APPENDIX B

Annotated Bibliography

Hannah R. Rothstein
Department of Management, Zicklin School of Business, Baruch College, New York, USA

Ashley Busing
Department of Psychology, Weissman School of Arts and Sciences, Baruch College, New York, USA

In this bibliography we list and describe a selection of materials that provide a more detailed look at various aspects of publication bias. Although some were selected independently by the authors of the appendix, many were chosen based on conversations with the chapter authors, who had been asked to nominate articles that were either 'classics' or 'cutting-edge'. By 'classics' we meant articles that described the research that established the legitimacy of publication bias as a topic of scientific inquiry, or that provided the methods by which this inquiry could be advanced; by 'cutting-edge' we meant articles that present the newest techniques or findings in the area. The references are presented in chronological order so that the reader will be able to get a sense of the way in which the scientific study of publication bias has developed over time.

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association*, **54**, 30–34.

This is possibly the earliest paper to provide (indirect) evidence of publication bias.

Rosenthal, R. (1979). The 'file drawer problem' and tolerance for null results. *Psychological Bulletin*, **86**, 638–461.

In addition to thanking the chapter authors, we would like to thank Doug Altman and Will Shadish for their suggestions about what to include in this bibliography.

Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments Edited by H.R. Rothstein, A.J. Sutton and M. Borenstein ISBN: 0-470-87014-1

This is the first study to suggest a method for assessing the potential impact of publication bias on meta-analytic results. It uses a 'failsafe' approach to test the robustness of a statistically significant effect by providing a formula to calculate the number of 'missing' studies averaging zero effect that would be needed to reduce a significant observed overall effect to non-significance.

Hemminki, E. (1980). Study of information submitted by drug companies to licensing authorities. *British Medical Journal*, **280**, 833–836.

This paper provides early evidence that industry-based clinical trials of new drugs may remain unpublished. It reports on clinical trials that were submitted by drug companies to regulatory authorities in Finland and Sweden during 1965–1975, and were never submitted for publication. Although trial design and quality varied among the studies, the author suggests that valuable information, including information about adverse effects, was lost because these studies were not published.

Smith, M.L. (1980). Publication bias and meta-analysis. *Evaluation in Education*, **4**, 22–24.

This was the first paper to use the term 'publication bias'.

Smith, M.L. (1980). Sex bias in counseling and psychotherapy. *Psychological Bulletin*, **87**, 392–407.

Smith separately meta-analysed published and unpublished studies in order to see whether they provided different answers to the question of whether counsellors and therapists were biased against women. (This was a hot political issue at the time of Smith's research.) Her results provide one of the earliest examples of publication bias. She found that published studies showed a small effect of bias against women, while unpublished studies showed the same magnitude of bias towards women. Her analysis also indicated that the degree of rigour in research design was the same in both published and unpublished studies.

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

In this paper, Orwin proposes an early method for evaluating the effects of putatively missing studies on meta-analytic results. His method is similar to Rosenthal's file-drawer method, except that it is focused on assessing the influence of missing studies on the magnitude of the combined mean effect rather than on its statistical significance.

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

This paper investigates the effects of extreme publication bias (where only statistically significant results are observed) on estimates of the standardized mean

difference. It derives the sampling distribution of the effect size estimates under extreme publication bias, shows that the bias depends on sample size and the true effect size, and shows that estimates can be biased by over 200 %. The maximum likelihood estimates of effect size assuming the extreme publication bias are also derived and shown to correspond to shrinking the observed (and biased) effect sizes towards zero. A table is provided to compute effect size estimates 'corrected' for extreme publication bias. Finally, an estimate of effect size derived from counts of positive and negative significant results is obtained which is valid under extreme publication bias. Sampling distributions and standard errors for all of the estimates are presented. Highly statistical.

Simes, R.J. (1986). Publication bias: The case for an international registry of clinical trials. *Journal of Clinical Oncology*, **4**, 529–541.

This paper provides one of the first illustrations of how a trials registry could reduce the problem of publication bias. The paper compares the results of published trials with the result of all trials appearing in a registry, whether published or not. For treatment of ovarian cancer with combination chemotherapy, pooled analysis of published clinical trials demonstrated a significant survival advantage for combination chemotherapy but no significant difference in survival when in the pooled analysis of all registered trials. For multiple myeloma, a pooled analysis of published trials demonstrated a significant survival advantage for combination chemotherapy, as did a pooled analysis of all registered trials; however, for all registered trials the estimated magnitude of the benefit was reduced.

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

This is the first comprehensive review of the publication bias literature presented to a medical statistics audience. The authors of this paper may have been the first to propose the idea of looking at sample size versus effect size in a formal, quantitative manner. They suggested that publication bias might be a bigger problem for new than for established areas of inquiry.

Iyengar, S. and Greenhouse, J.B. (1988). Selection models and the file drawer problem. *Statistical Science*, **3**, 109–135.

This is one of the first papers to advance the use of selection modelling to deal with publication bias. The authors use an example to demonstrate the differences between the failsafe *N* approach to dealing with publication bias and a maximum likelihood selection-modelling approach, and conclude that the maximum likelihood method offers several advantages over the failsafe *N* approach.

Meinert, C.L. (1988). Toward prospective registration of clinical trials. *Controlled Clinical Trials*, **9**, 1–5.

Along with the papers by Dickersin and Min (1993) and Simes (1986), this paper is among the earliest calls for a registry of clinical trials.

Berlin, J.A., Begg, C.B. and Louis, T.A. (1989) An assessment of publication bias using a sample of published clinical trials. *Journal of the American Statistical Association*, **84**, 381–392.

This is one of the earliest papers to demonstrate a relationship between sample size and effect size in the published clinical trials literature. The authors found a strong relationship between sample size and treatment effect in published cancer trials for three outcomes: overall patient survival, disease-free survival and tumour response rate. An examination of several study features showed that some were associated with bias, but that none could account for the impact of sample size on treatment effect. The authors present a carefully detailed description of the data set they used, the methodological problems they dealt with, the decisions they made, and the statistical analyses they conducted. Some of this information is statistically demanding, but other parts of the article should be easily understandable by non-statisticians.

Hetherington, J., Dickersin, K., Chalmers, I. and Meinert, C.L. (1989). Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. *Pediatrics*, **84**, 374–380.

The authors attempted to retrieve information about unpublished randomized trials from obstetricians and paediatricians, and were largely unsuccessful except for trials that had been completed within the most recent two years. They conclude that the problem of publication bias will not be resolved through retrospective identification of unpublished trials, and encourage the use of prospective registries.

Chalmers, I. (1990). Underreporting research is scientific misconduct. *Journal of the American Medical Association*, **263**, 1405–1408.

In this paper, the author contends that the failure to publish results of rigorously designed clinical trials is a form of scientific misconduct, noting that it can lead to inappropriate treatment decisions. He suggests that researchers, research ethics committees, funders and editors all share the responsibility for this problem, and that all must participate in reducing it. He suggests that prospective registration of clinical trials would ameliorate the problem substantially.

Easterbrook, P.J., Berlin, J.A., Gopalan, R. and Matthews, D.R. (1991). Publication bias in clinical research. *Lancet*, **337**, 867–872.

This is the first paper to look at a cohort of studies approved by an institutional review board (IRB) and follow them forward to see whether the studies that were eventually published had a greater likelihood of being statistically significant than those that were not published. The results showed that statistically significant studies were about twice as likely to be published as those that did not reach statistical significance. Higher likelihood of publication was also associated with the importance of the study results (as rated by the investigator), and with increasing sample size. The tendency towards publication bias was greater

with observational and laboratory-based experimental studies than with randomized clinical trials. The authors caution against overinterpretation of conclusions based on a review of only published data, and suggest the need for improved strategies to identify the results of unpublished studies.

Chalmers, I., Dickersin, K. and Chalmers, T.C. (1992). Getting to grips with Archie Cochrane's agenda. *British Medical Journal*, **305**, 786–788.

This short paper describes the work of Archie Cochrane and his focus on using the results of randomized clinical trials to develop health care practices and policies in Great Britain.

Dickersin, K., Min, Y.I. and Meinert, C.L. (1992). Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *Journal of the American Medical Association*, **267**, 374–378.

This is one of the first studies to demonstrate that publication bias originates for the most part with investigators rather than with journal editors. In a study of 737 studies that had received IRB approval, there were significant differences in the probability of publication for studies with significant or non-significant findings, but less than 5 % of the studies that had not been published were reported to have been rejected for publication.

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

This paper investigates the problem of modelling, and correcting for, publication selection in meta-analysis. It distinguishes between the concepts of selection models and effect size models. Selection models describe the probability that an effect size estimate is observed as a function of its level of statistical significance. Effect size models describe what the sampling distribution of effect size estimates would be if there were no selection. If the selection model were known, it would be possible to estimate the effect size model – that is, to estimate what the value of the mean effect size would be if there were no selection. This paper introduces a flexible, non-parametric, selection model that can be estimated (by maximum likelihood) from effect size data simultaneously with the effect size model. This provides information about the selection process that is operating (via the estimated selection model) and what the mean and between-studies variance component of the effect sizes would have been in the absence of selection (via the parameters of the effect size model). Statistical tests for the presence of selection and for mean and between-studies variance of effect sizes are given. Highly statistical.

Dickersin, K. and Min, Y.I. (1993). NIH clinical trials and publication bias. *Online Journal of Current Clinical Trials*, Doc. No. 50.

This is one of the earliest papers to systematically demonstrate that papers with statistically significant findings were more likely to be published than those with

non-significant findings. A meta-analysis of data from this and three similar studies yielded results that showed that significant papers were nearly three times as likely to be published as papers that did not reach statistical significance. Also of note is that this is one of the earliest papers to call for a registry of initiated trials as a means of preventing publication bias, as well as for financial support for this effort.

Stewart, L.A. and Parmar, M.K.B. (1993). Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet*, **341**, 418–422.

This is the first paper describing a formal comparison of the results and implications of two different approaches to meta-analysis. Using an example in ovarian cancer, an IPD meta-analysis was compared with a meta-analysis using data extracted from trial publications. Substantially more data were available with the IPD approach. In addition, longer-term follow-up was available and more appropriate time-to-event analysis of survival data could be done with the IPD data. The results of the meta-analysis of data from published reports were more positive, with an absolute estimate of treatment effect that was three times as large as the estimate from IPD. The results based on data extracted from trial reports were also statistically significant, whereas those from the IPD were not. The authors concluded that the results of each approach were likely to have a different clinical interpretation.

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

This paper introduces the first statistical test for the assessment of publication bias, which is based on the rank correlation between the treatment effect and the standard error for each study. The test is shown to be fairly powerful for large meta-analyses (75 or more studies), but has only moderate power for medium-sized meta-analyses (25 studies).

Bushman, B.J. and Wang, M.C. (1995). A procedure for combining sample correlation coefficients and vote counts to obtain an estimate and a confidence interval for the population correlation coefficient. *Psychological Bulletin*, **117**, 530–546.

In this article, the authors propose supplementing sample size procedures with vote-counting procedures as a means of salvaging studies that are relevant to the meta-analysis, but which are missing effect size estimates. They suggest that this will lead to less biased estimates of the population effect, and a narrower confidence interval around it than ignoring the studies without effect sizes. The article describes three vote-counting procedures that can be used to estimate the population effect for studies with missing sample correlations.

Stewart, L.A. and Clarke, M.J. on behalf of the Cochrane Working Party Group on Meta-analysis using Individual Patient Data (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*, **14**, 2057–2079.

This paper, based on experiences shared at a Cochrane Collaboration workshop, describes the rationale and practical methodology of IPD meta-analyses.

Vevea, J.L. and Hedges, L.V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, **60**, 419–435.

This paper presents a maximum likelihood based model of estimation of effect sizes in the presence of selection based on one-tailed *p*-values (publication bias, based on significance of results). It presents a test for the presence of publication bias, and a means of estimated corrected effect values. The authors provide an example based on the psychotherapy effectiveness literature. Highly statistical.

Bushman, B.J. and Wang, M.C. (1996). A procedure for combining sample standardized mean differences and vote counts to estimate the population standardized mean difference in fixed effect models. *Psychological Methods*, **1**, 66–80.

In this paper, the authors outline a procedure for handling missing estimates which combines effect size estimates and vote counts. They recommend this as the method of choice for a meta-analysis when some studies do not provide sufficient information to compute effect size estimates but do present the direction or statistical significance of results.

Hedges, L.V. and Vevea, J.L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, **21**, 299–332.

In this paper, the authors offer a procedure for dealing with publication bias, based on modelling its effects on random-effects meta-analytic results. The model, which is based on one-tailed *p*-values, can be used to assess the plausibility of the existence of publication bias, as well as to estimate effects, corrected for the operation of bias. The authors also provide the results of a simulation study used to test their model. These results indicate that the model is reasonably accurate under plausible conditions. Highly statistical.

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.

This paper introduces a very widely used statistical test for the assessment of funnel plot asymmetry. It presents a comparison of cases in which a single large trial and a meta-analysis were conducted on the same question. In no case where the large trial and the meta-analysis agreed was there evidence of funnel plot asymmetry. However, when the large trial and the meta-analysis disagreed, half the time there was funnel plot asymmetry. The paper also briefly reviews potential causes of asymmetry in addition to publication bias and cautions against assessing funnel plot asymmetry in meta-analysis with few studies.

Stern, J.M. and Simes, R.J. (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal*, **315**, 640–645.

This is one of the first studies to demonstrate the impact of publication bias on time to publication as well as on the likelihood of publication. Using a cohort of 743 studies submitted to a hospital ethics committee over 10 years, this study provides clear evidence that in addition to having a lower probability of publication, papers that fail to reach statistical significance experience serious publication delays compared with similar studies with statistically significant findings.

Whitehead, A. (1997). A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in Medicine*, **16**, 2901–2913.

This is one of the first studies to propose that prospective meta-analysis be applied to sets of broadly similar clinical trials that are being conducted at approximately the same time. The author suggests that the advantages of this approach include rapid answers to questions of safety or efficacy, and the facilitation of assessment of subgroup differences in treatment effects, as well as the early identification of adverse effects. The author indicates that a random-effects combined analysis, within a sequential framework, is the appropriate method for analysing these data.

Auger, C.P. (1998). *Information Sources in Grey Literature*, 4th edition. London: Bowker Saur.

This book is considered the classic introductory guide to grey literature. It describes what grey literature is and how it can be identified. The book discusses the types of publications included in the term 'grey literature' and covers collection/acquisition, bibliographic control, cataloguing/indexing, and distribution methods. It also discusses the grey literature in six specific content areas. This book may be of limited interest to those who have substantial experience with grey literature.

Ioannidis, J.P. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Journal of the American Medical Association*, **279**, 281–286.

This paper examines the relationship between statistical significance and time to publication in a series of HIV trials. The author found that statistically significant findings were submitted for publication significantly more rapidly after the trial was completed than was the case for trials that did not achieve statistically significant results, and that such findings, were published more rapidly after they were submitted.

Berlin, J.A. and Colditz, G.A. (1999). The role of meta-analysis in the regulatory process for foods, drugs, and devices. *Journal of the American Medical Association*, **281**, 830–834.

In this paper, the authors discuss the use of meta-analysis in evaluating the efficacy of drugs. The paper provides one of the earliest allusions to prospective meta-analysis. The authors claim that 'Preplanned meta-analysis of individual

trials with deliberately introduced heterogeneity may maximize the generalizability of results from randomized trials'. They suggest that meta-analysis may be particularly helpful in identifying adverse effects, and in identifying particular persons, settings and conditions in which a drug may be particularly effective or ineffective.

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

This paper introduces the trim and fill method as a means of estimating the number of 'missing studies' in a meta-analysis, and for adjusting the mean effect size and confidence intervals accordingly. Examples are given based on several data sets from medicine and epidemiology. Although the method itself is quite easy to grasp conceptually, the article contains a substantial amount of statistics.

Hahn, S., Williamson, P.R., Hutton, J.L., Garner, P. and Flynn, E.V. (2000). Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Statistics in Medicine*, **19**, 3325–3336.

In this paper, the authors discuss the potential for bias of meta-analytic results due to selective reporting of subgroup data. The authors present a method of sensitivity analysis that imputes data for missing subgroups that can be used to assess the robustness of the results and conclusions of systematic reviews that analyse subgroup data. They illustrate their method with reference to a published systematic review. The review in question addressed malaria chemoprophylaxis in pregnancy, and had concluded that benefits were limited to women who were pregnant for the first time. This conclusion was based on subgroup analysis using the three trials out of five which reported on subgroups. The authors' reanalysis suggested that the effect size reported in the original review probably overestimated the actual effect, and called into question the conclusion that the treatment benefited only first-time pregnant women.

Hutton, J.L. and Williamson, P.R. (2000). Bias in meta-analysis due to outcome variable selection within studies. *Applied Statistics*, **49**, 359–370.

This article is one of the earliest to focus specifically on publication bias within studies, rather than on entirely missing studies. The authors describe the potential effects of bias due to multiple testing of outcomes, and selective reporting based on significance levels or effect size. The authors demonstrate the operation of selective reporting by reanalysing two meta-analyses, where it was clear that more outcomes had been measured than were reported, and show that in one of the two cases selective reporting bias threatened the conclusions of the original meta-analysis.

McAuley, L., Pham, B., Tugwell, P. and Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet*, **356**, 1228–1231.

This paper examines whether the inclusion or exclusion of grey literature in meta-analyses affects the estimates of effects of interventions assessed in randomized trials. The authors found that one-third of the meta-analyses they examined included some grey literature, but that there was a large amount of variation in the proportion of 'grey' studies across meta-analyses. On average, published studies yielded significantly larger estimates of effects than did the 'grey' studies. The authors conclude that 'the exclusion of grey literature from meta-analyses can lead to exaggerated estimates of intervention effectiveness', and suggest that systematic reviewers should attempt to search for and include grey literature in their meta-analyses.

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.

This paper describes the results of simulations performed to assess the power of a weighted regression method and a rank correlation test in the presence of no bias, moderate bias and severe bias. The power to detect bias increased with increasing numbers of trials for both methods. The rank correlation test was less powerful than the regression method for both moderate and severe bias. On the other hand, the regression method produced higher than desirable false positive rates under some conditions. The authors suggest that when evidence of small-study effects is found, publication bias should not be the only possible explanation considered.

Sutton, A.J., Duval, S.J., Tweedie, R.L., Abrams, K.R. and Jones, D.R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal*, **320**, 1574–1577.

This study analysed 48 reviews from the Cochrane Database of Systematic Reviews that contained 10 or more individual studies and used a binary endpoint. The authors found that, using the trim and fill method, approximately half the reviews were missing studies and, in nearly 20 % of the reviews, the number missing was significant. In four cases, statistical inferences about the intervention's effect were altered after publication bias was adjusted for. Although in most cases publication biases did not affect the review's conclusions, the authors recommended that researchers should routinely check whether the conclusions of systematic reviews are robust to the operation of these biases.

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

This paper proposes a sensitivity analysis in which different patterns of selection bias can be tested against the fit to the funnel plot. The authors illustrate with two medical examples: passive smoking and coronary heart disease; and the effectiveness of prophylactic antibiotics in critically ill adults. An appendix lists the S-Plus code needed for carrying out the analysis. Highly statistical.

Lefebvre, C. and Clarke, M. (2001). Identifying randomised trials. In M. Egger, G.D. Smith and D.G. Altman (eds). *Systematic Reviews in Healthcare: Meta-analysis in Context*, pp. 69–86. London: BMJ Books.

This book chapter provides a useful overview of the sources that have contributed to the Cochrane Collaboration's main database, the Central Register of Controlled Trials. In addition, it describes how these sources have been searched. The chapter serves as a practical account of the issues to keep in mind when searching for randomized trials for systematic reviews.

Macaskill, P., Walter, S.D. and Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics In Medicine*, **20**, 641–654.

This study compares the performance of three methods of testing for bias: a rank correlation method; a simple linear regression of the standardized estimate of treatment effect on the precision of the estimate; and a regression of the treatment effect on sample size. These methods were tested using simulated meta-analyses of studies with binary endpoints. The results indicated that there was no 'winning' method. Performance varied depending upon the magnitude of the true treatment effect, distribution of study sizes and whether one- or two-tailed tests of significance test were used. In general, all methods suffered from low power in meta-analyses where the number of studies was typical of those found in the medical literature. Higher power was related to higher type I error rates. According to the authors, regressing the treatment effect on sample size, weighted by the inverse variance of the logit of the pooled proportion, is the preferred means of testing for bias.

Pham, B., Platt, R., McAuley, L., Klassen, T.P. and Moher, D. (2001). Is there a 'best' way to detect and minimize publication bias? An empirical evaluation. *Evaluation and the Health Professions*, **24**, 109–125.

This paper compares the performance of file-drawer analysis, the Begg test, the Egger test, trim and fill, and weighted estimation methods to assess the robustness of meta-analytic findings, and to detect and minimize publication bias effects in meta-analyses. The authors found that different approaches to dealing with publication bias reached different conclusions, when applied to the same set of meta-analytic data. This paper does not require advanced knowledge of statistics.

Pigott, T.D. (2001). Missing predictors in models of effect size. *Evaluation and the Health Professions*, **24**, 277–307.

The author of this paper reviews commonly used methods for dealing with missing data and their application to meta-analysis, such as complete case analysis and mean substitution, and suggests that they often yield biased estimates. The article briefly reviews the effects of missing predictors on the results of meta-analyses, discusses the strengths and weaknesses of commonly used missing-data methods, and suggests more desirable ways of handling missing predictors.

Specifically, the author recommends the use of maximum likelihood methods for multivariate normal data and multiple imputation methods.

Jennions, M.D. and Møller, A.P. (2002). Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. *Biological Review*, **77**, 211–222.

This article demonstrates the existence of publication bias in systematic reviews in ecology and evolution. The authors used the trim and fill method to examine the results of 40 published meta-analyses in this field. They found that for random-effects meta-analyses, 38 % had a significant number of 'missing' studies, and that after correcting for potential publication bias, approximately 20 % of weighted mean effects were no longer statistically significant. The authors conclude that in ecology and evolution, publication bias may affect the main conclusions of at least 15–20 % of published meta-analyses, and suggest that researchers routinely examine their results for the possible effects of publication bias.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley.

This is the classic reference for dealing with missing data. The authors focus on imputation, and therefore present a variety of imputation methods, in addition to the bootstrap, jackknife, and similar techniques. The theory behind each technique, as well as algorithms for implementation, is clearly laid out. It is written at the level of a graduate-school textbook. Although no consideration of how such techniques could be applied to meta-analysis is given, the book describes a lot of the statistical ideas underlying missing-data problems.

Manheimer, E. and Anderson, D. (2002). Survey of public information about ongoing clinical trials funded by industry: evaluation of completeness and accessibility. *British Medical Journal*, **325**, 528–531.

This paper reports a study of the completeness of on-line US-based trials registries. The authors examined whether ongoing trials of experimental drugs for colon and prostate cancer were reported in publicly accessible on-line trials registries. They found that 'a substantial proportion' of these trials were not contained in any of the publicly available registries. The authors concluded that there is a need for a comprehensive on-line registry of trials that includes Phase 3 industry-sponsored research.

Olson, C.M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J., Zhu, Q., Reiling, J. and Pace, B. 2002. Publication bias in editorial decision making. *Journal of the American Medical Association*, **287**, 825–828.

This paper provides evidence that a major reason for publication bias in medical research is that researchers are less likely to submit manuscripts reporting non-significant results to journals. Based on a sample of papers submitted to *JAMA*, the authors found no evidence that publication bias occurs after manuscripts have been submitted for publication.

Scherer R.W. and Langenberg, P. (2002). Full publication of results initially presented in abstracts (Cochrane Methodology Review). In *The Cochrane Library*, Issue 3, 2002. Oxford: Update Software.

The authors synthesized the results of 46 reports that examined the publication of articles based on abstracts that had been presented at scientific conferences. Overall, fewer than half of the abstracts were eventually published as full articles. Full publication of studies was more likely among abstracts with statistically significant results than among those with non-significant results. Abstracts of results of clinical research were less likely to be published than abstracts of the results of basic research, while abstracts of studies using randomized designs were published at higher rates than were other types of studies.

Song, F., Kahn, K.S., Dinnes, J. and Sutton, A.J. (2002). Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *International Journal of Epidemiology*, **31**, 88–95.

This is one of the first studies to assess the potential effects of publication bias in studies of diagnostic test research – empirical studies of publication having mainly focused on studies of treatment effect. The authors examined a sample of 28 meta-analyses of diagnostic accuracy from the Database of Abstracts of Reviews of Effectiveness (DARE). They found that, in general, the authors of the meta-analyses had not sufficiently considered either literature search strategies that would minimize publication bias or the impact of possible publication bias. Results showed that in a substantial proportion of the meta-analyses evaluated, smaller sample sizes were associated with greater diagnostic accuracy and greater funnel plot asymmetry. In addition, the fewer the literature databases searched, the greater the funnel plot asymmetry in meta-analyses. The authors suggest that authors of systematic reviews of diagnostic tests should perform more comprehensive literature reviews, and assess the likelihood and severity of publication bias.

Stewart, L.A. and Tierney, J.F. (2002). To IPD or not to IPD? *Evaluation and the Health Professions*, **25**, 76–97.

This paper describes the history, rationale and the pros and cons of the individual patient data approach to meta-analysis. These are discussed in terms of issues relating to data quality, analysis, and the organizational and collaborative approach. The authors conclude that reviewers should, at the outset, consider the methodological factors likely to influence or bias the outcome of their review, together with time and resource considerations, in order to take an active decision about the most appropriate approach.

Egger, M., Jüni, P., Bartlett, C., Holenstein, F. and Sterne, J. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment*, **7**, 1–76.

This paper describes the importance of comprehensive literature searches which cover the grey literature and all relevant databases and languages. It also makes an argument for the importance of assessing trial quality in systematic reviews.

Hopewell, S., McDonald, S., Clarke, M. and Egger, M. (2003). Grey literature in meta-analyses of randomized trials of health care interventions (Cochrane Methodology Review). In *The Cochrane Library*, Issue 4, 2003. Chichester: John Wiley and Sons, Ltd.

This paper describes a systematic review of studies comparing the impact of grey versus published literature on the overall results of meta-analyses. It shows that the exclusion of studies from the grey literature may lead to an exaggeration of the effects of treatment.

MacLean, C.H., Morton, S.C., Ofman, J.J., Roth, E.A. and Shekelle, P.G. (2003). How useful are unpublished data from the Food and Drug Administration in meta-analysis? *Journal of Clinical Epidemiology*, 56, 44–51.

This paper uses the example of research on non-steroidal anti-inflammatory drugs (NSAIDs) to examine whether studies summarized in Food and Drug Administration (FDA) reviews are eventually published, and to compare key characteristics of FDA-reviewed studies with those reported in published, peer-reviewed literature. The authors found that of 37 studies described in the FDA reviews, only one was published. They also found that there were no meaningful sample or methodological differences between FDA and published studies, and that the effect sizes found in both FDA and published studies were neither 'significantly [n]or practically' different. They concluded that FDA reviews could be a source of data for systematic reviews.

Melander, H., Ahlqvist-Rastad, J., Meijer, G. and Beermann, B. (2003). Evidence b(i)ased medicine – selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *British Medical Journal*, **326**, 1171–1173.

These authors examined publication bias related to multiple publication, selective publication and selective reporting in 42 studies of five selective serotonin reuptake inhibitors that were conducted by pharmaceutical companies in Sweden between 1983 and 1999. They compared reports of 42 studies as they were submitted to the Swedish drug regulatory authority with the published versions of these studies. Their results showed that 21 of the studies contributed to at least two publications each; that studies with significant effects were published as stand-alone publications more frequently than those with non-significant results; that many publications ignored the results of intention to treat analyses and presented only the more positive per-protocol analyses. However, the degree of multiple publication, selective publication and selective reporting differed across drugs. The authors concluded that publicly available data are likely to be biased.

Terrin, N., Schmid, C.H., Lau, J. and Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22, 2113–2126.

This study uses simulation to evaluate the accuracy of the trim and fill method when studies are heterogeneous. The results indicate that when studies are heterogeneous, trim and fill may wrongly adjust for publication bias when there is none. The authors suggest that funnel plots may not be appropriate for heterogeneous meta-analyses, and suggest that in cases of heterogeneity, selection modelling may be a better approach. The authors report that a selection model was superior to trim and fill in their simulations, although the results converged at times, under some conditions.

Bennett, D.A., Latham, N.K., Stretton, C. and Anderson, C.S. (2004). Capture–recapture is a potentially useful method for assessing publication bias. *Journal of Clinical Epidemiology*, **57**, 349–57.

This paper compares several approaches to the assessment of publication bias, including capture–recapture, visual examination of funnel plots, the Egger test, a funnel plot regression, trim and fill, and a selection model approach. In the illustrative example, all methods employed yielded broadly consistent results. Capture–recapture estimated that three relevant studies were missed, while trim and fill estimated that there were 16 missing studies. Both the Egger test and a funnel plot regression approach indicated the presence of publication bias, while selection modelling suggested that the observed funnel plot asymmetry observed was not entirely the result of publication bias. The authors suggest that capture–recapture is a potentially useful means of assessing publication bias, but that further simulation studies on all the methods should be conducted.

Chan, A.-W., Hrobjartsson, A., Haahr, M.T., Gøtzsche, P.C. and Altman, D.G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association*, **291**, 2457–2465.

In this paper, the authors demonstrate the prevalence of incompletely reported outcomes in Danish randomized trials from 1994–95, by comparing protocols with journal articles and by surveying researchers. They identified 102 trials with 122 published journal articles and 3736 outcomes. Overall, half of efficacy outcomes and nearly two-thirds of harm outcomes per trial were incompletely reported. Statistically significant outcomes were much more likely to be fully reported than non-significant outcomes for both efficacy and harm. When published articles were compared with their protocols, 62 % of trials altered, added or omitted at least one primary outcome. Eighty-six per cent of survey respondents (42/49) denied the existence of unreported outcomes, despite the evidence. The authors conclude that reporting of trial outcomes is often incomplete, biased and inconsistent with protocols. As a result, they caution, published articles, as well as reviews that include them, may overestimate the benefits of interventions. The authors issue a call for the registration of planned trials and public availability of protocols before trial completion.

Trikalinos, T.A., Churchill, R., Ferri, M., Leucht, S., Tuunainen, A., Wahlbeck, K. and Ioannidis, J.P.; EU-PSI project (2004). Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology*, **57**, 1124–1130.

The authors investigated whether the certainty and estimates of effect size of mental health interventions change over time, as additional trials appear on the same topic. This sort of evolution of effect sizes over time had previously been found in trials in genetics, but had not been examined in mental health intervention trials. Using cumulative meta-analysis and recursive cumulative meta-analysis, they examined 100 meta-analyses containing five or more trials each, published in at least three different years. Outcomes included death, relapse, failure or dropout. The authors found that eight meta-analyses reached statistical significance at some point, but became non-significant as more trials were published. In general, large effect sizes in early trials were reduced as further evidence accrued. The authors concluded that, in mental health as in other areas, evidence based on a small number of randomized subjects should be interpreted cautiously, and that early estimates of treatment effectiveness may be overly optimistic.

Chan, A.-W. and Altman, D.G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *British Medical Journal*, **330**, 753.

In this paper, the authors examine the extent of selective reporting of outcomes in published randomized trials in health care by reviewing publications, and then surveying their authors. They conclude that the published medical literature represents a biased subset of study outcomes. The sample for this study consisted of all journal articles reporting on randomized trials from journals that are indexed in PubMed, and which were published in December 2000. The authors identified 519 trials with 553 publications and 10 557 outcomes. They found that authors who responded to their survey were often unreliable about the presence of unreported outcomes. On average, over 20 % of the outcomes per trial were incompletely reported. The reasons most commonly given for omitting outcomes included space limitations, and lack of clinical or of statistical significance.

Ioannidis, J.P.A. and Trikalinos, T.A. (2005). Early extreme contradictory estimates may appear in published research: The Protens phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, **58**, 543–549.

In this paper, the authors make the argument that early hypothesis-generating research in areas where a lot of data can be generated quickly is prone to produce early studies with extreme, opposite results, due to the preferences of authors and editors.