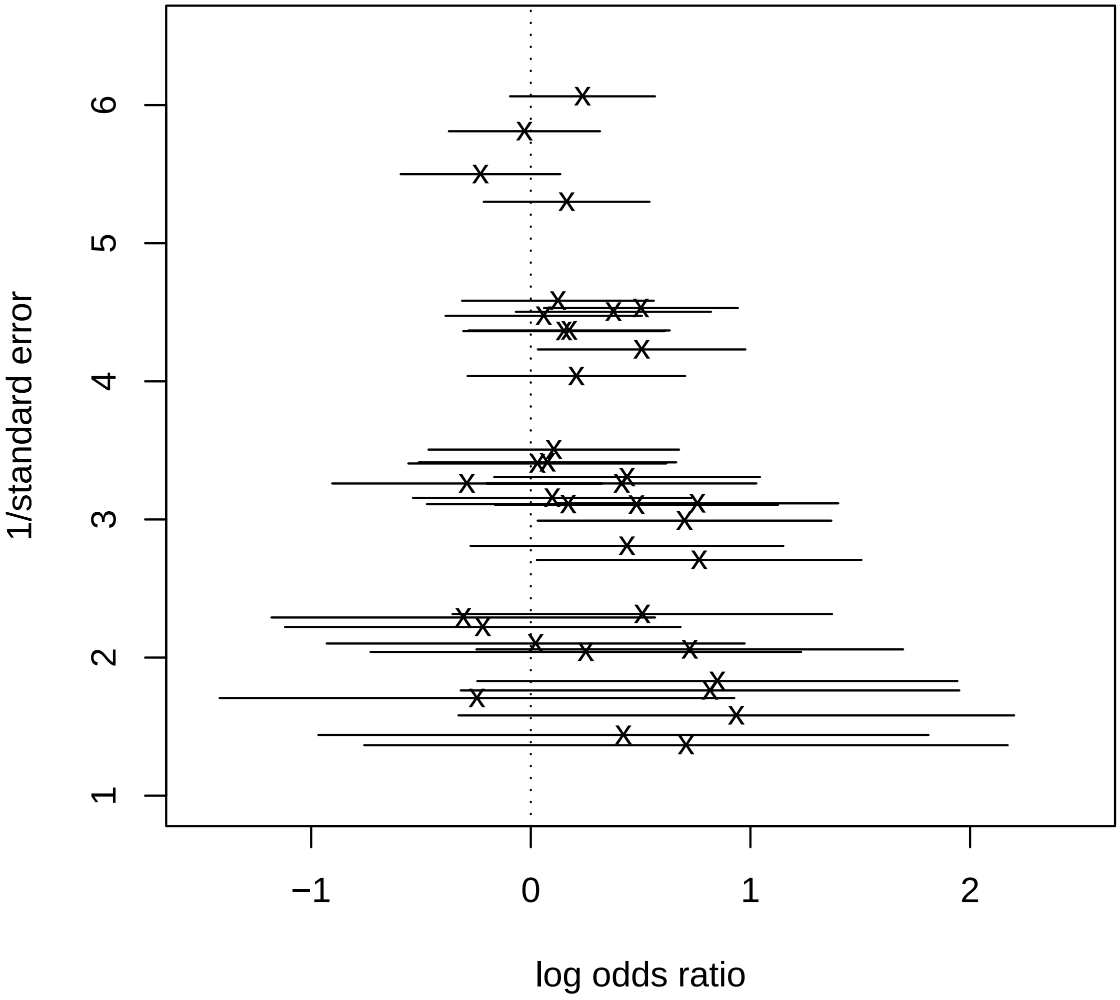

Figure 3: Funnel plot for passive smoking data. The horizontal bars through each point indicate the individual study confidence intervals.

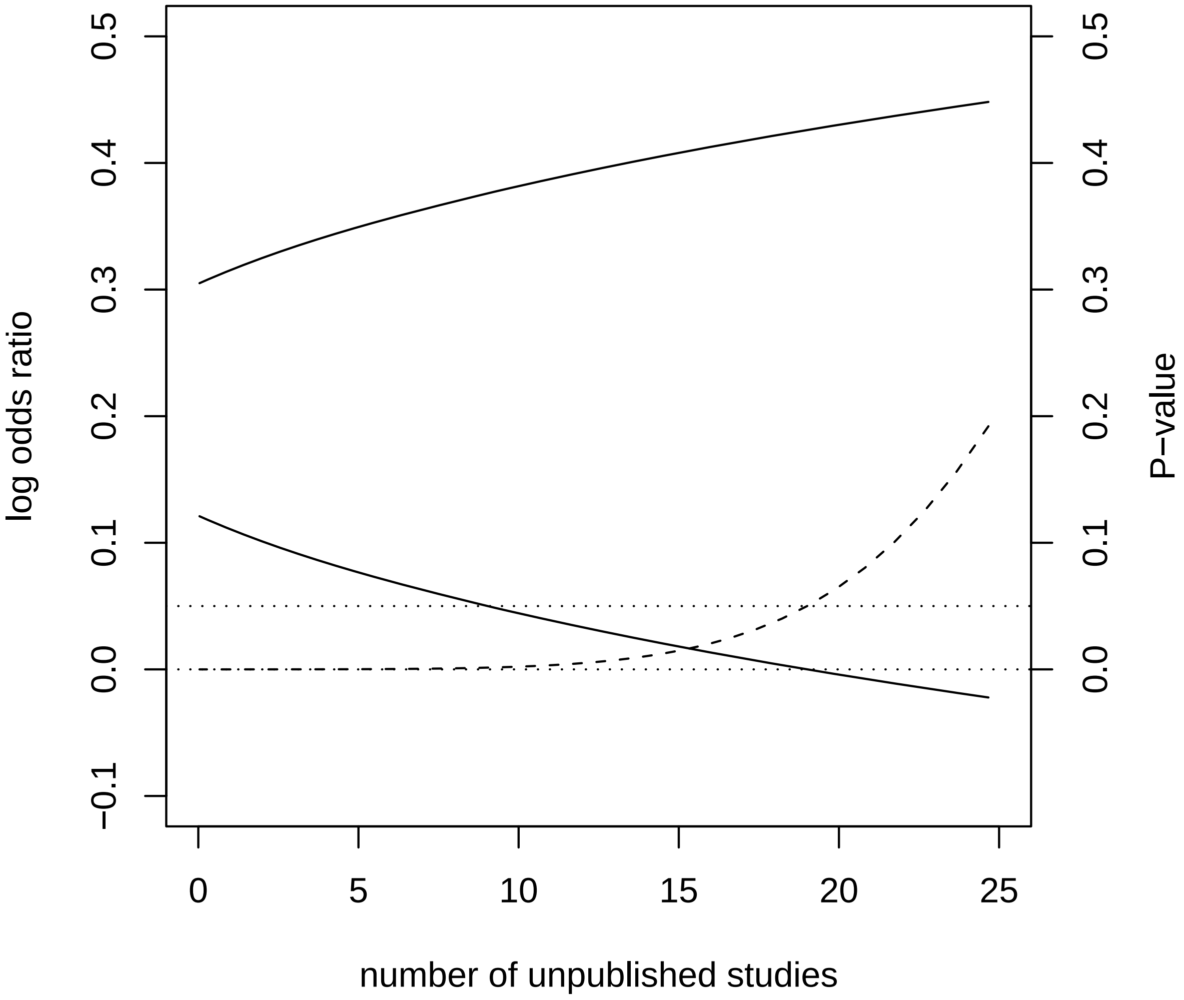

Figure 4: Confidence limits and P-values for passive smoking data. The solid lines show the upper and lower confidence limits, whereas the dashed line shows the bound for the P-value against the number of unpublished studies.