Cochrane Pregnancy and Childbirth Group
Methodological Guidelines

[Prepared by Simon Gates: July 2009, updated July 2012]

These guidelines are intended to aid quality and consistency across the reviews of the Pregnancy and Childbirth Group. They are intended to give broad guidance on some of the issues that arise in conducting reviews, and where the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins 2011[1]) does not give definite recommendations. They have been revised to take account of revisions to the Cochrane Handbook, the introduction of RevMan 5 (RevMan 2011[2]) and recent advances in systematic review methodology.

They should be used in conjunction with the revised Handbook, especially Chapters 5, 7, 8, 9 and 16, which give further details of many methods.

Types of study included (Handbook section 5.5)

The policy of the Pregnancy and Childbirth Group is to include only randomised or quasi-randomised studies. Observational studies (e.g. cohort or case-control studies) should not be included in meta-analyses or contribute to the results or conclusions of reviews, but may be discussed in the Background and Discussion where relevant.

Randomised controlled trials using all potentially valid types of design (such as parallel group, crossover and cluster randomisation) should be considered for inclusion. Special statistical methods will be needed for cluster and crossover trials (see below), but this should not prevent their inclusion.

Crossover trials are likely to be an invalid design for most Pregnancy and Childbirth Group reviews. They are most suitable for evaluating interventions with a temporary effect in the treatment of stable, chronic conditions, and hence are not suitable for most situations that Pregnancy and Childbirth reviews will address. If trials using a crossover design are found, review authors should consider whether this design is appropriate for the question being investigated; if not, they should be excluded. It may be appropriate to include the data from the first arm of a crossover trial.

Quasi-randomised trials may be included in reviews, if it is felt that they may make a useful contribution to it (for example, if the majority of studies are quasi-randomised, excluding them would discard most of the data). However, if they are included, there should also be a sensitivity analysis by trial quality, separating trials into those with high and low risk of bias (quasi-randomised trials will always be in the high risk of bias group), or a sensitivity analysis excluding trials of low quality.

Studies published only as abstracts present a problem. They may be poorly reported, and contain inadequate information about the studies' methods. Furthermore, there are often differences between data presented in abstracts and those in a subsequent full publication. Including them may therefore risk introducing bias because the studies were of poor quality or did not present the correct results. However, excluding them may also risk introducing bias, because some studies are only ever reported as abstracts, and studies that show no differences may be less likely to achieve full publication. Abstracts should be assessed in the same way as full papers. If there is sufficient information presented in the abstract to demonstrate that it meets the review's inclusion criteria and is of an acceptable methodological standard, it should be included in analyses. Many abstracts may provide inadequate or no information about aspects such as eligibility criteria, interventions, randomisation methods and withdrawals or post-randomisation exclusions. If there are doubts about the eligibility of the study or it is thought to be at risk of serious bias, it should be excluded with a note in the excluded studies table explaining that it will be reconsidered for inclusion once the full publication is available, or the authors have provided more information.

Types of outcomes (Handbook section 5.4)

The most important principles in deciding on the outcome measures for a review are that:

1. they should be kept to the minimum number necessary;
2. each outcome should have a clear rationale, explained in the Background.

Large numbers of outcome measures are likely to produce spuriously significant results (Handbook Section 16.7); moreover an excessive number of outcomes will make the review unwieldy and confusing, as conflicting results from different outcomes are likely to arise by chance. It will also become impossible to perform adequate investigations of issues such as publication bias and heterogeneity if the number of analyses in the review is too great. The Handbook (section 5.4.2) suggests that there should be no more than seven main outcomes for a review, which should be divided into a small number of primary outcomes (the Handbook suggests about three), which are the most important outcomes on which the review’s conclusions will be based, and additional secondary outcomes.

In Pregnancy and Childbirth reviews, there is often pressure to include a large number of outcomes because there may be outcomes for both mother and baby, short- and long-term effects are frequently both important, and interventions may have non-specific effects. For interventions that affect both mother and baby, it is reasonable to have a set of primary and secondary outcomes for each, as long as there is a clear justification for this.

Outcomes should be divided into:

1. primary outcomes: a small number of the most important outcomes. The review’s main conclusions and recommendations should be based on the primary outcomes;
2. secondary outcomes; the other prespecified outcomes;

3. non-prespecified outcomes: any additional outcomes that are included in the review but were not specified as primary or secondary outcomes. Non-prespecified outcomes should be clearly labelled as such in the Results and should not be used for the main conclusions.

Where interventions have non-specific effects and so may affect many outcomes, it may be preferable to use a composite outcome as a primary outcome for the review; for example, “serious neonatal morbidity” could be used to measure possible harms of an intervention, if there is not a strong reason to expect it to be associated with one particular adverse outcome. Individual components of the primary outcome could be included as secondary outcomes if necessary.

In the results, primary and secondary outcomes should be clearly identified (for example, by structuring the results so that primary and secondary outcomes are in separate sections), and the total number of meta-analyses performed in the review should be stated.

Assessing risk of bias (Handbook Chapter 8)

The risk of bias in the included studies must be taken into account when drawing conclusions from the studies in the review; therefore, thorough assessment of bias risk is essential. The *Cochrane Handbook for Systematic Reviews of Interventions* now includes detailed guidance on assessing studies for risk of bias. An assessment tool for risk of bias has been developed for The Cochrane Collaboration and is implemented in RevMan 5. This assesses bias risk in six domains (sequence generation, allocation concealment, blinding of participants, personnel and outcome assessors, incomplete outcome data, selective outcome reporting and other sources of bias), and includes criteria for judging studies to be at high or low risk of bias. A risk of bias table for each study should be included in the ‘Characteristics of included studies’.

It is reasonable to exclude outcome data from analyses if they have an unacceptably high risk of bias. What is judged “unacceptably high” risk of bias will vary between reviews, but the criteria for excluding data from analyses need to be specified in the protocol.

A commonly-used criterion for exclusion of outcomes from meta-analyses is missing data of more than 20% of the randomised sample. However, this is not based on any empirical evidence and cannot be a general recommendation. In some circumstances it may be acceptable to include studies with more than 20% missing data; for example, studies of disadvantaged populations, or long-term follow-up studies may frequently experience losses of greater than 20%, and it may be preferable to allow a greater bias risk in these analyses rather than exclude a large proportion of the existing data.

Measures of treatment effect

Summary statistics: dichotomous outcomes (Handbook Section 9.2.2)

Risk ratios (relative risks) are the preferred summary statistic for dichotomous outcomes because of their ease of interpretation. However, in some circumstances there may be good reasons for preferring a different statistic, e.g. the Peto odds ratio appears to perform best when data are very sparse.

Summary statistics: continuous outcomes (Handbook Section 9.2.3)

The mean difference should be used if the outcomes are measured in the same way between trials. The standardised mean difference (SMD) can be used to combine trials that measure the same outcome using different methods (e.g. two different scoring systems to measure developmental quotient (DQ)).

A frequent error in trial reports is presentation of the standard error of the mean (SEM) rather than the standard deviation. This is much smaller, and will, if used instead of the correct standard deviation in a meta-analysis, give far too much weight to that study. If one study appears to have a much smaller standard deviation than the others, it may be that the SEM has been erroneously reported as the standard deviation. SEM can be converted to standard deviation by multiplying it by the square root of the sample size. If standard deviations are not reported, it may be possible to calculate them from statistics given in the paper: the Handbook (section 7.7.3) presents methods for doing this.

If there is evidence of skew in continuous data (*see* Handbook section 9.4.5.3) the methods in RevMan may be unreliable. At present there are no straightforward methods for dealing with skewed data, so the problem should be noted in the text of the review. Alternatively, in some cases it may be possible to reduce skew by transforming the variable (though this may require either further information from the trialists, or individual patient data), or by converting a skewed continuous outcome into a dichotomous or other type of variable.

Unit of analysis issues

Cluster-randomised trials (Handbook Section 16.3)

Cluster-randomised trials should be included in analyses with individually randomised trials. The Handbook describes two methods for doing this (Sections 16.3.4 and 16.3.6). The method described in 16.3.6 is slightly preferable as it involves less approximation.

Cluster-randomised trials should never be included in meta-analyses without adjustments, as if they were individually randomised. Doing so will overestimate the sample size, give too much weight to the study and give confidence intervals for the overall estimate that are too narrow.

Incorporation of cluster-randomised trials requires an estimate of the intracluster correlation coefficient (ICC). Ideally this will be reported in the trial report for the outcome of interest. Often it will not be, in which case an estimate from (in descending order of desirability) (a) the same study but for a different outcome; (b) a similar randomised trial; (c) another study of a similar population; (d) an estimate; may be used. In all cases, a sensitivity analysis should be performed to investigate the effects of variation in the value of the ICC on the review’s results.

A subgroup analysis separating the included studies by type of design (individually randomised versus cluster-randomised) should also be performed to investigate possible relationships between treatment effect and randomisation unit.

Assessment of risk of bias for cluster-randomised trials needs slightly different methods, described in Handbook Section 16.3.2.

Crossover trials

There will be few situations in Pregnancy and Childbirth group reviews where crossover trials are an appropriate design, hence they are usually expected to be excluded. If they are included, the Handbook Section 16.4 has details of methods for bias risk assessment and analysis, which should be included in the protocol and full review.

Multiple pregnancies

Many trials and reviews include women with multiple pregnancies, and have outcomes both for the mother and baby. This raises a problem because when there are multiple pregnancies, the numbers of mothers and babies are different, meaning that different denominators could be used in the analysis. Moreover, babies from the same pregnancy cannot be regarded as independent.

In this situation, the review authors should consider for each outcome whether the appropriate denominator is the number of babies or the number of women. For most neonatal outcomes the number of babies will be the appropriate denominator, and for most maternal outcomes, the number of women will be appropriate. For example, "caesarean section" would usually be analysed most appropriately using the number of women as denominator, because each woman will have only one operation, regardless of how many babies are delivered this way. Counting one caesarean section as two outcomes for two twins would clearly not be correct. Conversely, neonatal outcomes such as sepsis would be most appropriately analysed by the number of babies, as each baby develops the condition separately.

Babies from multiple pregnancies may be more likely to develop the same outcomes (i.e. non-independence) so counting each as a separate data point may overestimate the sample size and make confidence intervals too narrow. This can be allowed for by using cluster trial methods with each woman regarded as a randomised cluster. As for other cluster randomised trials, however, the methods require an estimate of the ICC, which may not be available from the trial report. However, it may be possible to calculate an ICC from data included in the paper (consultation with a statistician may be necessary for this), or alternatively, it may be possible to use an ICC from another trial or review that included multiple pregnancies. Adjustments for multiple pregnancies will probably only make a substantial difference to reviews’ results if multiples make up a substantial proportion of the trial population. If it is not possible to obtain enough information to make any adjustment for the effects of multiple pregnancies, the data should be analysed as if babies from multiple pregnancies are independent, using the number of infants as the denominator. This will give an unbiased result but the width of the confidence intervals will be underestimated. As long as the proportion of multiple pregnancies in the analysis is fairly low, this is unlikely to make any substantial difference to the conclusions.

Dealing with missing data

Intention-to-treat analysis (Handbook Section 16.1) and Missing outcome data (Handbook Section 16.2)

Intention-to-treat (ITT) analysis means that:

1. all participants are included in the analysis;
2. all participants are analysed in the group to which they were allocated, regardless of whether or not they received the allocated intervention.

Reviews should attempt to fulfil both criteria for all trials. Published reports of trials frequently describe analyses that satisfy only the second criterion as “intention to treat”. Descriptions in published papers should therefore not be accepted uncritically, but the review should make its own judgement based on the information available.

If some participants were not analysed in the group to which they were randomised, there may be sufficient information in the trial report to restore them to the correct group. Alternatively, the original investigators may be able to provide the necessary information. If participants cannot be analysed in their allocated groups, this should be clearly stated in the review (in the Risk of Bias table and in the Risk of Bias section of the text).

Where there are missing outcome data in studies included in a review, the primary analysis should use the number of participants with data as the denominator (i.e. "available case" analysis). The possible effects of the missing data may be explored in sensitivity analyses (for example, assuming the "worst case" scenario of all missing participants having an outcome, or other scenarios).

Assessment of reporting biases (Handbook chapter 10)

The possibility of publication bias and related biases should be investigated if any meta-analyses are performed that involve 10 or more studies. The methods that will be used for this must be described in the protocol.

A recommended procedure for assessing and investigating publication and related biases is outlined below:

1. If there are 10 or more studies in an analysis, produce a funnel plot.
2. Perform a visual assessment of funnel plot asymmetry.
3. If evidence for funnel plot asymmetry is found, conduct exploratory analyses to investigate the possible causes (*see* Handbook section 10.4.4). These may include, for example, comparison of fixed-effect and random-effects estimates, comparison of treatments effects of small and large studies, or studies at low- and high-risk of bias, or investigation of the effects of methods to correct for publication bias. Statistical support from the Pregnancy and Childbirth Group is available to assist with any such analyses.

Data synthesis

Fixed-effect and random-effects analyses

The choice of fixed-effect or random-effects analysis and the conditions under which each should be used are a source of ongoing debate and disagreement among statisticians. It is therefore difficult to make clear-cut recommendations about when to use the two methods.

Fixed-effect and random-effects analyses address slightly different questions. Fixed-effect analysis assumes that the studies are estimating the same underlying treatment effect. It is therefore reasonable to use fixed-effect analysis if the studies are similar enough in their populations, interventions and methods to make this assumption reasonable. Random-effects analysis, however, assumes that the treatment effect differs between studies, and it estimates the centre of the distribution of effect sizes (i.e. the average treatment effect) and the uncertainty around it (Handbook Section 9.5.4). The estimate of the average treatment effect produced by a random-effects analysis will not always be clinically relevant; if there is a distribution of effect sizes, the average may not represent a treatment effect that would occur in any trial or population.

The choice of primary analytical method should be considered in the protocol, and not in response to the results. In most reviews fixed-effect analysis is most appropriate and this will usually be the preferred method. Under certain conditions random-effects analysis may be preferable:

(a) it is expected that the review will combine data from trials that have differences in design, population or intervention that are likely to result in different treatment effects, <u>and</u>

(b) the results of a random-effects analysis address a clinically relevant question.

If random-effects analyses are used, the results should include the tau-squared (τ^2) statistic produced by RevMan 5 and an estimated 95% range of underlying intervention effects (prediction interval), which can be calculated from τ^2 (*see* Handbook Section 9.5.4 and Higgins et al 2009[3]). The prediction interval is important and appropriate; as random effects analysis assumes there is a range of underlying treatment effects, it is important to estimate what that range is. The estimate of the average effect from a random-effects analysis should not be interpreted as a single underlying effect size as if it were an estimate from a fixed effects analysis. Statistical support is available from the Pregnancy and Childbirth Group to assist with these calculations.

Denominators for competing outcomes

A common situation in Pregnancy and Childbirth group reviews is that outcomes do not apply to all randomised participants. For a review of an intervention given during pregnancy, neonatal morbidity outcomes will only apply to those fetuses that survive to birth. Any that miscarry or are terminated would not be able to have any neonatal morbidity. In this situation there is a temptation to use as the denominators the number that could have experienced the outcome instead of the number randomised. However, this risks introducing bias because the comparison is then not between the randomised groups. If there is a difference in survival to birth, the populations “eligible” for neonatal morbidity in the two arms may be different, and the differences between the populations, rather than different effects of the intervention, may account for any differences in outcome.

The best option for dealing with the situation is to define the outcomes so that they apply to all randomised participants. For example, an outcome could be defined as “still birth or low cord blood pH” to avoid the confusing result that a group with more still births apparently has better neonatal outcomes. If this is not possible or is undesirable, the number randomised (rather than the number eligible for the outcome) should be used as the denominator for the primary analyses. A sensitivity analysis can be performed using the numbers eligible for the outcome as the denominator, to investigate whether this analysis suggests a different result, and the risks of the outcome among the eligible population may be discussed in the text.

Continuous outcomes that apply to only a subset of the participants present a particular problem and require careful consideration. The options are to assume a value for the participants who were not “eligible” for the outcome (usually, but not necessarily, the same value would be used for all of them), or restrict the analysis only to those for whom the outcome was measured. In some circumstances reasonable assumptions can be made about the value for those who did not have the outcome measure, but often there is no obvious sensible value to use for those ineligible for the outcome; a judgement will then need to be made on the best approach for each individual outcome. The strategy that is adopted should be documented in the review.

Heterogeneity

It is important to quantify the amount of heterogeneity in all meta-analyses that are performed, assess whether it is sufficiently large to demand investigation, and if so, perform appropriate actions to investigate its causes and account for it in analyses.

Measuring heterogeneity (Handbook Section 9.5.2)

There are several statistics available in RevMan for quantifying the amount of heterogeneity in a meta-analysis. The most useful of these are I^2 and τ^2 (tau-squared). These should be reported in the results of the meta-analysis. τ^2 is available in RevMan only for random-effects analyses, so this method will need to be used in addition to fixed effect in order to obtain this statistic.

Note that I^2 is a statistic, not a “test” for heterogeneity, and it does not give an unequivocal answer as to whether or not the amount of heterogeneity found is important. In particular, because I^2 measures the proportion of variation due to heterogeneity rather than sampling error, where there is low sampling error, for example if a meta-analysis contains only several large trials, high values of I^2 can result, which may not represent heterogeneity that is a concern (*see* Rücker et al 2008[4]). Assessing heterogeneity based on I^2 alone is therefore not recommended.

It is not possible to specify a threshold value of τ^2 representing substantial heterogeneity that could be used in all reviews, as its value depends on the type of outcome measure being used; it will be different for odds ratios, risk differences, mean differences etc. However, it is advisable to prespecify a threshold value for considering heterogeneity to be substantial (*see* Rucker et al 2008 for an example). Statistical support is available to help with this.

Reviews need to specify criteria for when heterogeneity is sufficient to require further investigation. Recommended criteria are that heterogeneity is a concern when:

(a) There is a high I^2 value (exceeding 30%); and

either

(b) There is inconsistency between trials in direction or magnitude of effects (judged visually), or a low (< 0.10) p-value in the chi-squared test for heterogeneity.

or

(c) the estimate of between-study heterogeneity (τ^2) is above a prespecified threshold value.

Investigation of heterogeneity

If substantial heterogeneity (according to the criteria defined by the review) is found, its causes should always be investigated using methods that are specified in the “Subgroup analyses and investigation of heterogeneity” section of the protocol. Several methods could be used for this, including subgroup analyses, meta-regression and sensitivity analyses.

In addition, consideration should be given, ideally, to whether it is meaningful to produce an overall estimate of the treatment effect. Not presenting an overall meta-analysis is a valid option and may be the best approach in some circumstances, for example, if the results are very variable or inconsistent in direction, or there is evidence from subgroup analyses or meta-regression that the heterogeneity is explained by variation in the treatment effect between different sorts of trials or participants. If an overall summary is considered meaningful, random effects should be used to incorporate the heterogeneity, and the results compared with the fixed-effect estimate.

A suggested summary process for investigation of heterogeneity, if fixed effect is the primary analytical method, is outlined below:

1. Assess whether each analysis has substantial heterogeneity using the criteria specified in the review’s protocol;
2. If there is substantial heterogeneity, perform the specified investigations (subgroup analyses, metaregressions, sensitivity analyses) ;
3. Consider whether an overall summary is required and meaningful;
4. If it is, use random-effects analysis to produce it;
5. Report the results of this analysis (the summary estimate and 95% confidence interval, I^2 and τ^2, 95% prediction interval;
6. Compare the results of the random-effects analysis with the fixed-effect analysis.

If random effects is the primary analytical method, comparison with the fixed-effect estimate is not necessary,

Subgroup analyses (Handbook Section 9.6.3 and 9.6.3.1)

Subgroup analyses have great potential to be misleading so should be undertaken and interpreted with caution. They should be pre-specified in the protocol, and only a small number that have a clear rationale should be conducted. Only in exceptional circumstances should subgroup analyses that are not pre-specified be conducted, and these must be clearly indicated as such in the text. Only a small number of the most important outcomes (usually the primary outcomes) should be included in subgroup analyses, and they should be pre-specified in the protocol. The restriction on the number of subgroup analyses and outcomes is to guard against the possibility of spuriously significant results.

Subgroup analyses should use a statistical interaction test to investigate whether there is a difference in treatment effects between the subgroups. An interaction test is implemented in RevMan 5.1, for both fixed-effect and random-effects analysis. For other types of analysis, the best option at present is to compare the confidence intervals around the treatment effect estimate in each subgroup; if the 95% confidence intervals do not overlap, there is good evidence that there is a real difference in treatment effect between the subgroups.

Differences between subgroups must not be judged by comparing the statistical significance of the results within each subgroup; this is extremely unreliable and may lead to serious errors.

Subgroup analyses are exploratory and should not be used for the main conclusions of reviews.

Results of subgroup analyses should quote the χ^2 statistic and p-value, and the interaction test I^2 value (this measures the proportion of the variation between subgroups that is due to heterogeneity).

Meta-regression

Meta-regression is another way of investigating differences in treatment effect between subgroups and other covariates. It is not implemented in RevMan so will require use of a package such as Stata or assistance from a statistician. Meta-regression is particularly suitable for investigating the effects of continuous covariates, for example dose-response relationships.

Sensitivity analyses (Section 9.7)

Sensitivity analyses test the robustness of the review's results to decisions that were taken during its conduct. The Handbook gives some examples of the types of decisions that might require investigation in sensitivity analyses. It is not possible to give a definitive list of sensitivity analyses that should be conducted, as they will be determined by the particular circumstances of each review. Sensitivity analyses should be carried out for any aspects of the review that could potentially make the results unreliable. Some that are commonly required are:

1. omission of studies at high risk of bias, such as quasi-randomised studies;
2. repeating analyses using fixed or random effects;

3. repeating analyses using a different summary statistic e.g. odds ratio instead of risk ratio;
4. if cluster-randomised trials are included, investigating the effect of different values of the ICC;
5. omission of studies published as abstracts or non-peer reviewed publications.

Sensitivity analyses can be used as part of an investigation of heterogeneity, for example repeating analyses using random effects or a different summary statistic, to determine the effects of these choices on the measures of heterogeneity or the results of the analysis.

1. Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

2.Review Manager (RevMan) Version 5.1. for Windows [Computer program]. Copenhagen: The Nordic Cochrane Centre: The Cochrane Collaboration, 2011.

References

[1]

[2]

[3] Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society. Series A, (Statistics in Society) 2009;172:137-59.

[4] Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I^2 in assessing heterogeneity may mislead. BMC Medical Research Methodology 2008;8:79.