STATISTICS IN MEDICINE

Statist. Med. 2000; **00**:1–12 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

A robust P-value for treatment effect in meta analysis with publication bias

John B. Copas*and Paul F. Malley

Deparment of Statistics, University of Warwick, Coventry, CV4 7AL, UK.

SUMMARY

Publication bias is a major and intractable problem in meta analysis. There have been several attempts in the literature to adapt methods to allow for such bias, but these are only possible if we are prepared to make strong assumptions about the underlying selection mechanism. We discuss the assumption that the probability that a paper is published may depend in some unspecified way on the P-value being claimed by that study. We suggest a new robust P-value for the overall treatment effect which turns out to be closely related to the correlation of the associated radial plot. Properties of the method are discussed and illustrated on two examples. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: Publication bias; Selection bias; Selection model; Sensitivity analysis; Unpublished studies.

*Correspondence to: J. B. Copas. jbc@stats.warwick.ac.uk; phone 02476523370; fax 02476524532

Contract/grant sponsor: EPSRC

17 January 2000

18 September 2002

1. INTRODUCTION

The usual fixed effects model in meta analysis is that we have k independent research studies, each of which gives an estimate $\hat{\theta}_i$ with

$$\hat{\theta}_i \sim N(\theta, \sigma_i^2), \quad i = 1, 2, \cdots, k. \tag{1}$$

Here, θ is the true treatment effect assumed common to all studies (the fixed effects assumption), and σ_i^2 is the (assumed known) within-study variance for the ith study. Under this simple model the maximum likelihood estimate of θ weights each study estimate inversely to its variance, giving the fixed effects estimate

$$\tilde{\theta} = \frac{\sum w_i \hat{\theta}_i}{\sum w_i}, \quad w_i = \frac{1}{\sigma_i^2}, \quad Var(\tilde{\theta}) = \frac{1}{\sum w_i}. \tag{2}$$

See Sutton *et al.* [1] for a good discussion of this and other methods of meta analysis.

In practice, evidence for a treatment effect is often measured in terms of a P-value. Assuming that the treatment has no effect under $H_0 : \theta = 0$, and is beneficial if $\theta > 0$, the one-sided P-value from the ith study is

$$P_i = \Phi\{-\sigma_i^{-1}\hat{\theta}_i\}, \tag{3}$$

where Φ is the standard normal cumulative distribution function. Similarly, the P-value from the meta analysis as a whole is

$$P = \Phi\{-(\sum w_i)^{\frac{1}{2}}\tilde{\theta}\}. \tag{4}$$

The *radial plot* [2] is a useful way of representing the data in a meta analysis. This plots the points y_i against x_i where

$$y_i = \frac{\hat{\theta}_i}{\sigma_i}, \quad x_i = \frac{1}{\sigma_i} = w_i^{\frac{1}{2}}. \tag{5}$$

In this notation, model (1) can be rewritten as

$$y_i \sim N(\theta x_i, 1), \tag{6}$$

Prepared using simauth.cls

so y_i has a linear regression on x_i through the origin, with slope θ and residual variance equal to one. The quantities in (2), (3) and (4) are now

$$\tilde{\theta} = \frac{\sum x_i y_i}{\sum x_i^2}, \quad P_i = \Phi\{-y_i\}, \quad P = \Phi\{-(\sum x_i^2)^{\frac{1}{2}}\tilde{\theta}\}.$$

Thus the y coordinates are the study specific P-values plotted on a probit scale, the x coordinates are the study precisions (one over the standard deviation), $\tilde{\theta}$ is the slope of the least squares regression of y on x through the origin, and P is the P-value for testing the significance of this slope.

The simplicity of this model disguises the fact that in practice there are often major problems which threaten its validity. Most intractable of these is *publication bias*, which recognizes that a systematic review does not cover all relevant studies in the area of interest, but only those which have been written up and published, or otherwise available to the reviewer. A reasonable conjecture is that studies which report a significant treatment effect are more likely to be published than studies where the results are inconclusive, which means that the studies in the review will be biased. The P-value in (4) will be too small, exaggerating the real evidence for a treatment effect.

The text edited by Rothstein *et al.* [3] gives an excellent review of the problem of publication bias and the various approaches in the literature which have attempted to overcome it. Prominent amongst these is the selection model approach, reviewed in the chapter by Hedges and Vevea [4]. Here we envisage the population of all relevant studies which have been done in the area of interest, only k of which are selected for the systematic review. If the selection of studies depends on their P-values (and only on their P-values), as conjectured above, then the probability that the ith study *in the population* is selected must be some function of P_i,

or equivalently

$$\mathrm{P}(i\text{th study selected}) \ = a(y_i) \tag{7}$$

for some function $a(y)$. Of course the precise form of this function is unknown. The aim of this paper is to suggest how (4) can be replaced by a *robust* P-value, robust in the sense that it remains valid for *all possible* selection functions $a(y)$.

2. ROBUST P-VALUES

2.1. A permutation P-value

If the selection of studies is made according to (7), then under H_0 the values of y_i for those studies which are selected will take the form of a random sample from the distribution with probability density function

$$f(y) = \frac{a(y)\phi(y)}{\int a(y)\phi(y)dy},$$

where ϕ is the density of the standard normal distribution. A basic property of random samples is that if we change the order of their values, their joint distribution remains exactly the same. Hence, under H_0, each member $Y = (Y_1, Y_2, \cdots, Y_k)$ of the permutation set

$$\mathcal{S} = \{Y | Y \text{ is a permutation of } y_1, y_2, \cdots, y_k\} \tag{8}$$

is equally likely. But each rearrangement of the order of the components of $y_1, y_2, \cdots, y_k$, holding the observed values $x_1, x_2, \cdots, x_k$ fixed, gives its own treatment estimate

$$\tilde{\theta}(Y) = \frac{\sum Y_i x_i}{\sum x_i^2}.$$

For a P-value we need the probability under H_0 that $\tilde{\theta}$ will be greater than or equal to that observed. This gives the permutation P-value

$$\tilde{P} = \mathrm{P}\{\tilde{\theta}(Y) \geq \tilde{\theta}(y)|H_0, Y \in \mathcal{S}\} = \mathrm{P}\{\sum \alpha_i Y_i \geq \sum \alpha_i y_i|H_0, Y \in \mathcal{S}\}, \tag{9}$$

where $\alpha_i = x_i - \bar{x}$.

Since the values of α_i are known, (9) can be calculated directly by evaluating $\sum \alpha_i Y_i$ for all $k!$ permutations of the observed values of y_i. The permutation P-value is the proportion of these permutations for which $\sum \alpha_i Y_i$ equals or exceeds the value for the observed values y_i. If k is large this calculation will be prohibitive, but we can replace complete enumeration in the obvious way by sampling random permutations of the y_is. With a sufficiently large sample we can estimate $\tilde{P}$ to arbitrarily high accuracy.

2.2. An approximate P-value

Provided the α_is are not too extreme the distribution of $\sum \alpha_i Y_i$ will be approximately normal, and so we can approximate $\tilde{P}$ in terms of the mean and variance of this linear sum. For this we need the means, variances and covariances of the components of a randomly selected permutation from (8). These follow using the methods of classical finite sampling theory, see for instance Cochran [5].

Let Y be a randomly chosen element of $\mathcal{S}$. Then for any fixed i, the ith element Y_i is equally likely to take any of the k values $y_1, y_2, \cdots, y_k$, and so

$$E(Y_i) = \frac{1}{k}\sum_a y_a = \bar{y}, \quad Var(Y_i) = \frac{1}{k}\sum_a (y_a - \bar{y})^2 = s_y^2,$$

say. Similarly, for any fixed pair i, j $(i \neq j)$, (Y_i, Y_j) is equally likely to be any of the $k(k-1)$

distinct (ordered) pairs (y_a, y_b). Thus

$$\begin{aligned} Cov(Y_i, Y_j) &= \frac{1}{k(k-1)} \sum_{a \neq b} (y_a - \bar{y})(y_b - \bar{y}) \\ &= \frac{1}{k(k-1)} \left[\{ \sum_a (y_a - \bar{y}) \}^2 - \sum_a (y_a - \bar{y})^2 \right] \\ &= -\frac{s_y^2}{k-1}. \end{aligned}$$

Hence

$$E(\sum \alpha_i Y_i) = \sum \alpha_i E(Y_i) = \sum \alpha_i \bar{y} = 0,$$

as $\sum \alpha_i = 0$. Similarly

$$Var(\sum \alpha_i Y_i) = \sum \alpha_i^2 Var(Y_i) + \sum_{i \neq j} \alpha_i \alpha_j Cov(Y_i, Y_j) = \frac{k^2}{k-1} s_y^2 s_x^2,$$

where

$$s_x^2 = \frac{1}{k} \sum \alpha_i^2 = \frac{1}{k} \sum (x_i - \bar{x})^2.$$

The normal approximation for $\sum \alpha_i Y_i$ under H_0 is therefore

$$\sum \alpha_i Y_i \sim N\left(0, \frac{k^2 s_y^2 s_x^2}{k-1}\right).$$

This gives the P-value

$$\mathrm{P}(\sum \alpha_i Y_i \geq \sum \alpha_i y_i | H_0, Y \in \mathcal{S}) \simeq \Phi\left(-\frac{(k-1)^{\frac{1}{2}} \sum \alpha_i y_i}{k s_y s_x}\right).$$

Thus $\tilde{P}$ in (9) is approximated by

$$\hat{P} = \Phi(-(k-1)^{\frac{1}{2}} r) \tag{10}$$

where

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{k s_x s_y}.$$

The approximate robust test takes a very simple form: all we have to do is calculate the sample correlation r of the points of the radial plot to give $\hat{P}$ in (10).

2.3. Least squares on the radial plot

Testing the size of the correlation r by $\hat{P}$ is like the usual test of significance in linear regression. If we have k observed pairs (x_i, y_i) from a standard linear regression model, significance of the dependence of y on x is judged by referring the standardized regression coefficient (least squares slope divided by its estimated standard error) to t_{k-2}, the t-distribution on $(k-2)$ degrees of freedom. This can be expressed in terms of the sample correlation r as

$$T = \left(\frac{(k-2)r^2}{1-r^2} \right)^{\frac{1}{2}} \sim t_{k-2}. \tag{11}$$

This gives the (one-tailed) P-value

$$P_{reg} = F_{k-2} \left\{ - \left[\frac{k-2}{(k-1)(1-r^2)} \right]^{\frac{1}{2}} (k-1)^{\frac{1}{2}} r \right\}, \tag{12}$$

where F_{k-2} is the cumulative distribution function of t_{k-2}. If r is relatively large, both $\hat{P}$ in (10) and P_{reg} in (12) are small. If r is sufficiently small for significance to be in doubt, the term in square brackets in (12) will be close to one and so the two P-values will be similar. More carefully, suppose that r is such that $\hat{P}$ is just significant at (one-tailed) level α. Then

$$\hat{P} = \Phi(-z_\alpha) \quad \text{and} \quad P_{reg} = F_{k-2} \left\{ - \left[\frac{k-2}{k-1-z_\alpha^2} \right]^{\frac{1}{2}} z_\alpha \right\}, \tag{13}$$

where z_α is the $100(1-\alpha)$ standard normal percentage point. For realistic values of k and α, the term in square brackets in (13) will be slightly bigger than one, offset by the fact that the t-distribution has longer tails than the normal. In practice $\hat{P}$ and P_{reg} are very similar.

These P-values are testing for zero slope in the regression of y on x. If the standard fixed effects model (6) is correct, then the intercept of this regression is also zero. This can also be tested in the usual way, giving P-value P_E, say. This is the P-value for the well-known Egger test for publication bias (Egger *et al.* [6]). The argument behind the Egger test is that if studies are selected in accordance with a selection function of the type (7), then the residual

variation of y around θx in (6) will be biased by different amounts depending on the values of x. This distortion will show up as an apparent non-zero intercept in the radial plot.

This suggests a simple interpretation of least squares on the radial plot and the two P-values P_E and P_{reg} routinely calculated by regression software. The intercept P-value P_E is testing for the presence of publication bias. The regression P-value P_{reg} is testing the treatment effect allowing for the possible presence of publication bias. If we *assume* there is no publication bias then we refit the regression line with the constraint that it passes through the origin, the new (single) P-value now testing for the treatment effect as in standard meta analysis. If we constrain this further to have residual variance equal to one, then we get the standard fixed effect P-value P in (4).

3. THE PRICE OF ROBUSTNESS: LOSS OF POWER

The conventional inference about θ given by (2) makes a very strong assumption about selection, which we call the *randomization assumption*: $a(y)$ in (7) is a constant, not depending on y. Relaxing this to the very much weaker assumption that selection merely depends on y (with $a(y)$ taking any form) inevitably leads to a loss of precision. There is no free lunch.

Under the fixed effects model, the usual test statistic for testing H_0 is

$$T_1 = \tilde{\theta}(\sum x_i^2)^{\frac{1}{2}}.$$

The test statistic corresponding to the approximate robust P-value $\hat{P}$ in (10) is

$$T_2 = r(k-1)^{\frac{1}{2}}.$$

These P-values are calculated on the assumption that T_1 and T_2 are standard normal under H_0. The power functions of the corresponding (one-tailed) level-α tests are the values when

$\theta \neq 0$ of the probabilities

$$A_1 = P(T_1 > z_\alpha) \quad \text{and} \quad A_2 = P(T_2 > z_\alpha).$$

Since T_1 remains normal even if $\theta \neq 0$, we have

$$A_1 = P(T_1 > z_\alpha) = \Phi\{-z_\alpha + \theta(\sum x_i^2)^{\frac{1}{2}}\}. \tag{14}$$

For the test based on r, we use standard linear regression theory as before, which shows that the test statistic T in (11) is now distributed as $t_{(k-2,\nu)}$, the non-central t-distribution on $(k-2)$ degrees of freedom and non-centrality parameter

$$\nu = \theta\{\sum (x_i - \bar{x})^2\}^{\frac{1}{2}}. \tag{15}$$

The power function for T_2 is thus the non-central extension of (13), namely

$$A_2 = P(T_2 > z_\alpha) = F_{(k-2,-\nu)}\left\{-\left(\frac{k-2}{k-1-z_\alpha^2}\right)^{\frac{1}{2}} z_\alpha\right\}, \tag{16}$$

where $F_{(k-2,-\nu)}$ is now the cumulative distribution function of $t_{(k-2,-\nu)}$. (The change of sign of the non-centrality parameter reflects the fact that we are interested in the right tail rather than the left tail of the distribution).

Let γ be the coefficient of variation of the observed x_is, namely $\gamma = s_x/\bar{x}$. Then from (14) and (15) we get

$$\nu = \left(\frac{\gamma^2}{1+\gamma^2}\right)^{\frac{1}{2}} \{z_\alpha + \Phi^{-1}(A_1)\}. \tag{17}$$

Substituting (17) into (16) gives the corresponding value of A_2 as a function of A_1.

The comparison between these two power functions is illustrated in Figure 1 which plots A_2 against A_1 for $k = 30$ and several different values of γ. When γ is small the loss of power is very substantial, reflecting the fact that the accuracy of the least squares slope in linear regression depends on the spread of the x values. An extreme case is if all the studies have

the same variance ($\gamma = 0$), when the regression slope is indeterminate despite the fact that $\tilde{\theta}$, the slope through the origin, may be quite accurately determined. If γ is large, which means that the observed values of x are heavily skewed to the right (a majority of small studies with a minority of much larger studies) the loss of power is much less. Plots for other values of k show a very similar pattern. Loss of power improves slightly as k becomes larger, but the size of γ remains the dominant factor.

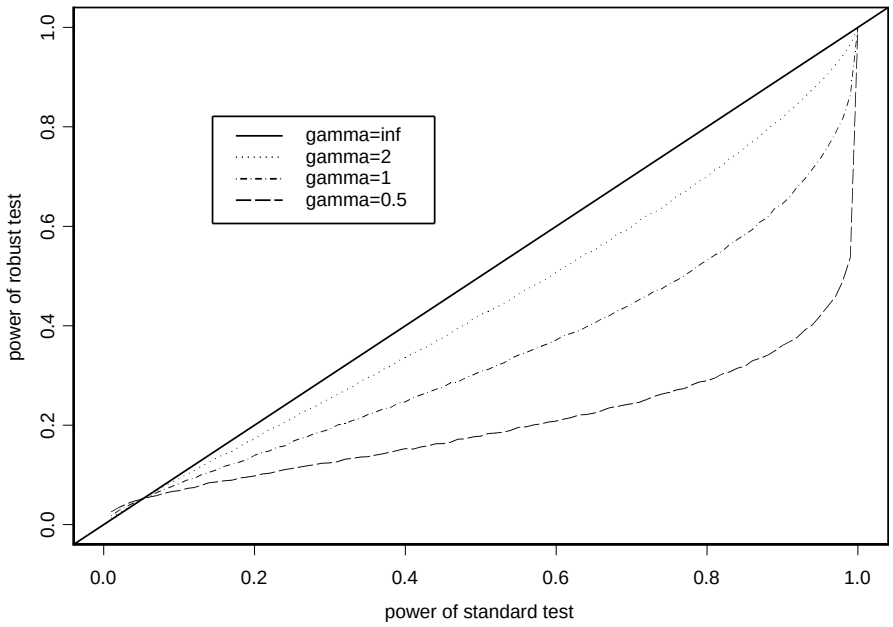

Figure 1. Power of robust test against power of standard test

4. ASSUMPTIONS ABOUT SELECTION

In common with all missing data problems, inference is impossible unless *some* assumptions are made about the underlying selection mechanism, assumptions which are intrinsically unverifiable. The proposed robust procedure is non-parametric in that it assumes nothing about the function $a(y)$ in (7), but modelling selection in terms of such a function is itself a strong assumption. In reality, selection in meta analysis (writing up, publishing, being accessible to the reviewer, satisfying the inclusion criteria for the systematic review) may depend on many factors for which the available data can only act as proxy. In this sense, selection could depend in some arbitrary way on both x and y. Unfortunately, even moderately flexible parametric models for such joint dependence are almost impossible to fit.

Henmi *et al.* [7] generalize (7) to an arbitray function $a(x, y)$ (or equivalently an arbitrary function $a(\hat{\theta}, \sigma)$) and develop a non-parametric "worst case" sensitivity analysis which controls on the marginal proportion of papers that are selected. In a related series of papers (Copas [8]; Copas and Shi [9, 10]; Shi and Copas [11]) a parametric sensitivity analysis is developed in which selection is modelled in terms of a latent propensity score which may be correlated with study outcome. Again inference is only possible if we control on aspects of the marginal selection process (marginal to study outcome but conditional on x). Other parametric approaches, such as modelling selection as a step function in y, are reviewed in Hedges and Vevea [4].

Because assumptions about selection in meta analysis can never be empirically verified, no single analysis can be taken as definitive, but rather as part of a sensitivity analysis. The sensitivity analysis can be implicit, by trying different assumptions about selection and comparing results, or explicit, by controlling on sensitivity parameters such in the methods

just reviewed. For example, we can compare the two P-values P in (4) and $\hat{P}$ in (10). If γ, the coefficient of variation of the x_is, is reasonably large, P and $\hat{P}$ can be similar, suggesting that assumptions about selection are not critical. However, if γ is small, $\hat{P}$ is usually much larger (less significant) than P, indicating that assumptions about selection are then crucially important. In some meta analysis problems, a significant treatment effect cannot be established unless one asserts as a statement of belief that the randomization assumption is valid, *i.e.* that the review is free of publication bias.

This discussion ignores the fact that the data do give us *some* information about selection. If we assume that selection is given by (7), and observe a marked non-zero intercept in the radial plot(or equivalently, clear asymmetry in the funnel plot), then we have evidence that $a(y)$ cannot be constant as required by the randomization assumption. The robust P-value $\hat{P}$ would then give a more appropriate indication of treatment effect than the naive (and misleading) P-value P. However, considerable caution is needed if this argument is used in reverse, *i.e.* if we are tempted to assume that a non-significant intercept of the radial means that there is no publication bias and hence that P is valid. Several papers have pointed out that the power of the Egger test (testing the intercept) for detecting publication bias can be disappointingly low (for example Schwarzer *et al.* [12]). For realistic values of k, levels of selection (dependence of $a(y)$ on y) which are sufficiently strong to lead to substantial bias have only a moderate chance of being detected from the radial (or funnel) plot.

5. RANDOM EFFECTS

Model (1) is based on the fixed effects assumption, that each study is estimating the same underlying treatment effect θ. When there is heterogeneity between studies (or heterogeneity

is suspected) the standard procedure is to fit the random effects model

$$\hat{\theta}_i \sim N(\theta, \sigma_i^2 + \tau^2), \tag{18}$$

where τ^2 is the random effects variance. This model is fitted in exactly the same way as the fixed effects model (when $\tau^2 = 0$) by replacing the weights w_i in (2) by

$$w_i = \frac{1}{\sigma_i^2 + \tau^2}. \tag{19}$$

In practice τ^2 is estimated from the data, usually using the method of DerSimonian and Laird [13].

The corresponding redefinition of the radial plot coordinates replaces (5) by

$$y_i = \frac{\hat{\theta}_i}{\sqrt{(\sigma_i^2 + \tau^2)}}, \quad x_i = \frac{1}{\sqrt{(\sigma_i^2 + \tau^2)}}. \tag{20}$$

The model (6) remains valid, as does the robust test proposed in Section 2. However, the selection model (7) now has a different interpretation, since the values of y_i are no longer a simple transformation of the within-study P-values (3). If τ^2 is small then selection on the basis of the new y_is is more or less equivalent to selection on the basis of the original y_is. But if τ^2 is large the two rank orders of the studies implied by the two versions of the y_is could be quite different, and so we lose the original intuition behind (7).

6. EXAMPLES

6.1. Passive smoking

Hackshaw *et al.* [14] reviewed 37 epidemiological studies of the lung cancer risk of passive smoking. The parameter being estimated is the log relative risk associated with prolonged exposure of non-smokers to environmental tobacco smoke; see the cited paper for an extended

discussion of these studies and potential sources of bias. Taking the data from Table 1 of Hackshaw *et al.* [14] gives the radial plot in Figure 2. The residual sum of squares from the line $y = \tilde{\theta}x$ is 48.0 on 36 degrees of freedom, suggesting that the fixed effect model being assumed here is reasonable.

For these data, $\tilde{\theta} = 0.184$ with standard error 0.038, implying very strong evidence of an increased risk. However, the intercept test (Egger test) gives $P_E = 0.021$, raising doubts about the selection of these studies and hence doubts about the validity of the conventional analysis. To implement the robust procedure of Section 2.1, the histogram in Figure 3 shows the distribution of $\sum \alpha_i Y_i$ for 10,000 random permutations of the y_i's, to be compared with the observed value $\sum \alpha_i y_i = 0.311$. This gives $\tilde{P} = 0.495$.

The similarity of Figure 3 to a normal distribution is striking. The correlation of the radial plot is $r = 0.0038$ giving $\hat{P} = 0.491$, almost the same as $\tilde{P}$. To this accuracy, this is exactly the same as P_{reg}, the P-value for significance of the least squares slope in (12). We see that allowing for selection functions of the form (7) has completely destroyed the evidence in these data. This is already evident in the radial plot (Figure 2): the regression through the origin has a clear positive slope, but if the origin constraint is removed the evidence for an upwards trend all but disappears.

This meta analysis has been extensively discussed in the literature, and several authors have noted the sensitivity of the estimated relative risk to selection bias. Henmi *et al.* [7] suggest how the usual P-value P can be extended to allow for selection of a given proportion of studies. They maintain that if 70% or more of comparable studies are included, then the evidence for risk can still be regarded as significant at the 5% level, but the evidence is no longer conclusive if more than 30% of eligible studies are selected out. It is not surprising that making only the

weaker assumption (7) also fails to give a significant result. Even if there were no selection bias, the price for making the inference robust to (7) in this example would still be high, as the coefficient of variation of the x_is is only $\gamma = 0.54$ implying a power curve near the lowest of those shown in Figure 1.

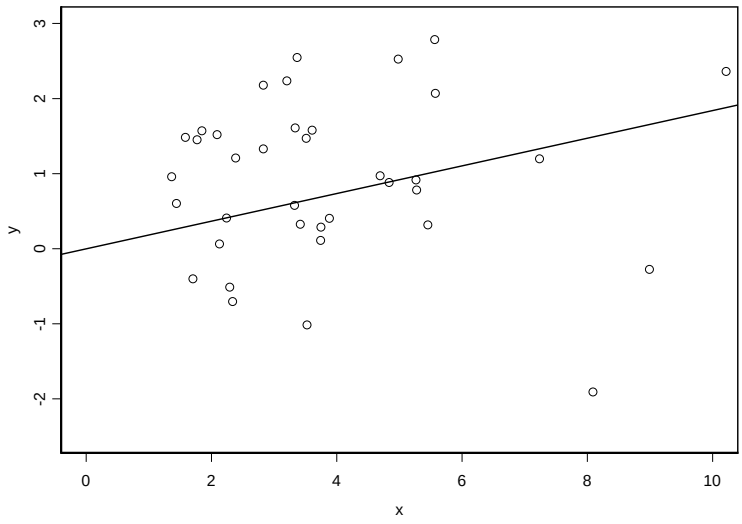

Figure 2. Radial plot of passive smoking studies

6.2. Intravenous streptokenase

The text on systematic reviews by Egger *et al.* [15] discusses a meta analysis of 22 clinical trials on the effectiveness of intravenous streptokinese following myodardial infarction (the data are on page 349). The radial plot for the mortality log-odds ratios observed in these studies is

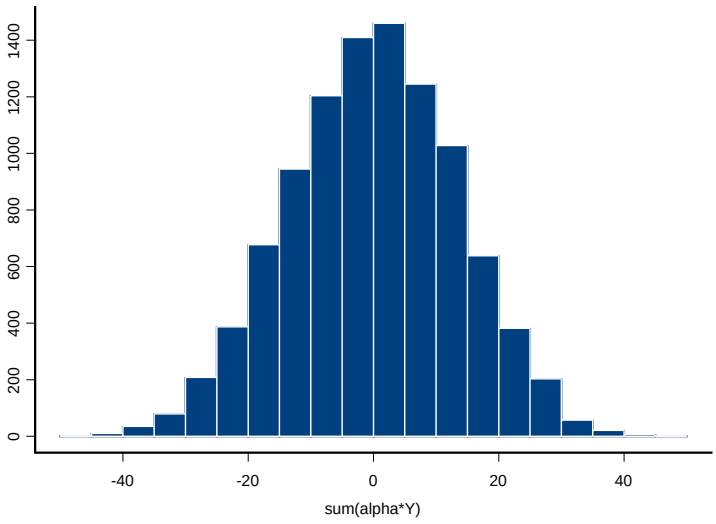

Figure 3. Permutation distribution of $\sum(\alpha_i Y_i)$

shown in Figure 4. Here the residual sum of squares from the line $y = \tilde{\theta}x$ is 31.5 on 21 degrees of freedom, again suggesting that the fixed effects model is reasonable. But in this example there is no evidence for selection bias, the least squares intercept is 0.117 with standard error 0.353, giving $P_E = 0.74$.

The standard analysis gives $\tilde{\theta} = -0.255$ with standard error 0.033, very clear evidence for a beneficial treatment effect. Simulating the random permutations in (9) gives $\tilde{P} = 0.00027$, again highly significant but less extreme than the conventional analysis (which gives the unbelievably small $P \approx 10^{-14}$). We get the same conclusion from the normal approximation in Section 2.2: the correlation in the radial plot is $r = -0.737$ giving $\hat{P} = 0.00037$. For these data $\gamma = 1.10$, so referring to Figure 1 we see that there is still a loss of power but not to the same extent as in the first example.

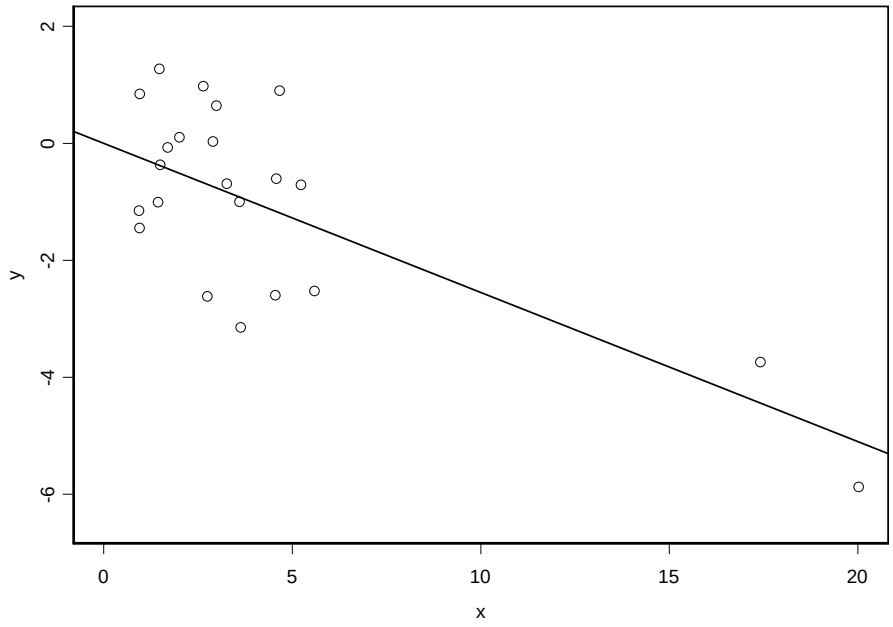

Figure 4. Radial plot of streptokinase studies

Prepared using simauth.cls

Acknowledgements

This research is supported by a grant from the *EPSRC* : Grant No. GR/S67807/01

REFERENCES

1. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-analysis in Medical Research.* John Wiley & Sons, Ltd: Chichester, England, 2000.
2. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine* 1988; **7**(8):889–894.
3. Rothstein HR, Sutton AJ, Borenstein M. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments.* John Wiley & Sons, Ltd: Chicester, England, 2005.
4. Hedges LV, Vevea J. Chapter 9: Selection Method Approaches in *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments.* eds Rothstein HR, Sutton AJ, Borenstein M. John Wiley & Sons, Ltd: Chicester, England, 2005.
5. Cochran WG. *Sampling Techniques.* John Wiley & Sons, Ltd: New York, USA, 1953.
6. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997; **315**(7109):629–634.
7. Henmi M, Copas JB, Eguchi S. Confidence intervals and P-values for meta analysis with publication bias. *Biometrics* 2007; **63**(2):475–482.
8. Copas J. What works? Selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Series A* 1999; **162**(1):95–109.
9. Copas JB, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* 2000; **1**(3):247–262.
10. Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research* 2001; **10**(4):251–265.
11. Shi JQ, Copas JB. Publication bias and meta analysis for 2×2 tables: an average Markov chain Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* 2002; **64**(2):221–236.
12. Schwarzer G, Antes G, Shumacher M. Inflation of Type I error rate in two statistical tests for the detection of publication bias in meta-analysis with binary outcomes. *Statistics in Medicine* 2002; **21**(17):2465–2477.

13. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.

14. Hackshaw AK, Law MR, Wald NJ. The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical Journal* 1997; **315**(7114):980–988.

15. Egger M, Davey Smith G, Altman DG. *Systematic Reviews in Health Care: Meta-analysis in context.* BMJ Publishing Group: London, 2001.