Outcome selection bias in meta-analysis

PR Williamson, C Gamble Centre for Medical Statistics and Health Evaluation, University of Liverpool, Liverpool, UK, DG Altman Centre for Statistics in Medicine, Oxford, UK and JL Hutton Department of Statistics, University of Warwick, Coventry, UK

Publication bias has been previously identified as a threat to the validity of a meta-analysis. Recently, new evidence has documented an additional threat to validity, the selective reporting of trial outcomes within published studies. Several diseases have several possible measures of outcome. Some articles might report only a selection of those outcomes, perhaps those with statistically significant results. In this article, we review this problem while addressing the questions: what is within-study selective reporting? how common is it? why is it done? how can it mislead? how can it be detected?, and finally, what is the solution? We recommend that both publication bias and selective reporting should be routinely investigated in systematic reviews.

1 Introduction: potential problem

A systematic review has been defined as ‘a review that has been prepared using a systematic approach to minimising biases and random errors which is documented in a materials and methods section’. Cochrane systematic reviews of randomized trials are internationally recognized as one among the best sources, if not the best source of reliable, up-to-date information on health care. The aim of the Cochrane Collaboration is ‘to help people to make well informed decisions about health care by preparing, maintaining and promoting the accessibility of systematic reviews’. Issue 3, 2004 of the Cochrane Library contains 2074 reviews and 1485 protocols. NICE technology appraisals provide guidance on the use of new and existing medicines and treatments within the NHS in England and Wales. During the year 2002–2003, 19 such reviews were published. Thus, the methodology of systematic reviewing is now widespread.

Meta-analysis of data, a statistical technique for combining results from several related studies, may or may not be undertaken in a particular review. Meta-analysis as part of a systematic review has led to important new findings in medical research, for example, the increased risk of death following administration of human albumin solution. This research, and concerns expressed about the choice of outcomes used in the meta-analysis, led to further studies and hence to official clinical guidelines in the UK (NICE Technology Appraisal Guidance 74, issued in January 2004, www.nice.org.uk/TA074guidance). Meta-analysis also often shows that there is evidence to support treatments not widely used or that evidence is lacking to support treatments that are widely used nowadays.

Address for correspondence: PR Williamson, Centre for Medical Statistics and Health Evaluation, Shelley’s Cottage, Brownlow Street, University of Liverpool, Liverpool L69 3GS, UK. E-mail: prw@liv.ac.uk
Empirical research consistently suggests that published work is more likely to be statistically significant ($P < 0.05$) than unpublished research. We call this phenomenon, by which an entire study is either published or not depending on its results, as between-study selective reporting bias. Such bias has been recognized as a potential threat to the validity of any meta-analysis. Evidence also suggests that published research without statistically significant results takes longer time to achieve publication than its statistically significant counterparts, further biasing evidence over time. There is now widespread acceptance of the need to search for unpublished studies, and several methods to look for evidence of possible publication bias in a set of studies included in a systematic review.

Within-study selective reporting bias has been defined as the selection, on the basis of the results, of a subset of the analyses undertaken to be included in a study publication. This may relate to the selection of outcomes, including outcome subscales, endpoint score versus change from baseline, the cutoff selected for dichotomizing a continuous measure, and the time point on which to focus when the same outcome has been measured at multiple time points. It may equally apply to the selection of subgroups, prognostic factors, first period results in crossover trials and so on.

There are often several alternative measures of the health of people, particularly those with chronic diseases, and reports of clinical trials may give results for only a subset of those recorded. For example, a review of schizophrenia trials reported that 640 different instruments were used. The authors found that ‘most trials used between one and five instruments, but greater numbers were common, with one trial using 17 different outcome scales.’ In a study, it may be that several significance tests were carried out and more significant outcomes were reported. Marshall et al. found that a treatment effect was more likely to be found when unpublished, rather than published, rating scales were used in schizophrenia trials. They hypothesised that data from unpublished scales may be less likely to be published when they are not significant or that following analysis, unfavourable items may have been dropped to create an apparent effect.

Systematic reviews of randomized trials can lead to important inferences, providing evidence to support treatments not widely used or showing that evidence is lacking to support treatments that are widely used nowadays and thus it is essential to identify any potential bias in the approach to the analysis. Although bias in meta-analysis arising from selective publication between studies has been investigated, within-study selection of results received little attention until recently. In the rest of this article, we review our current knowledge concerning the problem, with randomized clinical trials as our main focus of interest.

### 2 Evidence of within-study selective reporting

A small study involving a single Local Research Ethics Committee considered a complete cohort of applications approved in one year. The primary outcome was stated in only six of the protocols for the 15 publications obtained. Eight protocols made some reference to an intended analysis, but seven of the publications did not follow this analysis plan. Within-study selective reporting was evident or suspected in several trials included in a review of a cohort of five meta-analyses on the Cochrane
Other researchers have also concluded the problem to be major and deserving substantially more attention than it currently receives. However, the most convincing evidence has come to light only recently. This research provides direct empirical evidence for the existence of within-study selective reporting bias. In the first study, 102 trials with 122 publications and 3736 outcomes were identified. Overall, (a median of) 38% of efficacy and 50% of safety outcomes per parallel group trial were incompletely reported for meta-analysis. Statistically significant outcomes had a higher odds ratio of being fully reported when compared with non-significant outcomes for both efficacy (pooled odds ratio 2.4; 95% confidence interval 1.4–4.0) and safety (4.7, 1.8–12) data. In addition, when comparing publications with protocols, 62% of trials had at least one primary outcome that was changed, introduced or omitted. A second study of 48 trials funded by the Canadian Institutes of Health Research found closely similar results. A third study, involving a retrospective review of 519 trial publications and a follow-up survey of authors, concluded that 'Incomplete reporting of outcomes within published reports of a randomised trial is common and is associated with statistical non-significance.' For all the three studies, authors were asked whether there were unpublished outcomes, whether those showed significant differences and why those outcomes had not been published. The most common reasons for nonpublication of results were ‘lack of clinical importance’ or lack of statistical significance.

2.1 Identification of within-study selective reporting for an individual study

A study report may provide direct or indirect evidence of selective reporting. Direct evidence would include a statement that results were not reported because ‘$P > 0.05$’ or that several analyses of the same data were undertaken, but only the most significant were reported. Indirect evidence may suggest that the outcome was measured, although no or only partial data were reported, thus raising the level of suspicion. Ideally, the authors should be contacted to clarify the situation, however, where this is not possible, the following should be considered. The trial report should be scrutinized to see whether there is any indication that the data were collected and/or analysed. For example, if cause-specific mortality is reported, then overall mortality must have been measured, even if not reported. If the trial protocol is available, analyses specified in the protocol, but not reported, should be identified. Finally, knowledge of the clinical area may suggest that it is likely that the outcome was measured. For example, in a particular disease, a number of tests may be usually undertaken together, such that if one outcome is reported, but another is not then, one should be suspicious that the latter may have been selectively not reported. It is probable that in many cases, clinical judgement of whether the outcome was measured will be required.

If it has been established or is suspected that an outcome was measured, then the question of whether the results were selectively not reported needs to be considered. If the missing outcome is an efficacy outcome, and reported efficacy outcomes tended to be statistically significant, the level of suspicion would be raised.

It is important to distinguish between unreported and incompletely reported outcomes. The latter is still likely to cause problems in a meta-analysis but is less likely to result in bias.
2.2 Effect of and adjustment for within-study selective reporting of outcomes

Studies often measure several outcomes related to a particular domain. For example, measurements of urethral closure pressure may be taken in a prone or supine position and summarized in terms of the average or maximum value. Analyses of each such measurement may be undertaken but only the most significant were reported.

To understand the effect of selecting one outcome to report from several possible outcomes, such as various incontinent measurements, we consider the implications for estimates of the mean and the variance of the relative treatment effect.\textsuperscript{16} If we had two independent normally distributed outcomes, say X and Y, and we were comparing treatments, say A and B, which were equally effective, then the mean treatment effects would be equal. Hence, the difference in mean values would be zero. For simplicity, let the variances be one, and consider what difference in mean values we would get if we always select the larger of the mean differences for the two outcomes considered. Figure 1 shows the distribution of a standard Gaussian random variable and the distribution of the maximum of two independent \( N(0, 1) \) variables is concentrated above zero, with a mean of 0.56. That is, when there is no difference between treatments, selective reporting of one out of the two independent outcomes would give an estimate more than half a standard deviation above zero. The narrower bell of the distribution of the maximum shows that the standard deviation is smaller, so a test based on the ratio of the mean to the standard deviation would have a lower \( P \)-value (be more significant) than the correct test. If a meta-analysis was based on such estimates from several trials, the biases would be combined.\textsuperscript{16} The bias increases as the number of

![Figure 1: Density functions: \( N(0, 1) \) and maximum of two independent \( N(0, 1) \) variables.](image-url)
outcomes selected increases. The mean values of outcomes per trial in two recent studies\textsuperscript{23,24} were as high as 37 and 29.37.

It is reasonable to expect that different outcomes intended to measure aspects of the same condition will be positively correlated. This will reduce the bias arising from selective reporting. However, unless the outcomes are very highly correlated ($P > 0.95$), the bias will still substantially distort estimates of treatment effects and significance levels.\textsuperscript{16} Care is needed to ensure that similar results from different, correlated outcomes are not interpreted as separate evidence of effect.

When studies report different actual measurements, a shared statistic is required to facilitate meta-analyses. The effect size, that is, standardized mean difference, is widely used.\textsuperscript{26} The effect on a meta-analysis of selection of results under particular assumptions has been investigated theoretically and shown to be substantial when the relative weight given to the selectively reported trial is high, the number of variables selected from is high and the correlation between variables is low.\textsuperscript{16}

One approach is to adjust the mean and variance of the relative treatment effect in the study, where selective reporting is suspected.\textsuperscript{16} If selective reporting is thought likely to affect all studies, results can be compared with significance levels, on the basis of the distributions of maxima.\textsuperscript{16} Both significance testing and estimation of effect sizes in the presence of selection bias have been addressed, in terms of the robustness of the inferences drawn. As the actual biases will be generally unknown, the sensitivity analyses are not intended to provide ‘corrected’, let alone ‘correct’ results.

3 Within-study selective nonreporting of outcomes

Results may also be selectively nonreported due to lack of statistical significance.\textsuperscript{20} A review of beta-lactam and beta-lactam–aminoglycoside combination therapy for cancer patients with neutropenia found a significant benefit for combination treatment in terms of treatment failure (log odds ratio 0.22), but a marginally significant detrimental effect in terms of mortality (log odds ratio $-0.36$). Results for mortality, however, were based on data from only five of the nine eligible trials. Imputing missing data for four trials, under the assumption that selective non-reporting has occurred, shifts the estimate of effect on mortality to $-0.05$, no longer significant.\textsuperscript{20}

3.1 Within-study selection of subgroups

Subgroup analysis is frequently undertaken in randomized trials, but complete subgroup data are not always reported.\textsuperscript{27} Subgroup analyses with significant results are more likely to be reported. In a meta-analysis, reviewers usually calculate estimates, on the basis of those trials providing sufficient data for the subgroup of interest. Such estimates will be biased, if reporting of subgroups depends on the results of significance tests. Of course, if a trial did not collect data on the defining characteristics of the subgroups, then there is no mechanism to create bias.

It is sometimes possible to recreate the data from partial results on subgroups. This is the case for binary outcomes and binary subgroups.\textsuperscript{28} The sensitivity analysis for subgroup selection proposed by these authors involves imputing all possible data sets consistent with the information that is provided for each trial. This creates pseudo-data
for each subgroup, with associated significance levels. The assumption of the sensitivity analysis is that trials did not report subgroup results because no subgroup had statistically significant differences in outcome. Therefore, the region of interest contains those pseudo-data sets for which none of the subgroup analyses was significant at the 5% level.

Summary statistics, with significance levels, are calculated, which combine the reported subgroups with all possible combinations of the regions of interest for the trials without subgroup results. Histograms of the odds ratio and \( P \)-values illustrate the spread of results consistent with the available data. A meta-analysis of trials of malaria protection in pregnancy suggested that benefit was limited to primigravida women, but two trials did not report results for primigravidae and multigravidae subgroups. Hahn et al.\(^{28}\) found that the odds ratio of 0.51 for treatment benefit in primigravida women could increase to 0.7 with complete data.

3.2 Detection of within-study selective reporting in a meta-analysis

The first step is to examine whether there are studies eligible for the meta-analysis that do not provide data to be included in the meta-analysis. Clearly, if all eligible trials provide data on the outcomes of interest, no selective nonreporting has occurred. However, this does not rule out the possibility that the outcomes reported have been selected from a number of analyses of the same data, for example, at different time points.

A plot of the relative treatment effect estimate on a particular outcome against a measure of precision for all trials can be a useful second step. If no bias is present, a funnel shape should be evident. If the plot appears asymmetrical or hollow, bias may be suspected. However, it is difficult to distinguish between within- and between-study selective reporting and the other effects of small studies. The number of studies missing from the funnel plot can be estimated using the trim and fill methods,\(^{13,14}\) and this might be compared to the number of studies known to be eligible but with no outcome data reported.

A further useful graphical method, when two outcomes are of interest but one is less frequently reported, is to present the Forest plot for more frequently reported outcome, subgrouped by whether one or both outcomes were reported.\(^{20}\) If no selective reporting has occurred, the pooled estimates for the two subgroups should be similar.

3.3 Impact of within-study selective reporting on a meta-analysis

There is a need to establish how much of a threat within-study selective reporting is to the validity of meta-analysis. The impact of selective nonreporting bias has been examined in a small cohort of meta-analyses within Cochrane reviews.\(^{20}\) This study demonstrated that although within-study selection was evident or suspected in several trials, the impact on the meta-analysis conclusions was minimal. However, the study was limited because the cohort contained only nine meta-analyses and the amount of missing data was low (1–7%). Their motivating example highlighted how selective reporting can have a substantial effect when the amount of missing data is large (48%). However, selective reporting may be more likely in smaller trials which have potential...
for less impact in a meta-analysis, because such studies are given relatively less weight. Some supporting evidence linking sample size to bias exists, but further research is necessary.

4 What is the solution?

The best method of avoiding biases arising from within-study selective reporting is to obtain the original individual patients’ data or results, whenever possible. Two main issues arise: the existence and accessibility of original data and the cost in time and money of obtaining and analysing such data. If an individual patient data analysis is impossible, and there is evidence that some publications have selected the outcomes to report, then the sensitivity of conclusions to plausible biases should be investigated. We recommend sensitivity analyses, which combine commonsense and modelling, to allow the impact of possible selection of measures or subgroups for publication to be assessed.

Subtle bias may enter the review process if a reviewer’s choice of outcomes for a meta-analysis is influenced by their prior knowledge of what has been reported or what has been found to be interesting in the studies that they will meta-analyse.

It is highly desirable to have standard outcomes including standard definitions that are always reported regardless of other outcomes reported. As yet, few medical areas have agreed common outcomes. For example, a review of schizophrenia trials reported that 640 different instruments were used. In rheumatology, there is now an international project to improve outcome measurement (http://www.omeract.org). Common measures for epilepsy have been agreed by the International League Against Epilepsy (http://www.ilae-epilepsy.org). Registration of trial protocols and the extension of a trial amnesty to include an ‘outcome amnesty’ should also be considered. A drive to improve the quality of published research, which may include making the trial protocol available for review at the time the report is submitted to a journal, as previously suggested, may also decrease the likelihood of this form of bias in the future.

5 Discussion

The current evidence indicates that selective reporting in trials is widespread. A small study suggested that although it can have a substantial effect on the conclusions of a meta-analysis, when the amount of missing data is large, the impact on most meta-analysis conclusions was minimal due to a small amount of missing data. However, there is a need to assess the impact of within-study selective reporting in a larger cohort of meta-analyses.

Possibly more important is the need to understand how this bias arises. Medical doctors have been surveyed with respect to unpublished research, but the study did not address within-study sources of bias. Social psychologists have been surveyed, with respect to reporting, following-up or abandoning research when data do or do not reject the null hypothesis. This work ‘indicated a dysfunctional research–publication
system’, and although undertaken 30 years ago, recent evidence suggests that similar problems still exist.

Little research has been undertaken to understand the reasons why study protocols may change during a project, or simply not be adhered to, and researchers’ understanding of the potential sources of bias in the analysis and reporting of data. There are several reasons why a trial protocol may legitimately change during a study. It is important to understand the process from the viewpoint of the researcher, to address areas of misunderstanding or ignorance about the potential sources of bias in the decisions they make. For example, it has been hypothesized that researchers may misunderstand the phrase ‘primary outcome’, by not recognizing that it means the prespecified primary outcome.23 Changes to the study protocol should be reported in publications; none of the 150 studies did so in the two samples of Chan et al.23,24

It seems clear that many researchers fail to appreciate the ethical requirement for researchers to report their research completely and honestly. A recent consensus statement relating to the principles of registration and reporting of clinical trials includes the following statement:

All trial results should be registered and publicly available, along with sufficient protocol information to enable critical assessment of their validity.34

Wide adherence to this principle would have a clear benefit on systematic reviews and meta-analysis.

Systematic reviewers must consider the amount of, and reasons for data potentially missing from a meta-analysis. The most common reasons provided by trialists for not reporting outcomes were lack of statistical significance, journal space restrictions and lack of clinical importance.23

To boost confidence in the review, we recommend that the sensitivity of the conclusions to plausible biases should be investigated. If the conclusions from analyses making different assumptions vary, reviewers should present and interpret correctly both the original meta-analysis which assumes no selective reporting and the sensitivity analysis, including a description of the assumptions made concerning the nature of selection.

Methods developed to date to assess the impact of selective reporting and quantify the size of effect, address both selection of continuous and binary outcomes and incomplete reporting of subgroup analyses and outcomes.16,20,28 Dissemination of recent work, demonstrating direct evidence of selection bias,23,24 will encourage methodological research into models and sensitivity analyses that account for such bias. To inform this modelling, empirical evidence of the nature of the within-study selection process is exigently needed. Empirical evidence has been used to develop and inform models of the selection process that causes between-study publication bias.10–12,15 These methods rely on understanding the selection process to provide adjustments for the bias caused by censoring mechanisms. An increased understanding of the processes resulting in within-study selective reporting is essential to enable this methodology to be extended and developed further in this area.

The focus of this article has been evidence-based medicine, particularly randomized clinical trials. However, it is likely that this problem extends to non-RCTs where it may be even greater as in many such areas, for example prognostic studies, it is not usual to
have a detailed protocol, which may lead to more analyses being undertaken than were specified a priori. In social epidemiological research, often several social characteristics are measured, with the potential for similar bias. Effect sizes are widely used as summaries in social science. Thus evidence-based policy, which derives information from such studies, may also be affected. As evidence-based medicine extends more into chronic conditions and adverse events, within-study selection must receive as much attention as between-study selection.

References


