6 Assessment of study quality

Quality assessment of individual studies that are summarised in systematic reviews is necessary to limit bias in conducting the systematic review, gain insight into potential comparisons, and guide interpretation of findings. Factors that warrant assessment are those related to applicability of findings, validity of individual studies, and certain design characteristics that affect interpretation of results. Applicability, which is also called external validity or generalisability by some, is related to the definition of the key components of well-formulated questions outlined in section 4. Specifically, whether a review's findings are applicable to a particular population, intervention strategy or outcome is dependent upon the studies selected for review, and on how the people, interventions and outcomes of interest were defined by these studies and the authors (reviewers).

Interpretation of results is dependent upon the validity of the included studies and other characteristics. For example, a review may summarise twenty valid trials that evaluate the effects of antischismic agents on symptoms of chest pain in adults with prior myocardial infarction. However, the trials may examine different preparations and doses of antischismic agents and may have varying durations. These latter issues would affect interpretation though they may not be directly relevant to the internal validity of the trials. Examples of how to abstract data related to applicability and design factors likely to affect the interpretation are in section 7. The remainder of this section will focus on assessing the validity of individual studies included in a systematic review. As most Cochrane reviews focus on randomised trials, it concentrates on how to appraise the validity from these studies.

6.1 Validity

In the context of a systematic review, the validity of a study is the extent to which its design and conduct are likely to prevent systematic errors, or bias (Moher 1995). An important issue that should not be confused with validity is precision. Precision is a measure of the likelihood of chance effects leading to random errors. It is reflected in the confidence interval around the estimate of effect from each study and the weight given to the results of each study when an overall estimate of effect or weighted average is derived. More precise results are given more weight.

Variation in validity can explain variation in the results of the studies included in a systematic review. More rigorous studies may be more likely to yield results that are closer to the 'truth'. Quantitative analysis of results from studies of variable validity can result in 'false positive' conclusions (erroneously concluding an intervention is effective) if the less rigorous studies are biased toward overestimating an intervention's effectiveness. They might also come to 'false negative' conclusions (erroneously concluding no effect) if the less rigorous studies are biased towards underestimating an intervention's effect (Detsky 1992).

It is important to systematically complete critical appraisal of all studies in a review even if there is no variability in either the validity or results of the included studies. For instance, the results may be consistent among studies but all the studies may be flawed. In this case, the review's conclusions would not be as strong as if a series of rigorous studies yielded consistent results about an intervention's effect.
6.2 Sources of bias in trials of healthcare interventions

There are four sources of systematic bias in trials of the effects of healthcare: selection bias, performance bias, attrition bias and detection bias (see figure below). Unfortunately, we do not have strong empirical evidence of a relationship between trial outcomes and specific criteria or sets of criteria used to assess the risk of these biases (Moher 1995, Moher 1996b). There is, however, a logical basis for suspecting such relationships and good reason to consider these four potential biases when assessing studies for a review (Feinstein 1985).

![Diagram of sources of bias]

6.3 Selection bias

One of the most important factors that may lead to bias and distort treatment comparisons is that which can result from the way that comparison groups are assembled (Kunz 1998). Using an appropriate method for preventing foreknowledge of treatment assignment is crucially important in trial design. When assessing a potential participant's eligibility for a trial, those who are recruiting participants and the participants themselves should remain unaware of the next assignment in the sequence until after the decision about eligibility has been made. Then, after assignment has been revealed, they should not be able to alter the assignment or the decision about eligibility. The ideal is for the process to be impervious to any influence by the individuals making the allocation. This will be most securely achieved if an assignment schedule generated using true randomisation is administered by someone who is not responsible for recruiting subjects, such as someone based in a central trial office or pharmacy. If such central randomisation cannot be organised, then other precautions are required to prevent manipulation of the allocation process by those involved in recruitment.

The process of concealing assignment until treatment has been allocated has sometimes been referred to as 'randomisation blinding' (Chalmers 1983). This term does not clearly distinguish concealed allocation from blinding of patients, providers, outcome evaluators and analysts and is unsatisfactory for three reasons. First, the reason for concealing the assignment schedule is to eliminate selection bias. In contrast, blinding (used after the allocation of the intervention) reduces performance and detection biases. Second, from a practical standpoint, concealing allocation up to the point of assignment is always possible, regardless of the study question, but blinding after allocation may be impossible, as in trials comparing surgical with...
medical treatment. Third, control of selection bias is relevant to the trial as a whole, and thus to all outcomes being compared. In contrast, control of detection bias is often outcome-specific and may be accomplished successfully for some outcomes in a study but not others. Thus, blinding up to allocation and blinding after allocation are addressing different sources of bias, are inherently different in their practicability and may apply to different components of a study. To clearly distinguish these different forms and purposes of 'blinding', we will refer to the process of concealing assignments as allocation concealment and reserve blinding for measures taken to reduce bias after the intervention has been assigned.

Empirical research has shown that lack of adequate allocation concealment is associated with bias (Chalmers 1983, Schulz 1995, Moher 1998). Indeed, concealment has been found to be more important in preventing bias than other components of allocation, such as the generation of the allocation sequence (e.g., computer, random number table, alternation). Thus, studies can be judged on the method of allocation concealment. Information should be presented that provides some assurance that allocations were not known until, at least, the point of allocation. The method for assigning participants to interventions should be robust against patient and clinician bias and its description should be clear. The following are some approaches that can be used to ensure adequate concealment schemes.

- centralised (e.g. allocation by a central office unaware of subject characteristics)
- orphan pharmacy-controlled randomisation
- pre-numbered or coded identical containers which are administered serially to participants
- on-site computer system combined with allocations kept in a locked unreadable computer file that can be accessed only after the characteristics of an enrolled participant have been entered
- sequentially numbered, sealed, opaque envelopes

Other approaches may include approaches similar to ones listed above, along with reassurance that the person who generated the allocation scheme did not administer it. Some schemes may be innovative and not fit any of the approaches above, but still provide adequate concealment.

Approaches to allocation concealment that should be considered clearly inadequate include: alternation; the use of case record numbers, dates of birth or day of the week, and any procedure that is entirely transparent before allocation, such as an open list of random numbers. When studies do not report any concealment approach, adequacy should be considered unclear. Examples include merely stating that a list or table was used, only specifying that sealed envelopes were used and reporting an apparently adequate concealment scheme in combination with other information that leads the author to be suspicious. When authors enter studies into RevMan they are required to indicate whether allocation concealment was adequate (A), unclear (B), inadequate (C), or that allocation concealment was not used (D) as a criterion to assess validity.

### 6.4 Performance bias

Performance bias refers to systematic differences in the care provided to the participants in the comparison groups other than the intervention under investigation. To protect against unintended differences in care and placebo effects, those providing and receiving care can be 'blinded' so that they do not know the group to which the recipients of care have been allocated. Some research suggests that such blinding is important in protecting against bias (Karlowski 1975, Colditz 1989, Schulz 1995). Studies have shown that contamination (provision of the intervention to the control group) and cointervention (provision of unintended additional care to either comparison group) can affect study results (CCSG 1978,
Furthermore, there is evidence that participants who are aware of their assignment status report more symptoms, leading to biased results (Karlowski 1975). For these reasons, authors may want to consider the use of 'blinding' as a criterion for validity. This can be done with the following questions: Were the recipients of care unaware of their assigned intervention? Were those providing care unaware of the assigned intervention? A third question addressing blinding and detection bias is often added: Were persons responsible for assessing outcomes unaware of the assigned intervention? This addresses detection bias, as noted below.

Authors working on topics where blinding is likely to be important may want to develop specific criteria for judging the appropriateness of the method that was used for blinding. In some areas it may be desirable to use the same criterion across reviews, in which case a Collaborative Review Group (CRG) might want to agree to a standard approach for assessing blinding (Chalmers 1989, Schulz 1995, Jadad 1996, Moher 1996b).

6.5 Attrition bias
Attrition bias refers to systematic differences between the comparison groups in the loss of participants from the study. It has been called exclusion bias. It is called attrition bias here to prevent confusion with pre-allocation exclusion and inclusion criteria for enrolling participants. Because of inadequacies in reporting how losses of participants (e.g. withdrawals, dropouts, protocol deviations) are handled, authors should be cautious about implicit accounts of follow-up. The approach to handling losses has great potential for biasing the results and reporting inadequacies cloud this problem. What is reported, or more frequently implied, in study reports on attrition after allocation has not been found to be consistently related to bias (Schulz 1995). Thus authors should be cautious about using reported follow-up as a validity criterion, particularly when it is implied rather than explicitly reported. This is a general recommendation, however, and may not apply to certain topic areas that have higher quality reporting or where it is possible to obtain missing information from investigators.

6.6 Detection bias
Detection bias refers to systematic differences between the comparison groups in outcome assessment. Trials that blind the people who will assess outcomes to the intervention allocation should logically be less likely to be biased than trials that do not. Blinding is likely to be particularly important in research with subjective outcome measures such as pain (Karlowski 1975, Colditz 1989, Schulz 1995). However, at least two empirical studies have failed to demonstrate a relationship between blinding of outcome assessment and study results. This may be due to inadequacies in the reporting of studies (Reitman 1988).

Bias due to the selective reporting of results is somewhat different from bias in outcome assessment. This source of bias may be important in areas where multiple outcome measures are used, such as evaluations of treatments for rheumatoid arthritis (Gotzsche 1989). Therefore, authors may want to consider specification of predefined primary outcomes and analyses by the investigators as indicators of validity. Alternatively, selective reporting of particular outcomes could be taken to suggest the need for better reporting and efforts by authors to obtain missing data.
6.7 Approaches to summarising the validity of studies

6.7.1 Simple approaches

There are several ways to rate validity. One is to rate individual criteria as 'met', 'unmet', or 'unclear' and to use individual criteria, such as adequacy of allocation concealment, in sensitivity analyses (see section 8.10). However, if several explicit criteria are used to assess validity, it is desirable to summarise these so as to derive an overall assessment of how valid the results of each study are. A simple approach to doing this is to use three categories such as the following:

<table>
<thead>
<tr>
<th>Risk of bias</th>
<th>Interpretation</th>
<th>Relationship to individual criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Low risk of bias</td>
<td>Plausible bias unlikely to seriously alter the results</td>
<td>All of the criteria met</td>
</tr>
<tr>
<td>B. Moderate risk of bias</td>
<td>Plausible bias that raises some doubt about the results</td>
<td>One or more criteria partly met</td>
</tr>
<tr>
<td>C. High risk of bias</td>
<td>Plausible bias that seriously weakens confidence in the results</td>
<td>One or more criteria not met</td>
</tr>
</tbody>
</table>

The relationships suggested above will most likely be appropriate if only a few assessment criteria are used and if all the criteria address only substantive, important threats to the validity of study results. In general and when possible, authors should obtain further information from the authors of a report when it is unclear whether a criterion was met.

6.7.2 'Quality' scales and checklists

David Moher and his colleagues identified 25 scales and 9 checklists that have been used to assess the validity and 'quality' of randomised controlled trials (Moher 1995, Moher 1996b). These scales and checklists include anywhere from 3 to 57 items and take from 10 to 45 minutes to complete. Almost all of the items in the instruments are based on suggested or 'generally accepted' criteria that are mentioned in clinical trial textbooks. Many of the instruments are liable to confuse the quality of reporting with the validity of the design and conduct of a trial. Moreover, scoring is based on whether something was reported (such as how participants were allocated) rather than whether it was done appropriately in the study. Many also contain items that are not directly related to validity, such as whether a power calculation was done (an item that relates more to the precision of the results) or whether the inclusion and exclusion criteria were clearly described (an item that relates more to applicability than validity).

Because there is no 'gold standard' for the 'true' validity of a trial, the possibility of validating any proposed scoring system is limited. While it is possible to apply basic principles of measurement to the development of a scale for assessing the validity of randomised trials, the relationship between such a score and the degree to which a study is free from bias is not obvious. None of the currently available scales for measuring the validity or 'quality' of trials can be recommended without reservation. If authors or CRGs choose to use such a scale, it must be with caution.

Most of the available scales for assessing the validity of randomised controlled trials derive a summary score by adding the scores (with or without differential weights) for each item.
While this approach offers appealing simplicity, it is not supported by empirical evidence (Emerson 1990, Schulz 1995). Notably, scales with multiple items and complex scoring systems take more time to complete than simple approaches. They have not been shown to provide more reliable assessments of validity (Jüni 1999). They may carry a greater risk of confusing the quality of reporting with the validity of the study. They are more likely to include criteria that do not directly measure internal validity, and they are less likely to be transparent to users of the review. For these reasons, it is preferable to use simple approaches for assessing validity that can be fully reported (i.e. how each trial scored on each criterion).

6.8 Bias in non-experimental studies

The Non-randomised Studies Methods Group are preparing guidance on the use of non-randomised studies in Cochrane reviews (Appendix 6a). In the meantime, this section describes some issues that should be considered in assessing the validity of non-randomised studies. The logical reason for focusing on randomised controlled trials in Cochrane reviews is that randomisation is the only means of allocation that controls for unknown and unmeasured confounders as well as those that are known and measured. Differences between comparison groups in prognosis, responsiveness to treatment or exposure to other factors that affect outcomes can distort the apparent magnitude of effects of the intervention of interest. It is possible to control or adjust for confounders that are known and measured in observational studies, such as case-control and cohort studies. However, it is not possible to adjust for those factors that are not known to be confounders or that were not measured. Unfortunately it can rarely, if ever, be assumed that all important factors relevant to prognosis and responsiveness to treatment are known, and for those that are known difficulties can arise in measuring and accounting for them in analyses. Empirical evidence supports these logical concerns (Kunz 1995). Selection bias can distort effects in either direction, causing them to appear either larger or smaller than they are. It is generally not possible to predict the magnitude, and often not even the direction of this bias in specific studies. However, on average, selection bias tends to make treatment effects appear larger than they are and the size of these distortions can be as large or larger than the size of the effects that are being measured (Kunz 1995).

Despite these concerns, there is sometimes good reason to rely on observational studies for information about the effects of healthcare interventions, and to include such studies in Cochrane reviews. For example, well designed observational studies have provided useful data regarding the effects of interventions such as mandatory use of helmets by motorcyclists, screening for cervical cancer, dissemination of clinical practice guidelines to change professional practice and rare adverse effects of medication.

Various criteria have been suggested to critically appraise the validity of observational studies (Horwitz 1979, Feinstein 1982, Levine 1994, Bero 1999). In general, the same four sources of bias noted above can be applied to other types of comparative studies, as illustrated below:

<table>
<thead>
<tr>
<th>Source of bias</th>
<th>Cohort studies</th>
<th>Case-control studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection bias</td>
<td>Control for confounders</td>
<td>Matching</td>
</tr>
<tr>
<td>Performance bias</td>
<td>Measurement of exposure</td>
<td>Measurement of exposure</td>
</tr>
<tr>
<td>Attrition bias</td>
<td>Completeness of follow-up</td>
<td>Completeness of follow-up</td>
</tr>
<tr>
<td>Detection bias</td>
<td>Blinding</td>
<td>Case definition</td>
</tr>
</tbody>
</table>

Concerns about attrition bias are similar in randomised trials, cohort studies and case-control studies and relate to the extent that all participants in a study are appropriately accounted for in its results. Concerns about detection bias are also similar for cohort studies, and are related to the case definition that is used in case-control studies (since people are entered into such studies based on knowledge of the outcome of interest). The major difference between randomised trials and observational studies has to do with selection bias and the need to identify and account for potential confounders in observational studies. To do this authors
must make judgements about what confounders are important and the extent to which these were appropriately measured and controlled for. Assessing 'performance bias' is also more difficult in observational studies since it is necessary to measure exposure to the intervention of interest and ensure that there were not differences in the exposure of the comparison groups to other factors that could affect outcomes. In addition to considerations of blinding, which are similar to those in randomised trials, it is important to consider whether exposure was measured in a similar and unbiased way in the groups being compared. So, for example, in addition to concerns about bias due to confounders in cohort and case control studies of the effects of post-menopausal hormone replacement therapy, investigators and authors must ensure that use of hormones was measured in an unbiased way.

In summary, a great deal of judgement is necessary in assessing the validity of observational studies. Judgement is also needed when the validity of randomised trials is assessed, but the nature of observational studies makes them even more difficult to critically appraise. This requires a thorough understanding of both the problem that is the focus of the review and methodological considerations. Caution is needed.

### 6.9 Application of quality assessment criteria

Several basic decisions must be made regarding the assessment of studies, similar to those made regarding the process of selecting studies (section 5.2.3). A prime consideration is the number of authors. Should there be one or more than one? How many are necessary and how many are too many? Will authors review the same articles to maximise reliability or mutually exclusive sets of reports to minimise workload? A concomitant consideration is the backgrounds of the different authors and whether previous training and experience in study design or critical appraisal will be required.

Conducting systematic reviews with multiple authors is a two-sided coin. On the one hand it may limit bias, minimise errors and improve reliability of findings, but having more than one creates the potential for disagreement among authors. When multiple authors will be involved, there should be an explicit procedure or decision rule identified \textit{a priori} for identifying and resolving disagreement. As a general rule, we recommend that at least two authors assess information that involves subjective interpretation and information that is critical to the interpretation of results (e.g., outcome data). Section 7 describes methods for reaching and monitoring consensus when more than one author is used.

Regardless of the number of authors, it is important to first test any assessment criteria that are planned using a pilot sample of articles to ensure that the appraisal criteria can be applied consistently. Three to six papers that span a range of low to high risk bias might provide a suitable sample for this.

Should authors be especially trained in research methods, the content area of a review or both? Although experts in content areas may have pre-formed opinions that can bias their assessments (Oxman 1993b), they may nonetheless give more consistent assessments of the validity of studies than persons without content expertise (Jadad 1996). They may also have valuable insights that are different than those that someone with methodological expertise alone would have. It would seem intuitively desirable to use both content experts and non-experts and to ensure that both have an adequate understanding of the relevant methodological issues.

Authors must also decide whether those assessing study validity will be blinded to the names of the authors, institutions, journal and results of a study when they assess its methods. Some empirical evidence suggests that blind assessment of reports might produce lower and more consistent scores than open assessments (Jadad 1996). Other empirical evidence suggests little benefits from blind assessments (Berlin 1997). However, blinded assessments are very time consuming. Authors must weigh their potential benefits against the costs involved when
deciding whether or not to blind the authors. Further research is underway comparing blind and open assessments of study validity and these results may help guide this decision.

6.10 Incorporating assessments of study validity in reviews

There are several ways in which validity assessments can be used in a review:

- as a threshold for inclusion of studies
- as a possible explanation for differences in results between studies
- in sensitivity analyses
- as weights in statistical analysis of the study results

Failure to meet one or more validity criteria may indicate such a high risk of bias in some reviews that it constitutes grounds for exclusion of those studies. For example, for highly subjective outcomes such as pain, authors may decide to include only studies that prevent 'performance bias' by blinding participants. The decision about where to set the cut point for inclusion can be conceptualised as existing on a continuum between 'free from bias' and 'undoubtedly biased' as illustrated below:

If authors raise the methodological cut-point for including studies, there will be less variation in validity among the included reports. Assessments of validity would then categorise studies by the risk of bias within the range above the inclusion cut-point. With a sufficiently high cut-point, any variation in validity among included studies may be too small to be important.

There are several methods to examine whether validity may explain differences among study results (Detsky 1992). Visual plots of the results arranged in order of their validity can be used. A second approach is to analyse subgroups of studies above a methodological cut-point, which should, preferably, be specified a priori, in the protocol of the review. This approach can be used whether or not the study results are heterogeneous, by doing a sensitivity analysis to determine if the overall results are the same when only studies with little risk of bias are included in the analysis. A third approach is to combine the results of each study sequentially in order of their assessed validity ('cumulative meta-analysis'), examining the impact on the overall results as trials of decreasing validity are included (see section 8.11.6).
A fourth approach is to use statistical methods to weight studies according to their assessed validity or to use 'meta-regression' to explore the relationship between validity and the magnitude of effect across studies (see section 8.8.1). Statistical methods for combining the results of studies generally weight the influence of each study by the inverse of the variance for the estimated measure of effect. In other words, studies with more precise results (narrower confidence intervals) are given more weight. It is also possible to weight studies according to validity so that more valid studies have more influence on the summary result. The main objection to this approach is that there is no empirical basis for determining how much weight to assign to different validity criteria or for quantitatively relating differences on 'quality' scales to differences in the risk of bias between studies.

It is possible using RevMan 4.0 to order studies according to either adequacy of concealment of allocation or 'user defined' assessments of validity. Subgroup analyses based on assessments of validity can be done, although a test of statistical significance of differences between subgroups of studies has not been implemented. RevMan does not include an option for weighting studies by methodological validity and meta-regression is not possible using RevMan 4.0.

6.11 Limitations of quality assessment

There are two major difficulties with assessing the validity of studies. The first is inadequate reporting of trials (SORT 1994, Schulz 1994, WGRR 1994, Begg 1996). It is possible to assume if something was not reported it was not done. However, this is not necessarily correct. Authors should attempt to obtain additional data from investigators as necessary, but this may be difficult. The application of standards for reporting trials (SORT 1994, WGRR 1994, Begg 1996) should facilitate the assessment of study validity in the future.

The second limitation, which in part is a consequence of the first, is limited empirical evidence of a relationship between parameters thought to measure validity and actual study outcomes. As noted above, there is empirical evidence suggesting that, on average, both inadequate concealment of allocation and lack of double blinding result in over-estimates of the effects of treatment. Clearly much more research needs to be done to establish which criteria for assessing validity are indeed important determinants of study results and when. Improved reporting of methods will facilitate such research. Meanwhile, authors should avoid the use of 'quality scores' and undue reliance on detailed quality assessments. It is not supported by empirical evidence, it can be time-consuming and it is potentially misleading.

6.12 References


