CHAPTER 11

Software for Publication Bias

Michael Borenstein
Biostat, Inc., USA

KEY POINTS

- Various procedures for addressing publication bias are discussed elsewhere in this volume. The goal of this chapter is to show how these different procedures fit into an overall strategy for addressing bias, and to discuss computer programs that can be used to implement this strategy.
- To address publication bias the researcher should proceed through a logical sequence of analyses. First, forest plots and funnel plots provide a visual sense of the data. Then, rank correlation and regression procedures ask whether or not there is evidence of bias, while the failsafe *N* and its variants ask if the *entire* effect size may be attributed to bias. Finally, trim and fill and the cumulative forest plot offer a more nuanced perspective, and ask how the effect size would *shift* if the apparent bias were to be removed.
- The computer programs discussed are Comprehensive Meta Analysis, Stata, MetaWin, and RevMan. We show how the researcher would address publication bias using each of these programs in turn. The chapter closes with a subjective assessment of each program's strengths and weaknesses.
- The appendix includes a list of additional resources, including web links to data sets and to step-by-step instructions for running the publication bias procedures with each of the programs.

INTRODUCTION

This chapter discusses computer programs for meta-analysis and publication bias. The author of this chapter is also one of the developers of Comprehensive Meta Analysis (CMA), one of the programs discussed below. While I have endeavored to make this chapter as accurate as possible, the chapter will reflect the same sensibilities that shaped the development of CMA, and as such is not entirely objective. Additional detail on this potential conflict of interest, and also of steps taken to ensure accuracy, are included at the end of the chapter.

Where other chapters in this volume discuss the rationale and interpretation of specific procedures for assessing publication bias, the goal of this chapter is to present an overall strategy for working with these different procedures, and then to discuss computer programs that can be used to implement this strategy. The overview that follows, and then the discussion of each program in turn, proceeds through a logical sequence of analyses: Forest plots and funnel plots provide a visual sense of the data. Rank correlation and regression procedures can test for the presence of bias. The failsafe *N* and its variants ask if we need to be concerned that the *entire* observed effect may be an artifact of bias. Finally, trim and fill and the cumulative forest plot offer a more nuanced perspective, and ask how the treatment effect (or other effect size) would *shift* if the apparent bias were to be removed.

Step-by-step instructions for each of the programs, and copies of the data sets, are available on the book's website, http://www.meta-analysis.com/publication-bias.

STATISTICAL PROCEDURES

Getting a sense of the data

Forest plots
A key element in any meta-analysis is the forest plot, which serves as the visual representation of the data. In this plot each study as well as the combined effect is depicted as a point estimate bounded by its confidence interval. The plot, as suggested by its appellation, allows the researcher to see both the forest and the trees. It shows if the overall effect is based on many studies or a few; on studies that are precise or imprecise; whether the effects for all studies tend to line up in a row, or whether they vary substantially from one study to the next. The plot puts a face on the statistics, helping to ensure that they will be interpreted properly, and highlighting anomalies such as outliers that require attention.

While the forest plot is more closely associated with the core meta-analysis than with publication bias, an examination of this plot is a logical first step in any analysis. Before turning to the funnel plot or statistical tests to look for bias, the researcher should study the forest plot to get a sense of the data.

Funnel plots
The funnel plot, in its traditional form, is a plot of a measure of study size on the vertical axis as a function of effect size on the horizontal axis (Chapter 5). Large studies appear toward the top of the graph, and tend to cluster near the mean

effect size. Smaller studies appear toward the bottom of the graph, and (since there is more sampling variation in effect size estimates in the smaller studies) will be dispersed across a range of values. This pattern tends to resemble a funnel, which is the basis for the plot's name (Light and Pillemer, 1984; Light *et al.*, 1994).

In the absence of publication bias, the studies will be distributed symmetrically about the combined effect size. By contrast, in the presence of bias, we would expect that the bottom of the plot would show a higher concentration of studies on one side of the mean than on the other. This would reflect the fact that smaller studies (which appear toward the bottom) are more likely to be published if they have larger than average effects, which makes them more likely to meet the criterion for statistical significance.

Sterne, Becker and Egger (Chapter 5) explain that the selection of an effect index (e.g., the odds ratio or the risk difference) will have an impact on the symmetry of the funnel plot. Therefore, the selection of an index should reflect the fundamental questions being asked in the analysis. They note also that the index used to represent study size will affect the way studies are dispersed on the plot, and that the proper index will facilitate the researcher's ability to detect bias. They recommend the use of the standard error (rather than its inverse) for this purpose. Finally, they suggest that it may be helpful to superimpose guidelines on the funnel plot to show the expected distribution of studies in the absence of bias. These guidelines can help to identify outliers and facilitate the process of detecting asymmetry (see also Egger *et al.*, 1997; Sterne and Egger, 2001).

Is there evidence of bias?

The funnel plot offers a visual sense of the relationship between effect size and precision, but the interpretation of the plot is largely subjective. Two tests are commonly used in an attempt to quantify the amount of bias captured by the funnel plot.

Begg and Mazumdar's rank correlation test

Begg and Mazumdar's rank correlation test reports the rank correlation (Kendall's tau) between the standardized effect size and the variances (or standard errors) of these effects. Tau would be interpreted much the same way as any correlation, with a value of zero indicating no relationship between effect size and precision, and deviations from zero indicating the presence of a relationship (Begg and Berlin, 1988; Begg and Mazumdar, 1994; Begg, 1994).

If asymmetry is caused by publication bias we would expect that high standard errors (small studies) would be associated with larger effect sizes. If larger effects are represented by low values (e.g., odds ratio for preventing lung cancer) tau would be positive, while if larger effects are represented by high values (e.g., risk ratio for passive smoking) tau would be negative. Since asymmetry could appear in the reverse direction however, the significance test is two-sided. (Kendall's tau should not be confused with the convention of using τ^2 to denote a variance component in random-effect meta-analysis models. There is no relationship between the two.)

Sterne and Egger (Chapter 6) caution against using the test unless the meta-analysis includes a range of study sizes, including at least one of 'medium' size. Otherwise, the result will be driven primarily by noise. They also note that the test has low power unless there is severe bias, and so a non-significant tau should not be taken as proof that bias is absent (see also Sterne *et al.*, 2000, 2001b, c).

Egger's regression

Egger's linear regression method (Egger *et al.*, 1997; Sterne *et al.*, 2001b, c), like the rank correlation test, is intended to quantify the bias captured by the funnel plot. This differs from Begg and Mazumdar's test in that Egger uses the actual values of the effect sizes and their precision, rather than ranks. In the Egger test, the standard normal deviate is regressed on precision, defined as the inverse of the standard error. The intercept in this regression corresponds to the slope in a weighted regression of the effect size on the standard error. As was true for the rank correlation test, the significance test should be two-tailed.

Sterne and Egger (Chapter 6) discuss this method in more detail. They report that power for this test is generally higher than power for the rank correlation method, but is still low unless there is severe bias or a substantial number of studies. As is true of all significance tests, the point estimate and confidence interval are more informative, and less likely to be misinterpreted, than the significance test. As was true for the rank correlation method, the Egger test should only be used if the analysis includes a range of study sizes and at least one of 'medium' size.

Can we be confident that the observed effect is not entirely an artifact of bias?

The failsafe *N*

Rosenthal suggested that we could compute the number of missing studies (with mean effect of zero) that would need to be added to the analysis before the combined effect would no longer be statistically significant. Rosenthal referred to this as a 'File drawer' analysis (this being the presumed location of the missing studies), and Harris Cooper (1979) suggested that the number of missing studies needed to nullify the effect should be called the 'failsafe *N*' (Rosenthal, 1979; Begg, 1994).

While Rosenthal's work was critical in focusing attention on publication bias, this approach is of limited utility for a number of reasons. First, it focuses on the question of statistical significance rather than clinical significance. That is, it asks how many hidden studies are required to make the effect not *statistically* significant, rather than how many hidden studies are required to reduce the effect to the point that it is not *clinically* important. Second, the formula assumes that the mean effect size in the hidden studies is nil, when in fact it could be negative (which would require fewer studies to nullify the effect) or positive. Third, the failsafe *N* is based on significance tests that combine *p*-values across studies, as was common at the time that Rosenthal suggested the method. Today, the common practice is to compute a *p*-value for the combined effect, and the failsafe algorithm would not work with this approach.

Therefore, as Becker (Chapter 7) suggests, the failsafe N might best be seen as a heuristic which can help to put the question of publication bias in perspective. If someone is concerned that the observed effect is entirely an artifact of publication bias, the statement that 'we would need 5000 studies with an effect size of zero to nullify the observed effect' may help to move the conversation along to more subtle and relevant issues about publication bias (see also Carson *et al.*, 1990).

The algorithm for the failsafe N involves computing a combined *p*-value for all studies in the analysis, and then determining how many additional studies with (an average) z of zero would be required to yield a non-significant *p*-value. The criterion *p*-value for non-significance could be either one-tailed or two-tailed. While this issue has not been addressed in discussions of this method, it seems that the use of a one- or two-tailed criterion for the failsafe N should match the criterion used in the meta-analysis.

Orwin's failsafe N

As noted, two problems with Rosenthal's method are that it focuses on statistical significance rather than clinical significance, and that it assumes that the mean effect size in the missing studies is nil.

Orwin (1983) proposed a variant on the Rosenthal formula which addresses both of these issues. First, Orwin's method allows the researcher to determine how many hidden studies would bring the overall effect to a specified level other than zero. The researcher could therefore select a value that would represent the smallest effect deemed to be clinically important, and ask how many missing studies it would take to bring the combined effect below *this* point. Second, it allows the researcher to specify the mean effect in the hidden studies as being some value other than nil. This would allow the researcher to model a series of other distributions for the missing studies (see Becker, Chapter 7, this volume; Begg, 1994).

What is the effect size after we adjust for bias?

Duval and Tweedie's trim and fill

As discussed above, the key idea behind the funnel plot is that in the absence of bias the plot would be symmetric about the summary effect. If there are more small studies on the right than on the left, our concern is that there may be studies missing from the left. The trim and fill procedure (Chapter 8) builds directly on this idea by imputing the missing studies, adding them to the analysis, and then recomputing the effect size.

Trim and fill uses an iterative procedure to remove the most extreme small studies from the positive side of the funnel plot, recomputing the effect size at each iteration, until the funnel plot is symmetric about the (new) effect size. While this 'trimming' yields the adjusted effect size, it also reduces the variance of the effects, yielding a too narrow confidence interval. Therefore, the algorithm then adds the original studies back into the analysis, and imputes a mirror image for each. This 'fill' has no impact on the point estimate but serves as a correction to the variance (Duval and Tweedie, 1998, 2000a, 2000b).

A major advantage of this approach is that it yields an effect size estimate that is adjusted for the funnel plot asymmetry. That is, rather than ask whether or not *any* bias exists, or whether the *entire* effect can be attributed to bias, it provides an estimate of the effect size after the bias has been taken into account. In many cases the adjusted effect will be essentially similar to the original effect. In other cases the magnitude of the effect size will shift but the core finding (e.g., that the treatment is, or is not, effective) will remain intact. In others, the core finding will be called into question. However, these are precisely the kinds of questions that we should be asking.

Another nice feature of this approach is that it lends itself to an intuitive visual display. The computer programs that incorporate trim and fill are able to create a funnel plot that includes both the observed studies and the imputed studies, so the researcher can see how the effect size shifts when the imputed studies are included. If this shift is trivial, then one can have more confidence that the reported effect is valid.

The problem with this approach is that it relies on a number of assumptions. The key assumption is that the observed asymmetry is due to publication bias rather than a 'small-study effect' (see below). If this assumption is incorrect, the idea of imputing the missing studies cannot be supported. Even if the assumption is correct, the remaining problem is that this procedure, like many of the others, assumes that publication bias in the real world follows a neat pattern, and uses this pattern to identify the number of missing studies. As Duval explains (Chapter 8), the algorithm seems to work well in many circumstances, but not all. Variants on the algorithm have been proposed, with one appearing to be more robust than the others (see Chapter 8).

Cumulative meta-analysis

A cumulative meta-analysis is a meta-analysis run with one study, then repeated with a second study added, then a third, and so on. Similarly, in a cumulative forest plot, the first row shows the effect based on one study, the second row shows the cumulative effect based on two studies, and so on. These procedures have traditionally been used to show shifts in the cumulative weight of the evidence over time. Typically, the studies were sorted chronologically, and the plot showed the effect size based on cumulative data though 1970, 1980, 1990, and so on (see, for example, Lau *et al.*, 1995).

The same mechanism, however, can be used to assess the potential impact of publication bias or of a small-study effect. For this purpose the studies would be sorted in the sequence of largest to smallest (or of most precise to least precise), and a cumulative meta-analysis performed with the addition of each study. If the point estimate has stabilized with the inclusion of the larger studies and does not shift with the addition of smaller studies, then there is no reason to assume that the inclusion of smaller studies had injected a bias (i.e., since it is the smaller studies in which study selection is likely to be greatest). On the other hand, if the point estimate does shift when the smaller studies are added, then there is at least a *prima facie* case for bias, and one would want to investigate the reason for the shift.

This approach also provides an estimate of the effect size based solely on the larger studies. And, even more so than trim and fill, this approach is entirely transparent: We compute the effect based on the larger studies and then determine if and how the effect shifts with the addition of the smaller studies (a clear distinction between larger and smaller studies will not usually exist, but is not needed).

An important caveat

Sterne and Egger (Chapter 6) discuss the fact that while these procedures may detect a relationship between sample size and effect size, they cannot assign a causal mechanism to it. That is, the effect size *may* be larger in small studies because we retrieved a biased sample of the smaller studies, but it is also possible that the effect size *really is* larger in smaller studies – perhaps because the smaller studies used different populations or different protocols than the larger ones. Sterne and Egger use the term 'small-study effect' to capture these and other potential confounds (see also Sterne *et al.* 2001b, c).

This caveat figures into the interpretation of all the procedures, but in different ways. Some procedures are based entirely on a model of *publication* bias and would have no meaning for a 'small-study effect'. These include the failsafe *N* and Orwin (which look for numbers of missing studies) and trim and fill (which imputes missing studies). Other procedures (funnel plot, rank correlation, regression) look for a relationship between sample size and effect size. These will 'work' for a small-study effect as well as for publication bias, but the caveat still applies. If the apparent bias is actually a small-study effect then the larger effect size in the smaller studies reflects legitimate heterogeneity in the effect sizes. Similarly, if the cumulative meta-analysis shows a shift in the effect sizes with the inclusion of smaller studies, this could reflect the fact that these studies used a different population or protocol than the larger ones. It may be possible to identify the nature of these populations or protocols by studying the moderator variables, and this would be a logical next step.

ILLUSTRATIVE EXAMPLE

This chapter uses data from the passive smoking meta-analysis that serves as one of the running examples throughout this book (Hackshaw *et al.*, 1997). The meta-analysis includes data from 37 studies that reported on the relationship between 'second-hand' (passive) smoking and lung cancer. The paper reported that exposure to 'second-hand' smoke increased the risk of lung cancer by about 20 %. There was concern that these 37 studies could have been a biased sample of all studies, since studies which had found a statistically significant relationship would have been more likely to be published than those which had found no (statistically significant) evidence of a relationship. The intent of the publication bias analyses is to assess the likely extent of the problem, and to determine what conclusions can be drawn despite the potential for bias.

In the original paper the data are presented in the form of a risk ratio and confidence interval for each study. Some of the programs will accept the data in this format, but others require that the data be used to compute a point estimate and standard error (or variance), which would then be entered into the program. As is standard practice, all programs run the analyses (both the core meta analyses and the publication bias analyses) using the log of the risk ratio. The analyses in the following examples use the inverse variance formula and a fixed-effect model, but can be extended to other computational options and to the random-effects model.

THE SOFTWARE

Four computer programs are discussed in this chapter. These programs – Comprehensive Meta Analysis, Stata, RevMan Analyses, and Metawin – were included because they are in widespread use for meta-analysis, and offer some of the best options for addressing publication bias. Other programs, including SPSS and SAS, are discussed briefly toward the end of the chapter. This chapter focuses on *what* the programs can do. Detailed instructions for running the analyses are available on the website (see the appendix to this chapter).

Comprehensive Meta Analysis (version 2.0)

Comprehensive Meta Analysis (CMA) is a stand-alone program for meta-analysis. CMA was developed in collaboration with people working in medicine, epidemiology, and the social sciences. This program is able to create a forest plot and a funnel plot, to compute the rank correlation test, Egger's test, the failsafe *N* and Orwin's variant, trim and fill, and to display a cumulative forest plot sorted by precision (Borenstein *et al.*, 2005).

Data entry and conversion
CMA will accept data in close to 100 formats including the number of events and sample size in each group, means and standard deviations, correlations, or point estimates and confidence intervals. The passive smoking data, provided as risk ratios and confidence intervals, can therefore be pasted directly into the program.

Forest plot
Figure 11.1 is a forest plot produced by CMA. The program will display the data using either the *log* relative risk or the relative risk (as in this example). The graphical part of the plot (the point estimate and confidence interval) is displayed toward the center of the screen, with points to the right of 1.0 indicating an increased risk for persons exposed to passive smoking. The overwhelming majority of studies show an increased risk, but the 95 % confidence intervals for most studies include the null value of 1.0, and therefore fail to meet the 0.05 criterion for statistical significance. The last row in the spreadsheet shows the summary data for the fixed-effect model. The risk ratio is 1.204 and the 95 % confidence interval is from 1.120
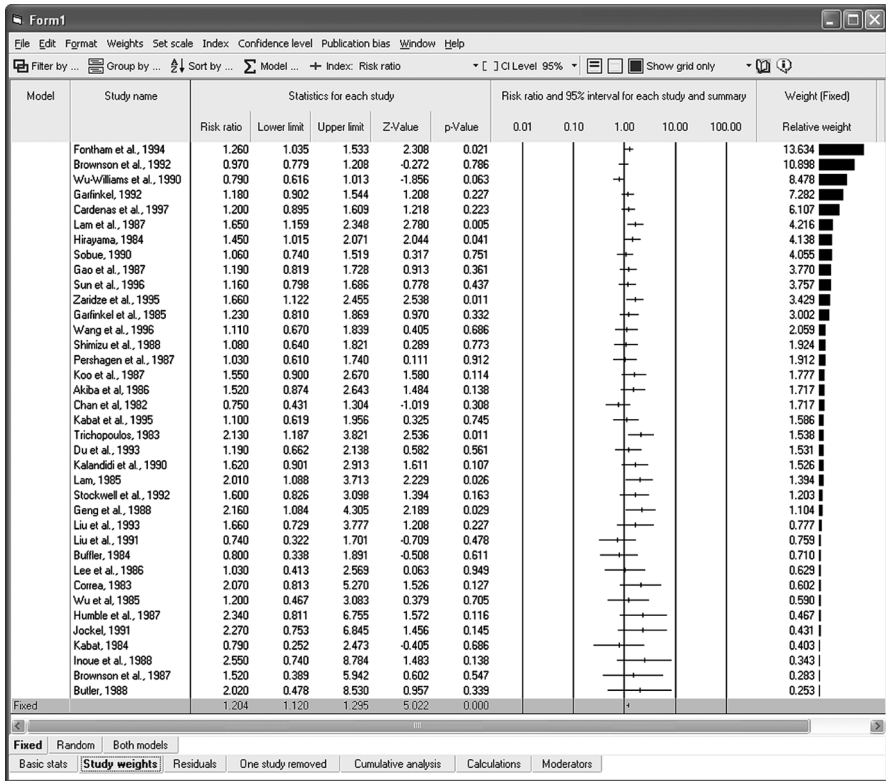
Form1

File Edit Format Weights Set scale Index Confidence level Publication bias Window Help

Filter by ... Group by ... Sort by ... ∑ Model ... + Index: Risk ratio   ▾ [ ] CI Level 95% ▾ ☐ ☐ ■ Show grid only   ▾ 🔲 ⓘ

Model	Study name	Statistics for each study					Risk ratio and 95% interval for each study and summary					Weight (Fixed)
		Risk ratio	Lower limit	Upper limit	Z-Value	p-Value	0.01	0.10	1.00	10.00	100.00	Relative weight
	Fontham et al., 1994	1.260	1.035	1.533	2.308	0.021						13.634
	Brownson et al., 1992	0.970	0.779	1.208	-0.272	0.786						10.898
	Wu-Williams et al., 1990	0.790	0.616	1.013	-1.856	0.063						8.478
	Garfinkel, 1992	1.180	0.902	1.544	1.208	0.227						7.282
	Cardenas et al., 1997	1.200	0.895	1.609	1.218	0.223						6.107
	Lam et al., 1987	1.650	1.159	2.348	2.780	0.005						4.216
	Hirayama, 1984	1.450	1.015	2.071	2.044	0.041						4.138
	Sobue, 1990	1.060	0.740	1.519	0.317	0.751						4.055
	Gao et al., 1987	1.190	0.819	1.728	0.913	0.361						3.770
	Sun et al., 1996	1.160	0.798	1.686	0.778	0.437						3.757
	Zaridze et al., 1995	1.660	1.122	2.455	2.538	0.011						3.429
	Garfinkel et al., 1985	1.230	0.810	1.869	0.970	0.332						3.002
	Wang et al., 1996	1.110	0.670	1.839	0.405	0.686						2.059
	Shimizu et al., 1988	1.080	0.640	1.821	0.289	0.773						1.924
	Pershagen et al., 1987	1.030	0.610	1.740	0.111	0.912						1.912
	Koo et al., 1987	1.550	0.900	2.670	1.580	0.114						1.777
	Akiba et al, 1986	1.520	0.874	2.643	1.484	0.138						1.717
	Chan et al, 1982	0.750	0.431	1.304	-1.019	0.308						1.717
	Kabat et al., 1995	1.100	0.619	1.956	0.325	0.745						1.586
	Trichopoulos, 1983	2.130	1.187	3.821	2.536	0.011						1.538
	Du et al., 1993	1.190	0.662	2.138	0.582	0.561						1.531
	Kalandidi et al., 1990	1.620	0.901	2.913	1.611	0.107						1.526
	Lam, 1985	2.010	1.088	3.713	2.229	0.026						1.394
	Stockwell et al., 1992	1.600	0.826	3.098	1.394	0.163						1.203
	Geng et al., 1988	2.160	1.084	4.305	2.189	0.029						1.104
	Liu et al., 1993	1.660	0.729	3.777	1.208	0.227						0.777
	Liu et al., 1991	0.740	0.322	1.701	-0.709	0.478						0.759
	Buffler, 1984	0.800	0.338	1.891	-0.508	0.611						0.710
	Lee et al., 1986	1.030	0.413	2.569	0.063	0.949						0.629
	Correa, 1983	2.070	0.813	5.270	1.526	0.127						0.602
	Wu et al, 1985	1.200	0.467	3.083	0.379	0.705						0.590
	Humble et al., 1987	2.340	0.811	6.755	1.572	0.116						0.467
	Jockel, 1991	2.270	0.753	6.845	1.456	0.145						0.431
	Kabat, 1984	0.790	0.252	2.473	-0.405	0.686						0.403
	Inoue et al., 1988	2.550	0.740	8.784	1.483	0.138						0.343
	Brownson et al., 1987	1.520	0.389	5.942	0.602	0.547						0.283
	Butler, 1988	2.020	0.478	8.530	0.957	0.339						0.253
Fixed		1.204	1.120	1.295	5.022	0.000						

Fixed | Random | Both models

Basic stats | Study weights | Residuals | One study removed | Cumulative analysis | Calculations | Moderators

Figure 11.1 Forest plot (CMA).

to 1.295. Toward the left, the program shows the name and additional detail for each study. Toward the right, the program shows the relative weight assigned to each study.

In this plot, the studies have been listed from most precise to least precise, so that larger studies appear toward the top and smaller studies appear toward the bottom. This has no impact on the analysis, but allows us to get a sense of the relationship between sample size and effect size. Note that the point estimates become more widely dispersed but also shift toward the right as we move toward the bottom of the plot, where the smaller studies (note the wider confidence intervals) are located. This will be more evident in the funnel plot, which is designed to highlight this relationship.

Funnel plot

CMA will create a funnel plot of any effect size index on the X-axis by either the standard error (Figure 11.2) or precision (not shown) on the Y-axis. The program allows the user to include a vertical line at the summary effect, and guidelines for the 95 % confidence interval. In this example studies at the bottom are clustered
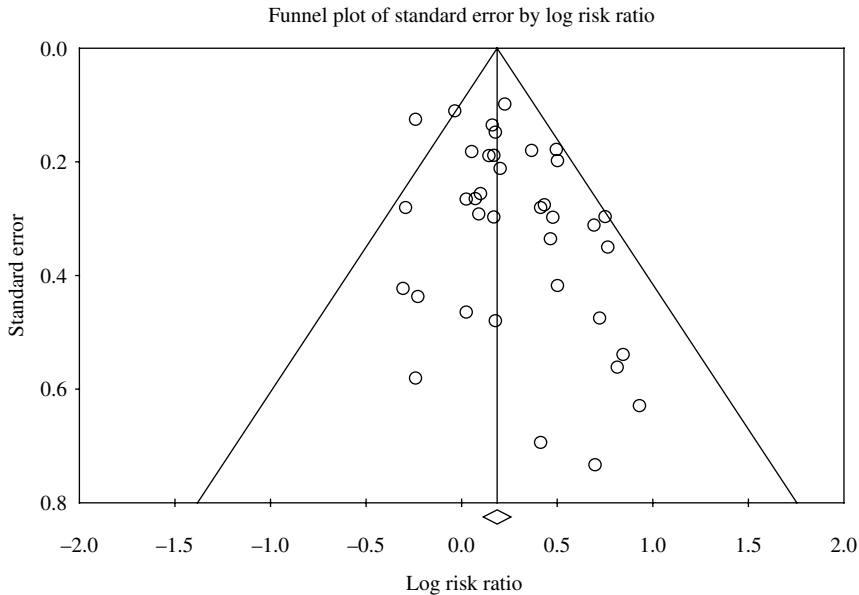
Funnel plot of standard error by log risk ratio

Figure 11.2   Funnel plot (CMA).

toward the right-hand side of the graph, this asymmetry suggesting the possibility of publication bias.

Statistical tests
Results for the various statistical tests are presented in tabular form by the program. For the rank correlation test, Kendall's tau is 0.144 with one-tailed $p = 0.102$. For Egger's test, the intercept (b) is 0.892, with a 95 % confidence interval from 0.127 to 1.657, and a one-tailed p-value of 0.024. The failsafe N is 398 using a one-tailed criterion (or 269 using a two-tailed criterion). Orwin's failsafe N is 103, assuming a mean risk ratio of 1.0 in the missing studies, with a 'trivial' effect defined as a risk ratio of 1.05.

The program creates a detailed text report summarizing all of these analyses. To save space, only the portion of this report which addresses trim and fill is shown in Figure 11.3.

Finally, the program redisplays the funnel plot, taking into account the trim and fill adjustment. Here, the observed studies are shown as open circles, and the observed point estimate in log units is shown as an open diamond at 0.185 (0.113, 0.258), corresponding to a risk ratio of 1.204 (1.120, 1.295). The seven imputed studies are shown as filled circles, and the imputed point estimate in log units is shown as a filled diamond at 0.156 (0.085, 0.227), corresponding to a risk ratio of 1.169 (1.089, 1.254). The 'adjusted' point estimate suggests a lower risk than the original analysis. Perhaps the key point, though, is that the adjusted estimate is fairly close to the original – in this context, a risk ratio of 1.17 has the same substantive implications as a risk ratio of 1.20.

Duval and Tweedie's Trim and Fill

If the meta analysis had captured all the relevant studies we would expect the funnel plot to be symmetric. If the funnel plot is actually asymmetric, with a relatively high number of small studies falling toward the right (representing a large treatment effect) and relatively few falling toward the left, we are concerned that these left-hand studies exist, but are missing from the analysis.

Duval and Tweedie developed a method that allows us to impute these studies. That is, we determine where the missing studies are likely to fall, add them to the analysis, and then recompute the combined effect. The method is known as 'Trim and Fill' as the method initially trims the asymmetric studies from the right side to identify the unbiased effect (in an iterative procedure), and then fills the plot by re-inserting the trimmed studies as well as their imputed counterparts.

In this case the method suggests that 7 studies are missing. Under the fixed effects model the point estimate and 95 % confidence interval for the combined studies is 1.204(1.120, 1.295). Using Trim and Fill the imputed point estimate is 1.169(1.089, 1.254).

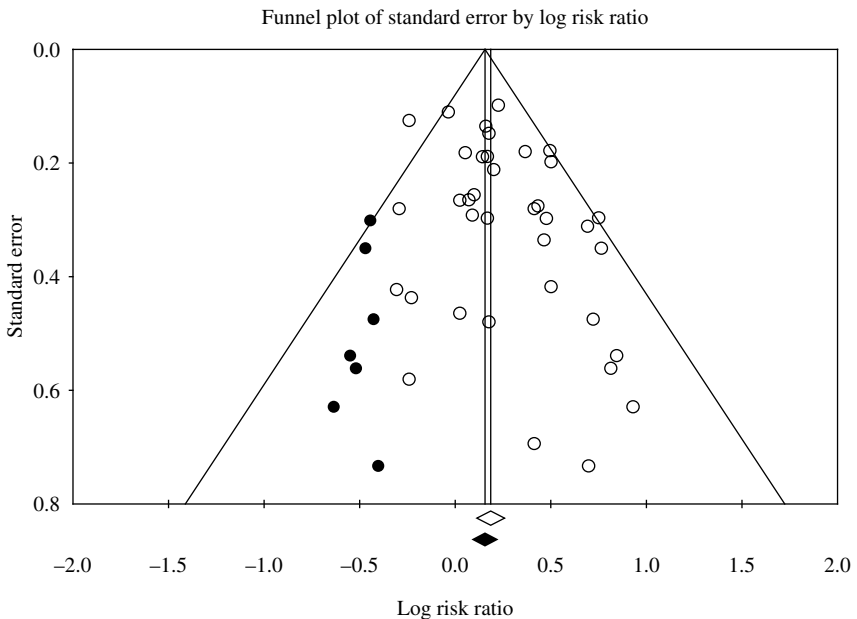**Figure 11.3** Portion of text report created by CMA.

Funnel plot of standard error by log risk ratio

Figure 11.4 Trim and fill (CMA).

Cumulative forest plots

Figure 11.5 is a cumulative forest plot produced by CMA. Note the difference between the cumulative plot and the standard version shown earlier. Here, the
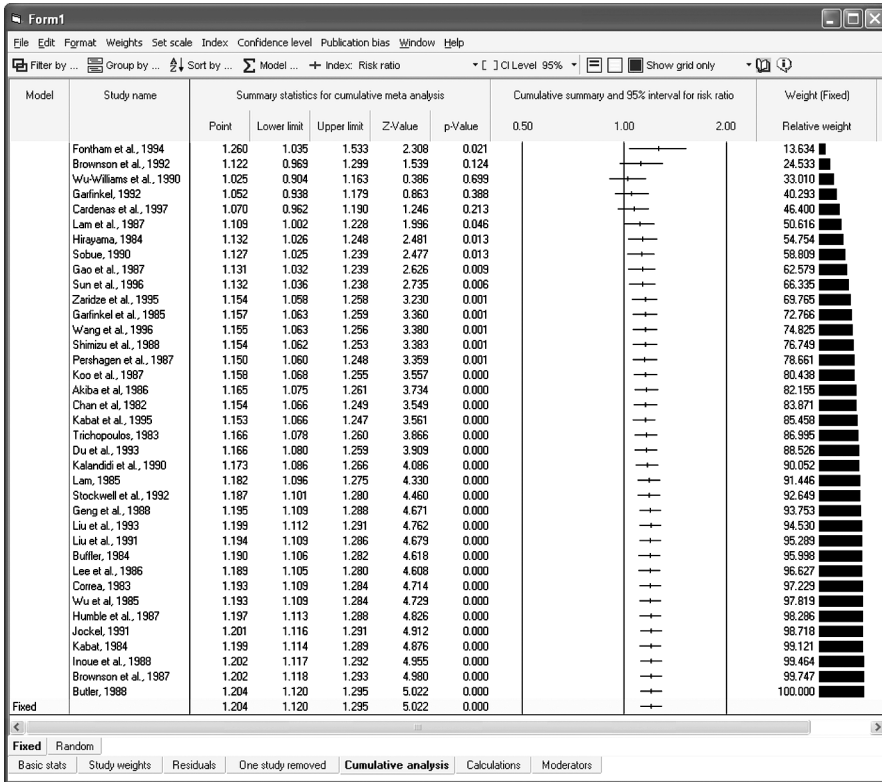
Form1

File Edit Format Weights Set scale Index Confidence level Publication bias Window Help

Filter by ...   Group by ...   Sort by ...   Σ Model ...   + Index: Risk ratio   ▾ [ ] CI Level 95% ▾ ☰ ☐ ▣ Show grid only   ▾ ⬚ ⓘ

Model	Study name	Summary statistics for cumulative meta analysis					Cumulative summary and 95% interval for risk ratio			Weight (Fixed)
		Point	Lower limit	Upper limit	Z-Value	p-Value	0.50	1.00	2.00	Relative weight
	Fontham et al., 1994	1.260	1.035	1.533	2.308	0.021				13.634
	Brownson et al., 1992	1.122	0.969	1.299	1.539	0.124				24.533
	Wu-Williams et al., 1990	1.025	0.904	1.163	0.386	0.699				33.010
	Garfinkel, 1992	1.052	0.938	1.179	0.863	0.388				40.293
	Cardenas et al., 1997	1.070	0.962	1.190	1.246	0.213				46.400
	Lam et al., 1987	1.109	1.002	1.228	1.996	0.046				50.616
	Hirayama, 1984	1.132	1.026	1.248	2.481	0.013				54.754
	Sobue, 1990	1.127	1.025	1.239	2.477	0.013				58.809
	Gao et al., 1987	1.131	1.032	1.239	2.626	0.009				62.579
	Sun et al., 1996	1.132	1.036	1.238	2.735	0.006				66.335
	Zaridze et al., 1995	1.154	1.058	1.258	3.230	0.001				69.765
	Garfinkel et al., 1985	1.157	1.063	1.259	3.360	0.001				72.766
	Wang et al., 1996	1.155	1.063	1.256	3.380	0.001				74.825
	Shimizu et al., 1988	1.154	1.062	1.253	3.383	0.001				76.749
	Pershagen et al., 1987	1.150	1.060	1.248	3.359	0.001				78.661
	Koo et al., 1987	1.158	1.068	1.255	3.557	0.000				80.438
	Akiba et al., 1986	1.165	1.075	1.261	3.734	0.000				82.155
	Chan et al., 1982	1.154	1.066	1.249	3.549	0.000				83.871
	Kabat et al., 1995	1.153	1.066	1.247	3.561	0.000				85.458
	Trichopoulos, 1983	1.166	1.078	1.260	3.866	0.000				86.995
	Du et al., 1993	1.166	1.080	1.259	3.909	0.000				88.526
	Kalandidi et al., 1990	1.173	1.086	1.266	4.086	0.000				90.052
	Lam, 1985	1.182	1.096	1.275	4.330	0.000				91.446
	Stockwell et al., 1992	1.187	1.101	1.280	4.460	0.000				92.649
	Geng et al., 1988	1.195	1.109	1.288	4.671	0.000				93.753
	Liu et al., 1993	1.199	1.112	1.291	4.762	0.000				94.530
	Liu et al., 1991	1.194	1.109	1.286	4.679	0.000				95.289
	Buffler, 1984	1.190	1.106	1.282	4.618	0.000				95.998
	Lee et al., 1986	1.189	1.105	1.280	4.608	0.000				96.627
	Correa, 1983	1.193	1.109	1.284	4.714	0.000				97.229
	Wu et al., 1985	1.193	1.109	1.284	4.729	0.000				97.819
	Humble et al., 1987	1.197	1.113	1.288	4.826	0.000				98.286
	Jockel, 1991	1.201	1.116	1.291	4.912	0.000				98.718
	Kabat, 1984	1.199	1.114	1.289	4.876	0.000				99.121
	Inoue et al., 1988	1.202	1.117	1.292	4.955	0.000				99.464
	Brownson et al., 1987	1.202	1.118	1.293	4.980	0.000				99.747
	Butler, 1988	1.204	1.120	1.295	5.022	0.000				100.000
Fixed		1.204	1.120	1.295	5.022	0.000				

Fixed   Random

Basic stats   Study weights   Residuals   One study removed   **Cumulative analysis**   Calculations   Moderators

Figure 11.5   Cumulative forest plot (CMA).

first row is a 'meta'-analysis based only on the Fontham *et al.* study. The second row is a meta-analysis based on two studies (Fontham *et al.* and Brownson *et al.*), and so on. The last study to be added is Butler, and so the point estimate and confidence interval shown on the line labeled 'Butler' are identical to that shown for the summary effect on the line labeled 'Fixed'.

The studies have been sorted from the most precise to the least precise (roughly corresponding to largest to smallest). With the 18 largest studies in the analysis (inclusive of Chan) the cumulative relative risk is 1.15. With the addition of another 19 (smaller) studies, the point estimate shifts to the right, and the relative risk is 1.20. As such, our estimate of the relative risk has increased. This could be due to publication bias or it could be due to small-study effects. Again, the key point is that even if we had limited the analysis to the larger studies, the relative risk would have been 1.15 (1.07, 1.25) and the clinical implications probably would have been the same.

Note also that the analysis that incorporates all 37 studies assigns nearly 84 % of its weight to the first 18 (see the bar graph in the right-hand column). In other words, if small studies are introducing a bias, we are protected to some extent by the fact that small studies are given less weight.

Stata (version 8.2)

Stata is a general purpose statistical package. While Stata itself does not include routines for meta-analysis, macros written by experts in meta-analysis are freely available (they can be located and downloaded by accessing the web from within the program) and provide a level of functionality comparable to the better stand-alone programs. With these macros in place Stata is able to create a forest plot and a funnel plot, to compute the rank correlation test, Egger's test, trim and fill, and to display a cumulative forest plot sorted by precision. Instructions for downloading, installing, and using the macros are on this book's website.

Data entry and conversion

Since the various macros were developed by different people, they tend to be somewhat idiosyncratic in the formats accepted for effect size data. Most will accept the number of events and sample size in each group, means and standard deviations, a point estimate and confidence interval, or a point estimate and standard error. The passive smoking data can therefore be pasted directly into the program.

Forest plot

Figure 11.6 is a forest plot produced by Stata. The user can elect to display the data using either the *log* relative risk or the relative risk (as in this example). Points to the right of 1.0 indicate an increased risk for persons exposed to passive smoking. The overwhelming majority of studies show an increased risk, but the 95 % confidence
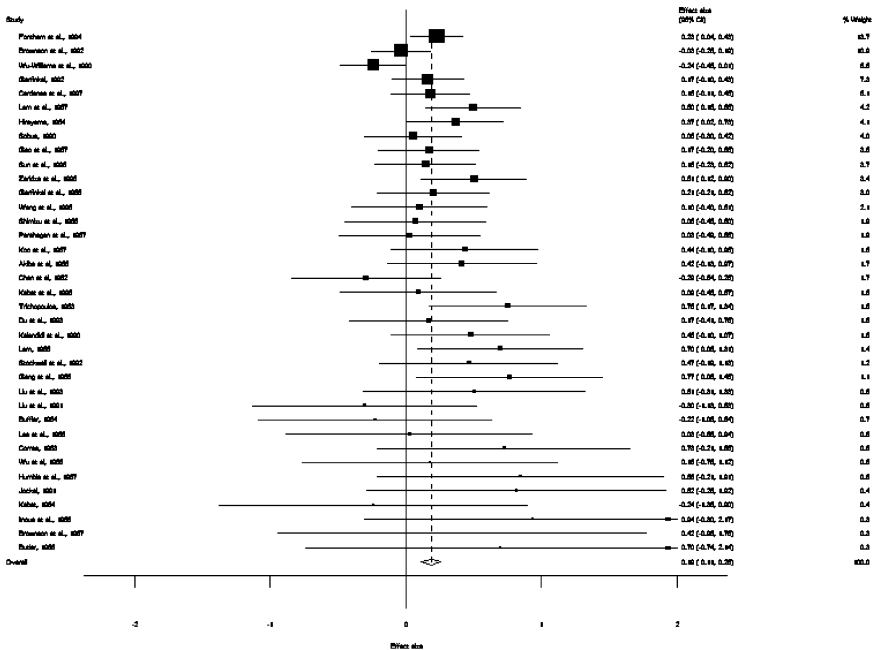
Figure 11.6   Forest plot (Stata).

intervals for most studies include the null value of 1.0, and therefore fail to meet
the criterion for statistical significance. The last row in the spreadsheet shows the
summary data for the fixed-effect model. The risk ratio is 1.204 and the 95 %
confidence interval is from 1.120 to 1.295.

In this plot, the studies have been listed from most precise to least precise, so
that larger studies appear toward the top and smaller studies appear toward the
bottom. This has no impact on the analysis, but allows us to get an initial sense of
the relationship between sample size and effect size. Note that the point estimates
become more widely dispersed but also shift toward the right as we move toward the
bottom of the plot, where the smaller studies (note the wider confidence intervals)
are located. This will be more evident in the funnel plot, which is designed to
highlight this relationship.

Funnel plot

Stata will create a funnel plot of effect size on the *X*-axis by standard error on
the *Y*-axis. In Figure 11.7, where larger studies are plotted at the top and smaller
ones at the bottom, the latter cluster toward the right, suggesting the possibility of
publication bias. There are several options within Stata for creating funnel plots, the
most flexible being the metafunnel macro (Sterne and Harbord, 2004) which was
used to create this figure as well as the figures in Chapter 5. This macro allows the
user to customize the funnel plot by reversing the scale, and by including guidelines
and annotations. It will also display studies from different subgroups in different
colors/symbols, and will include the regression line corresponding to the Egger
test.

For Kendall's test Stata reports a corrected *z*-value of 1.24 and *p*-value of 0.214. It
reports Egger's test as 0.8922, standard error 0.3767, $t = 2.37$, $p = 0.024$, CI (0.127,
1.657).

Funnel plot of passive smoking data
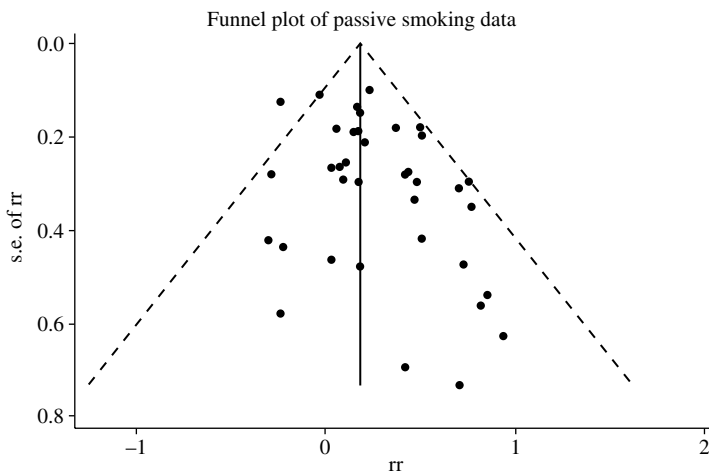
Figure 11.7   Funnel plot (Stata).

Duval and Tweedie's trim and fill

Stata runs the trim and fill algorithm and produces the results shown in Figure 11.8. In this example the trimming was based on the fixed-effect model. The original point estimate in log units had been 0.185, corresponding to a risk ratio of 1.20. The adjusted point estimate in log units is 0.156, corresponding to a risk ratio of 1.17. Perhaps the key point, though, is that the adjusted estimate is fairly close to the original – in this context, a risk ratio of 1.17 has the same substantive implications as a risk ratio of 1.20. The macro will allow the user to select any of the three trimming algorithms discussed in Duval and Tweedie's original paper (Duval and Tweedie, 2000a, 2000b) – linear, quadratic, and run. The results shown here are for the linear model, which is the default model in Stata and which Duval (personal communication) has reported as the most robust. The program also creates a funnel plot, with original points shown as circles and imputed points shown as circles encased in squares (see Figure 11.8).

Cumulative forest plot

Stata will produce a cumulative forest plot as shown in Figure 11.9. Note the difference between this cumulative forest plot and the standard version shown earlier. Here, the first row is a 'meta'-analysis based only on the Fontham *et al.* study. The second row is a meta-analysis based on two studies (Fontham *et al.* and Brownson *et al.*), and so on. The last study to be added is Butler, and so the point estimate and confidence interval shown on that line are identical to that shown for the summary effect.

The studies have been sorted from the most precise to the least precise (roughly corresponding to largest to smallest). With the 18 largest studies in the analysis (inclusive of Chan) the cumulative relative risk is 1.15. With the addition of another
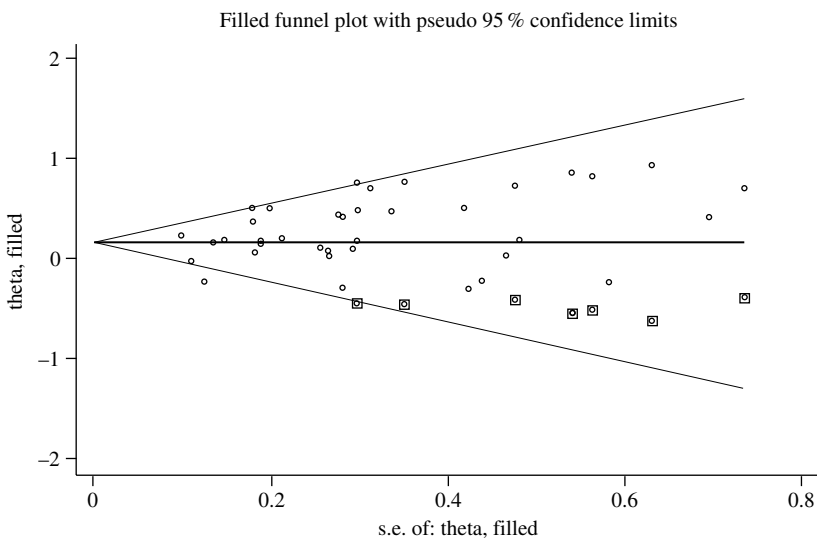
Figure 11.8   Trim and fill (Stata).
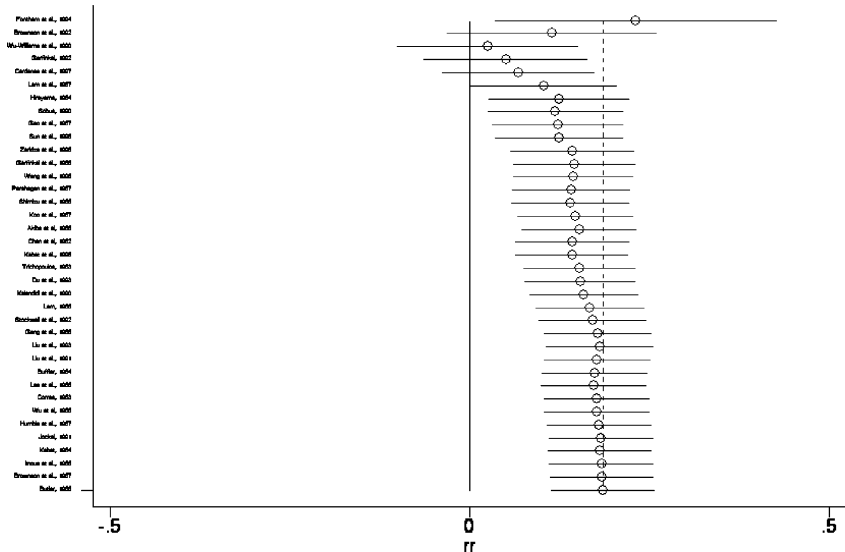
Figure 11.9 Cumulative forest plot (Stata).

19 (smaller) studies, the point estimate shifts to the right, and the relative risk is 1.20. As such, our estimate of the relative risk has increased. This could be due to publication bias or it could be due to small-study effects. Again, the key point is that even if we had limited the analysis to the larger studies, the relative risk would have been 1.15 (1.07, 1.25) and the clinical implications probably would have been the same.

This report used the procedures meta (Sharp, Sterne), metan (Bradburn, Deeks, Altman), metabias (Steichen, Egger, Sterne), metatrim (Steichen) metafunnel (Sterne and Harbord), metacum (Sterne), and meta-dialog (Steichen). Technical details for the Stata meta-analysis macros are given in Bradburn *et al.* (1999a, 1999b), Sharp and Sterne (1998a, 1998b, 1999), StataCorp. (2003), Steichen (1998, 2001a, 2001b, 2001c), Steichen *et al.* (1998, 1999), Sterne (1998), and Sterne and Harbord (2004). Perhaps the most comprehensive resource for using Stata in meta-analysis is Sterne *et al.* (2001a). The book that includes this chapter (Egger *et al.*, 2001) can be found at www.systematicreviews.com, and the chapter itself is available online as a PDF file at http://www.blackwellpublishing.com/medicine/bmj/systreviews/pdfs/chapter18.pdf.

Review Manager (version 4.2)

Review Manager (RevMan) is software provided by the Cochrane Collaboration for researchers whose reviews are intended for inclusion in the Cochrane Library. RevMan's statistical procedures are encapsulated in a module called RevMan Analyses (formerly MetaView) and invoked from within the larger program. RevMan will create a forest plot and a funnel plot but will not run any of the other procedures

discussed in this chapter. Details for this program are provided in Clarke and Oxman, 2003.

Data entry and conversion

RevMan will accept data in two formats (the number of events and sample size in each group, means and standard deviations), as well as a generic format (point estimate and standard error). RevMan will not accept the passive smoking data in the format provided (risk ratios and confidence intervals), so the researcher would need to compute the log risk ratio and its standard error using another program such as Excel and then enter these data into RevMan.

Forest plot

Figure 11.10 is a forest plot produced by RevMan. The graphical part of the plot (the point estimate and confidence interval) is displayed toward the center of the screen, with points to the right of 1.0 indicating an increased risk for persons exposed to passive smoking. The overwhelming majority of studies show an increased risk, but the 95 % confidence intervals for most studies include the null value of 1.0,

Study or sub-category	log[Risk ratio] (SE)	Risk ratio (fixed) 95% Cl	Weight %	Risk ratio (fixed) 95% Cl
Fontham et al.	0.2310 (0.1000)		13.67	1.26 [1.04, 1.53]
Brownson et al.(b)	−0.0300 (0.1120)		10.90	0.97 [0.78, 1.21]
Wu-Williams et al.	−0.2360 (0.1270)		8.48	0.79 [0.62, 1.01]
Garfinkel	0.1660 (0.1370)		7.28	1.18 [0.90, 1.54]
Cardenas et al.	0.1820 (0.1500)		6.08	1.20 [0.89, 1.61]
Lam et al.	0.5010 (0.1800)		4.22	1.65 [1.16, 2.35]
Hirayama	0.3720 (0.1820)		4.13	1.45 [1.02, 2.07]
Sobue	0.0580 (0.1840)		4.04	1.06 [0.74, 1.52]
Gao et al.	0.1740 (0.1900)		3.79	1.19 [0.82, 1.73]
Sun et al.	0.1480 (0.1910)		3.75	1.16 [0.80, 1.69]
Zaridze et al.	0.5070 (0.2000)		3.42	1.66 [1.12, 2.46]
Garfinkel et al.	0.2070 (0.2130)		3.01	1.23 [0.81, 1.87]
Wang et al.	0.1040 (0.2580)		2.05	1.11 [0.67, 1.84]
Shimizu et al.	0.0770 (0.2670)		1.92	1.08 [0.64, 1.82]
Pershagen et al.	0.0300 (0.2670)		1.92	1.03 [0.61, 1.74]
Koo et al.	0.4380 (0.2770)		1.78	1.55 [0.90, 2.67]
Akiba et al.	0.4190 (0.2820)		1.72	1.52 [0.87, 2.64]
Chan et al.	−0.2880 (0.2820)		1.72	0.75 [0.43, 1.30]
Kabat et al.	0.0950 (0.2940)		1.58	1.10 [0.62, 1.96]
Trichopoulos	0.7560 (0.2980)		1.54	2.13 [1.19, 3.82]
Du et al.	0.1740 (0.2990)		1.53	1.19 [0.66, 2.14]
Kalandidi et al.	0.4820 (0.2990)		1.53	1.62 [0.90, 2.91]
Lam	0.6980 (0.3130)		1.40	2.01 [1.09, 3.71]
Stockwell et al.	0.4700 (0.3370)		1.20	1.60 [0.83, 3.10]
Geng et al.	0.7700 (0.3520)		1.10	2.16 [1.08, 4.31]
Liu et al.(b)	0.5070 (0.4190)		0.78	1.66 [0.73, 3.77]
Liu et al.(a)	−0.3010 (0.4250)		0.76	0.74 [0.32, 1.70]
Buffler	−0.2230 (0.4390)		0.71	0.80 [0.34, 1.89]
Lee et al.	0.0300 (0.4660)		0.63	1.03 [0.41, 2.57]
Correa	0.7280 (0.4770)		0.60	2.07 [0.81, 5.27]
Wu et al.	0.1820 (0.4810)		0.59	1.20 [0.47, 3.08]
Humble et al.	0.8500 (0.5410)		0.47	2.34 [0.81, 6.76]
Jockel	0.8200 (0.5630)		0.43	2.27 [0.75, 6.84]
Kabat	−0.2360 (0.5820)		0.40	0.79 [0.25, 2.47]
Inoue et al.	0.9360 (0.6310)		0.34	2.55 [0.74, 8.78]
Brownson et al. (a)	0.4190 (0.6960)		0.28	1.52 [0.39, 5.95]
Butler	0.7030 (0.7350)		0.25	2.02 [0.48, 8.53]
Total (95 % Cl)			100.00	1.20 [1.12, 1.29]

Test for heterogeneity: Chi^2 = 47.52, df = 36 (P = 0.09), I^2 = 24.2%
Test for overall effect: Z = 5.02 (P < 0.00001)

0.1 0.2   0.5   1   2   5   10

Figure 11.10   Forest plot (RevMan).

and therefore fail to meet the criterion for statistical significance. The last row in the spreadsheet shows the summary data for the fixed-effect model. The risk ratio is 1.20 and the 95 % confidence interval is from 1.12 to 1.29. Toward the left, the program shows the name and log risk ratio (SE) for each study. Toward the right, the program shows the relative weight assigned to each study, the risk ratio and confidence interval.

In this plot, the studies have been listed from most precise to least precise, so that larger studies appear toward the top and smaller studies appear toward the bottom. This has no impact on the analysis, but allows us to get an initial sense of the relationship between sample size and effect size. Note that the point estimates become more widely dispersed but also shift toward the right as we move toward the bottom of the plot, where the smaller studies (note the wider confidence intervals) are located. This will be more evident in the funnel plot, which is designed to highlight this relationship.

Funnel plot

RevMan Analyses will display a funnel plot of effect size on the *x*-axis by standard error on the *y*-axis (Figure 11.11). It allows the user to include a vertical line at the summary effect, and guidelines for the 95 % confidence interval as shown. In RevMan these are based on the fixed-effect model. In this example studies toward the bottom of the plot tend to cluster toward the right, suggesting the possibility of publication bias.

Aside from the funnel plot, RevMan Analyses is not able to run any of the publication bias procedures nor to create a cumulative forest plot. The best option
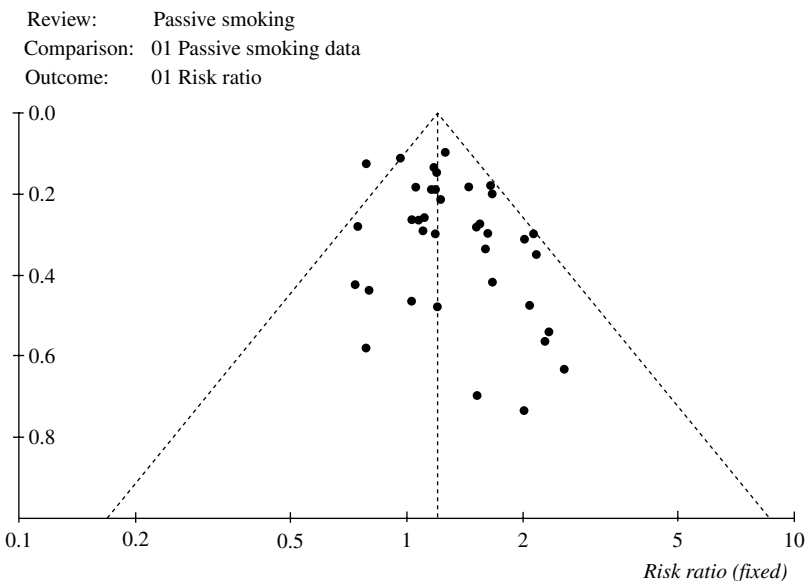
Figure 11.11   Funnel plot (RevMan).

for running additional analyses would therefore be to export the data to one of the other programs. RevMan Analyses does not support cut-and-paste of data using the Windows© clipboard, but does allow the user to export data to an ASCII (text) file. Step-by-step instructions are provided on the website.

MetaWin (version 2.0)

MetaWin is a stand alone program for meta-analysis. It was developed by people working in the field of ecology, but its features are general enough to have application in most other fields as well. MetaWin is able to create a forest plot and a funnel plot. It will compute the rank correlation test, the failsafe N and Orwin's variant, and display a cumulative forest plot sorted by precision (see Rosenberg *et al.*, 2000).

Data entry and conversion
MetaWin will accept data in three formats (the number of events and sample size in each group, means and standard deviations, correlations), as well as a generic format (point estimate and variance). MetaWin will not accept the passive smoking data in the format provided (risk ratios and confidence intervals), so the researcher would need to compute the log risk ratio and its variance using another program such as Excel and then cut and paste these data into MetaWin.

Forest plot
The forest plot (Figure 11.12) has been sorted according to each study's precision, with larger studies appearing toward the bottom and smaller studies appearing
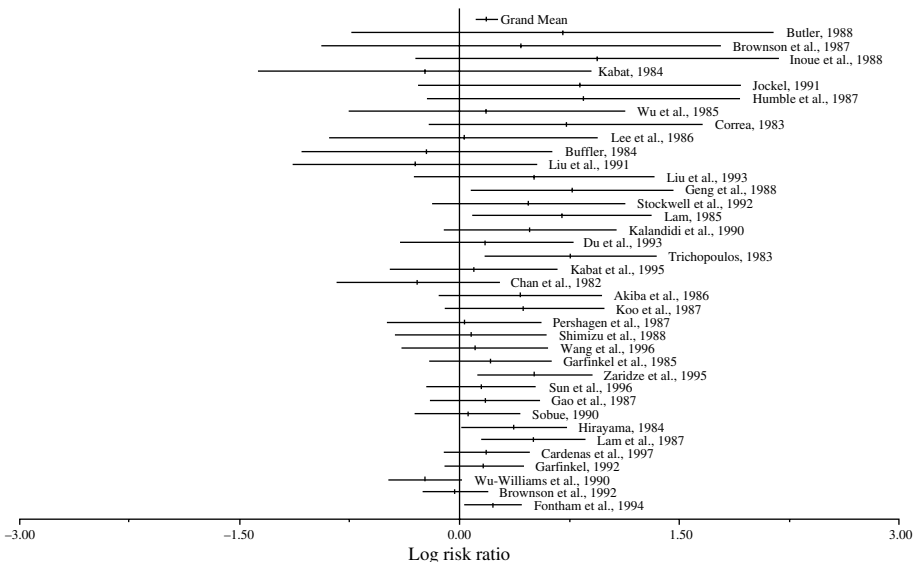
Figure 11.12 Forest plot (MetaWin), read from bottom to top.

toward the top (note that this is the reverse of the other programs). This has no impact on the analysis, but allows us to get an initial sense of the relationship between sample size and effect size. The summary effect of 0.186 (0.113, 0.259), corresponding to a risk ratio of 1.204 (1.120, 1.295) is shown at the top, and labeled 'Grand mean'. Note that the point estimates become more widely dispersed but also shift toward the right as we move toward the top of the plot, where the smaller studies (note the wider confidence intervals) are located. This will be more evident in the funnel plot, which is designed to highlight this relationship.

Funnel plot
In MetaWin the funnel plot (Figure 11.13) is shown with the effect size on the Y-axis and the variance (as in this example) or the sample size on the X-axis. In this example smaller studies (which appear toward the right) are clustered toward the top (representing larger effects). MetaWin does not offer the option of including a line to denote the summary effect, nor guidelines for the 95 % confidence interval. MetaWin will also let the user create a scatter plot of any two continuous variables, so it would be possible to plot the effect size against any other index of precision that the researcher had included in the data sheet.

Statistical tests
The failsafe N is given as 397.3. For computing Orwin's failsafe N, MetaWin assumes the mean value in the missing studies is zero but the user is allowed to define the value of a negligible effect. If this is defined as a log odds ratio of 0.05, Orwin's failsafe N is computed as 103, assuming a risk ratio of 1.0 in missing studies and a defining a trivial effect as a risk ratio of 1.05 (log risk ratio of 0.049).
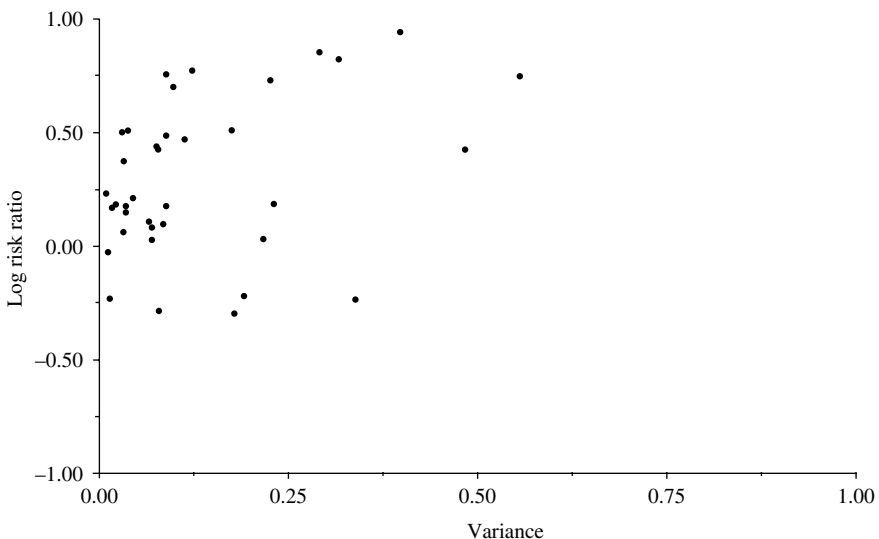
Figure 11.13   Funnel plot (MetaWin).

MetaWin reports Kendall's tau as 0.144, $z = 1.256$, p (one-tailed) $= 0.21$. MetaWin will not computer Egger's test of the intercept.

Cumulative forest plot

MetaWin will produce a cumulative forest plot as shown in Figure 11.14. While the program does include various options for sorting the data, these will not work with the cumulative forest plot. Therefore, the user must ensure that the studies have been entered into the data sheet in the sequence of most precise to least precise studies in order for this option to work.

As was true for the basic forest plot, MetaWin plots the studies from the bottom up. The meta-analysis for the first study (Fontham *et al.*) is omitted (since it would include only one study) and so the bottom row shows the analysis for the first two studies (Fontham *et al.* and Brownson *et al.*). The second row from the bottom is a meta-analysis based on three studies, and so on. The last study to be added is Butler, at the top.

The studies have been sorted from the most precise to the least precise (roughly corresponding to largest to smallest). With the 18 largest studies in the analysis (inclusive of Chan) the cumulative relative risk is 1.15 (the graph shows this in log units). With the addition of another 19 (smaller) studies, the point estimate shifts to the right, and the relative risk is 1.20. As such, our estimate of the relative risk has increased. This could be due to publication bias or it could be due to small-study effects. Perhaps the key point, though, is that even if we had limited the analysis to the larger studies, the relative risk would have been 1.15 (1.07, 1.25) and the clinical implications probably would have been the same.
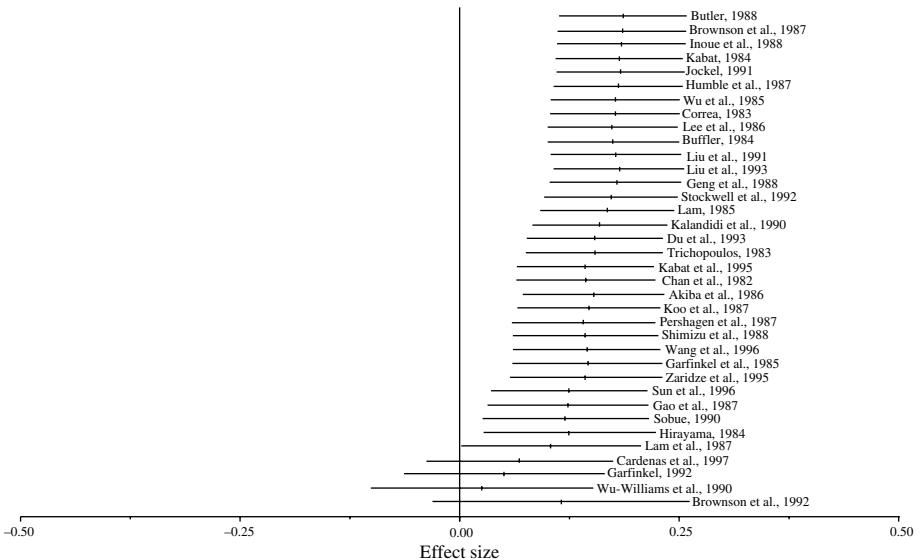
Figure 11.14   Cumulative forest plot (MetaWin), read from bottom to top.

Summary of the findings for the illustrative example

The various statistical procedures approach the problem of bias from a number of directions. One would not expect the results of the different procedures to 'match' each other since the procedures ask different questions. Rather, the goal should be to synthesize the different pieces of information provided by the various procedures.

Is there evidence of bias?

The funnel plot is noticeably asymmetric, with a majority of the smaller studies clustering to the right of the mean. This impression is confirmed by Egger's test which yields a statistically significant p-value. The rank correlation test did not yield a significant p-value, but this could be due to the low power of the test, and therefore does not contradict the Egger test. The body of evidence suggests that smaller studies reported a higher association between passive smoking and lung cancer than did the larger studies.

Is it possible that the observed relationship is entirely an artifact of bias?

The failsafe N is 398, suggesting that there would need to be nearly 400 studies with a mean risk ratio of 1.0 added to the analysis, before the cumulative effect would become statistically non-significant. Similarly, Orwin's failsafe N is 103, suggesting that there would need to be over 100 studies with a mean risk ratio of 1.0 added to the analysis before the cumulative effect would become trivial (defined as a risk ratio of 1.05). Given that the authors of the meta-analysis were able to identify only 37 studies that looked at the relationship of passive smoking and lung cancer, it is unlikely that nearly 400 studies (Rosenthal) or even 103 studies (Orwin) were missed. Therefore, while the actual risk of passive smoking may be smaller than reported, it is unlikely to be zero.

What is the risk ratio adjusted for bias?

We can attempt to classify the bias as falling into one of three categories: (a) the bias, if any, is probably trivial; (b) the bias is probably non-trivial, but we can have confidence in the substantive conclusions of the meta-analysis; or (c) the bias is potentially severe enough to call the substantive findings into question.

   In this example the meta-analysis based on all studies showed that passive smoking was associated with a 20% increase in risk of lung cancer. The cumulative meta-analysis showed that the increased risk in the larger studies was 16%. Similarly, the trim and fill method suggested that if we trimmed the 'biased' studies, the increased risk would be imputed as 15%. As such, this meta-analysis seems to fall squarely within category (b). There is evidence of bias, and the risk is probably somewhat lower than reported, but there is no reason to doubt the validity of the core finding, that passive smoking is associated with a clinically important increase in the risk of lung cancer.

WHERE DO WE GO FROM HERE?

As noted, the exaggerated effect in the small studies could be due to publication bias but could also be due to a small-study 'bias'. If the latter, this 'bias' reflects a source of heterogeneity among studies which, like any other heterogeneity, should be investigated.

COMPARING THE PROGRAMS

Table 11.1 shows the publication bias procedures incorporated in the four programs. As indicated in this table, three of the four programs will carry out most of the publication bias procedures. Aside from this, however, the programs have surprisingly little in common, regarding their approaches to data entry and to the core meta-analysis, and even the publication bias procedures are quite different. This section is meant to give the reader a better sense of the 'look and feel' of the four programs, and as such is bound to be subjective. Therefore, I want to remind the reader of the conflict of interest mentioned at the outset.

Comprehensive Meta Analysis (2.0) offers the advantage of being able to perform a wide range of tasks required for the analysis. It will accept data in virtually any format, automatically compute the effect size, and (where needed) convert data from different formats to a common effect size index. CMA allows the researcher to run the meta-analysis with one click, but allows the advanced user to select

Table 11.1   Publication bias procedures in meta-analysis programs.

	CMA (2.0)	Stata (8.2)	RevMan (4.2)	MetaWin (2.0)
Data entry formats				
Effect size and standard error	●	●	●	●
Events and sample size/non-events	●	●	●	●
Means and standard deviation	●	●	●	●
Point estimate and confidence interval	●	●		
Get a sense of the data				
Forest plot	●	●	●	●
Funnel plot	●	●	●	●
Is there evidence of bias?				
Begg and Mazumdar's rank correlation	●	●		●
Egger's regression	●	●		●
Can effect be entirely an artifact of bias?				
Failsafe *N*	●			●
Orwin's failsafe *N*	●			●
What is the adjusted effect size?				
Duval and Tweedie's trim and fill	●	●		
Cumulative forest plot	●	●		●
Miscellaneous				
Text report with detailed narrative	●			

from an array of computational options. The program features advanced options for graphics, including the ability to extensively customize the forest plot (including the use of proportionally sized symbols) and to export high-resolution plots to other programs such as Word and PowerPoint. The interface is similar to that of programs such as Excel™, so that the learning curve is minimal, and the functions are intuitive.

Stata (8.2) is a general purpose statistical program. To use Stata for meta-analysis the researcher would incorporate a series of macros (code segments) that have been written by experts in meta-analysis. These macros can be located and downloaded by accessing the web from within Stata. Stata differs from the other three programs in that it is command-driven (i.e., the user writes code, or uses a dialog box to create code), rather than being directly menu-driven, and therein lie both its advantages and disadvantages. For someone who is familiar with Stata, and who is relatively expert in meta-analysis, Stata offers the advantage that the user has complete control over every element of the analysis, from the computation of effect size to the precise weighting of each study. Code can be written from scratch, the provided macros can be viewed and modified, and all code can be saved and reused with other data sets. Someone who is not familiar with Stata, on the other hand, would need to learn the basic Stata language before proceeding to meta-analysis, and would encounter problems working with data formats that were not automatically handled by the Stata macros. One disadvantage of Stata, even for the advanced user, is that the forest plot may be difficult to read and relatively little customization is currently possible (see Figure 11.6, for example).

RevMan Analyses (4.2) is meant to function as part of a larger package whose goal is not only to run the meta-analysis but also to track other aspects of the systematic review and to ensure that the data can be uploaded to the Cochrane Library. The advantage of this program is that it is compatible with the Cochrane Library. The disadvantage is that data entry is not intuitive, and is limited to only a few formats (2×2 tables, means and standard deviations, and a generic point estimate and standard error). While the forest plot can be customized only slightly, the default format is well designed and incorporates all key data. Anyone working with the Cochrane Collaboration will need to enter their data into RevMan, so the question for these people is whether or not to use another program as well. If they want to produce higher quality graphics, or to run additional publication bias analyses, they should consider exporting the data to CMA or to Stata. RevMan does not allow copy and paste of data, but it does allow the user to export the entire data file in ASCII format, and this file can then be read by either CMA or Stata (the appendix to this chapter lists a website with instructions).

MetaWin 2.0 has taken the same general approach as CMA in that it offers a spreadsheet interface with menus and toolbars. However, MetaWin offers a much more limited version of this approach. The data input will work with only a few formats. The forest plot is relatively difficult to read and offers few options for customizing. A strength of this program is that it allows the user to work with the response ratio, an effect size index that is used in ecology and is not found in any of the other programs. The program also provides resampling methods (bootstrap and randomization) for Q-statistics and confidence intervals of cumulative effect sizes.

Computational accuracy

In choosing among programs, accuracy of results is not an issue. The programs all give the same results, both for the core meta-analyses and also for the publication bias analyses. The literature includes two versions of Egger's test, the original using a weighting scheme and the later version (at Egger's suggestion) omitting the weight. All programs that include this test use the newer version of the formula. Two of the programs reviewed here compute the failsafe N, and they both use the Stouffer method (Stouffer *et al.*, 1949), which is the one originally used by Rosenthal. Note, however, that the programs will yield the same results only if they are using the same options, and that different programs may use different computational options as the default. For example, the default in CMA is to use a two-tailed test for failsafe N, while the default in MetaWin is to use a one-tailed test. CMA offers a button to set all the defaults to match those used in RevMan. None of the programs implements the selection models discussed by Hedges and Vevea (Chapter 9).

Other programs

The discussion in this chapter does not include all programs used for meta-analysis. SPSS and SAS are general purpose statistical packages which have often been pressed into service for meta-analysis by researchers who write their own code or who incorporate the macros written by David Wilson (Lipsey and Wilson, 2001). Additional detail for these programs, including instructions for downloading the macros, are included on the book's website. These programs were not included in the discussion because there are no widely available macros to address publication bias (see also Wang and Bushman, 1999; Arthur *et al.*, 2001).

REFERENCES

Arthur, W., Bennett, W. and Huffcutt, A.J. (2001). *Conducting Meta-analysis Using SAS*. Mahwah, NJ: Lawrence Erlbaum Associates.

Begg, C.B. (1994). Publication bias. In H.M. Cooper and L.V. Hedges (eds), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Begg, C.B. and Berlin, J.A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A*, **151**, 419–463.

Begg, C.B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, **50**, 1088–1101.

Borenstein, M., Hedges, L., Higgins, J. and Rothstein, H. (2005) *Comprehensive Meta Analysis, Version 2*. Englewood, NJ: Biostat.

Bradburn, M.J., Deeks, J.J. and Altman, D.G. (1999a). Metan – an alternative meta-analysis command. *Stata Technical Bulletin Reprints*, **8**, 86–100.

Bradbury, M.J., Deeks, J.J. and Altman, D.G. (1999b). Correction to funnel plot. *Stata Technical Bulletin Reprints*, **8**, 100.

Cooper, H.M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, **37**, 131–146.

Clarke, M. and Oxman, A.D. (eds) (2003). Cochrane Reviewers' Handbook 4.2.0 [updated March 2003]. In *The Cochrane Library, Issue 2, 2003*. Oxford: Update Software. Updated quarterly.

Carson, K.P., Schriesheim, C.A. and Kinicki, A.J. (1990). The usefulness of the 'fail-safe' statistic in meta-analysis. *Educational and Psychological Measurement*, **50**, 233–243.

Duval, S.J. and Tweedie, R.L. (1998). Practical estimates of the effect of publication bias in meta-analysis. *Australasian Epidemiologist*, **5**, 14–17.

Duval, S. and Tweedie, R. (2000a) A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, **95**, 89–99.

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

Egger, M., Davey Smith, G., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, **315**, 629–634.

Egger, M., Davey Smith, G., and Altman, D. (eds) (2001). *Systematic Reviews in Health Care: Meta-analysis in Context*. London: BMJ.

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.

Lau, J., Schmid, C.H. and Chalmers, T.C. (1995). Cumulative meta-analysis of clinical trials: builds evidence for exemplary medical care. *Journal of Clinical Epidemiology*, **48**, 45–57.

Light, R. & Pillemer, D. (1984). *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press, Cambridge.

Light, R.J., Singer, J.D. and Willett, J.B. (1994). The visual presentation and interpretation of meta-analyses. In H.M. Cooper and L.V. Hedges (eds), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation

Lipsey, M.W. and Wilson, D.B. (2001) *Practical Meta-analysis*. Thousand Oaks, CA: Sage.

Orwin, R.G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, **8**, 157–159.

RevMan Analyses (2002) [Computer program]. Version 1.0 for Windows. In Review Manager (RevMan) 4.2. Oxford, England. The Cochrane Collaboration.

Rosenberg, M.S., Adams, D.C. and Gurevitch, J. (2000) *MetaWin: Statistical Software for Meta-Analysis Version 2.0*. Sunderland, MA: Sinauer Associates, Inc.

Rosenthal, R. (1979). The 'file-drawer problem' and tolerance for null results. *Psychological Bulletin*, **86**, 638–641.

Sharp, S.J. and Sterne, J. (1998a). Meta-analysis. *Stata Technical Bulletin Reprints*, **7**, 100–106.

Sharp, S.J. and Sterne, J. (1998b). New syntax and output for the meta-analysis command. *Stata Technical Bulletin Reprints*, **7**, 106–108.

Sharp, S.J. and Sterne, J. (1999). Corrections to the meta-analysis command. *Stata Technical Bulletin Reprints*, **8**, 84.

StataCorp. (2003). Stata Statistical Software: Release 8.0. College Station, TX: StataCorp LP.

Steichen, T.J. (1998). Tests for publication bias in meta-analysis. *Stata Technical Bulletin Reprints*, **7**, 125–133.

Steichen, T.J. (2001a). Update of tests for publication bias in meta-analysis. *Stata Technical Bulletin Reprints*, **10**, 70.

Steichen, T.J. (2001b). Tests for publication bias in meta-analysis: erratum. *Stata Technical Bulletin Reprints*, **10**, 71.

Steichen, T.J. (2001c). Update to metabias to work under version 7. *Stata Technical Bulletin Reprints*, **10**, 71–72.

Steichen, T.J., Egger, M. and Sterne, J. (1998) Modification of the metabias program. *Stata Technical Bulletin*, **STB-44**(sbe19.1), 3–4.

Steichen, T.J., Egger, M. and Sterne, J. (1999). Tests for publication bias in meta-analysis. *Stata Technical Bulletin Reprints*, **8**, 84–85.

Sterne, J. (1998). Cumulative meta-analysis. *Stata Technical Bulletin Reprints*, **7**, 143–147.

Sterne, J. A. & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis, *Journal of Clinical Epidemiology*, **54**, 1046–1055.

Sterne, J.A. and Harbord, R.M. (2004) Funnel plots in meta analysis. *Stata Journal*, **4**, 127–141.

Sterne, J.A.C., Gavaghan, D. and Egger, M. (2000) Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, **53**, 1119–1129.

Sterne, J. A., Bradburn, M. J., & Egger, M. (2001a). Meta-Analysis in Stata. In M. Egger, G. Davey Smith and D. G. Altman (eds.), *Systematic Reviews in Health Care: Meta-analysis in Context* (pp. 347–369). London: BMJ Books.

Sterne, J. A. C., Egger, M. and Davey Smith, G. (2001b). Investigating and dealing with publication and other biases. In M. Egger, G. Davey Smith and D. G. Altman (eds), *Systematic Reviews in Health Care: Meta-analysis in Context.*, 2nd edn. London: BMJ Books

Sterne, J. A., Egger, M., and Davey Smith, G. (2001c). Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis, *British Medical Journal*, **323**, 101–105.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. and Williams, R.M., Jr. (1949). *The American Soldier: Adjustment during Army Life* (Vol. 1). Princeton, NJ: Princeton University Press.

Wang, M.C. and Bushman, B.J. (1999). *Integrating Results through Meta-analytic Review Using SAS Software*. Cary, NC: SAS Institute Inc.

APPENDIX

Updates and data sets

Updates to this chapter and also the data sets used as examples in this volume, will be posted on the book's website, http://www.meta-analysis.com/publication-bias.

Conflicts of interest

The development of Comprehensive Meta Analysis was funded by the National Institutes of Health in the United States. The editors of this book, as well as many of the chapter authors, have been involved in the development of that program, and some have a financial interest in its success. The author of this chapter is the lead developer on the program and has a financial interest in its success.

To help ensure the accuracy of the information in this chapter, developers and users of the various programs were asked to review a draft of this chapter. Jonathan Sterne, who collaborated on the development of the Stata macros and has written a chapter on using Stata for meta-analysis, provided feedback on Stata. Julian Higgins,

a member of the RevMan Advisory Group, provided comments on RevMan and on Stata. Dean Adams, one of the developers of MetaWin, provided comments on that program. Alex Sutton, Hannah Rothstein, and Steven Tarlow provided extensive and very helpful comments on earlier drafts of this chapter.

Trademarks

Comprehensive Meta Analysis is a trademark of Biostat, Inc. MetaWin is a trademark of that program's developers. RevMan is a trademark of the Cochrane Collaboration. Stata is a trademark of StataCORP.

Contacts

Comprehensive Meta Analysis
Biostat, Inc., 14 North Dean Street, Englewood, NJ 07631, USA
Phone + 1 201 541 5688
Fax + 1 201 541 5526
MBorenstein@Meta-Analysis.com
www.Meta-Analysis.com

MetaWin

Sinauer Associates, Inc., 23 Plumtree Road, P.O. Box 407, Sunderland, MA 01375-0407, USA
Phone + 1 413 549 4300
Fax + 1 413 549 1118
www.metawinsoft.com

Stata

4905 Lakeway Drive, College Station, TX 77845, USA
Phone + 1 979 696 4600 (USA 800 782 8272)
Fax + 1 979 696 4601
www.Stata.com

RevMan

http://www.cochrane.org/software/download.htm