

# Chapter 12: Interpreting results and drawing conclusions

---

Authors: Holger J Schünemann, Andrew D Oxman, Gunn E Vist, Julian PT Higgins, Jonathan J Deeks, Paul Glasziou and Gordon H Guyatt on behalf of the Cochrane Applicability and Recommendations Methods Group.

Copyright © 2008 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd under “The Cochrane Book Series” Imprint.

This extract is made available solely for use in the authoring, editing or refereeing of Cochrane reviews, or for training in these processes by representatives of formal entities of The Cochrane Collaboration. Other than for the purposes just stated, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the copyright holders.

Permission to translate part or all of this document must be obtained from the publishers.

This extract is from *Handbook* version 5.0.1. For guidance on how to cite it, see Section 12.8. The material is also published in Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on [www.wiley.com](http://www.wiley.com).

## Key points

- The GRADE approach, adopted by The Cochrane Collaboration, specifies four levels of quality (high, moderate, low and very low) where the highest quality rating is for a body of evidence based on randomized trials. Review authors can downgrade randomized trial evidence depending on the presence of five factors and upgrade the quality of evidence of observational studies depending on three factors.
- Quality ratings are made separately for each outcome.
- Methods for computing, presenting and interpreting relative and absolute effects for dichotomous outcome data, including the number needed to treat (NNT), are described in this chapter.
- For continuous outcome measures, review authors can present pooled results for studies using the same units, the standardized mean difference and effect sizes when studies use the same construct but different scales, and odds ratios after transformation of the standardized mean differences.
- Review authors should not describe results as ‘not statistically significant’ or ‘non-significant’, but report the confidence interval together with the exact P value.
- Review authors should not make recommendations, but they can – after describing the quality of evidence and the balance of benefits and harms – highlight different actions that might be consistent with particular patterns of values and preferences.

## 12.1 Introduction

The purpose of Cochrane reviews is to facilitate healthcare decision-making by patients and the general public, clinicians, administrators, and policy makers. A clear statement of findings, a considered discussion and a clear presentation of the authors' conclusions are important parts of the review. In particular, the following issues can help people make better informed decisions and increase the usability of Cochrane reviews:

- information on all important outcomes, including adverse outcomes;
- the quality of the evidence for each of these outcomes, as it applies to specific populations, and specific interventions; and
- clarification of the manner in which particular values and preferences may bear on the balance of benefits, harms, burden and costs of the intervention.

A 'Summary of findings' table, described in Chapter 11 (Section 11.5), provides key pieces of information in a quick and accessible format. Review authors are encouraged to include such tables in Cochrane reviews, and to ensure that there is sufficient description of the studies and meta-analyses to support their contents. The Discussion section of the text should provide complementary considerations. Authors should use five subheadings to ensure they cover suitable material in the Discussion section and that they place the review in an appropriate context. These are 'Summary of main results (benefits and harms)'; 'Overall completeness and applicability of evidence'; 'Quality of the evidence'; 'Potential biases in the review process'; and 'Agreements and disagreements with other studies or reviews'. Authors' conclusions are divided into 'Implications for practice' and 'Implications for research'.

Because Cochrane reviews have an international audience, the discussion and authors' conclusions should, so far as possible, assume a broad international perspective and provide guidance for how the results could be applied in different settings, rather than being restricted to specific national or local circumstances. Cultural differences and economic differences may both play an important role in determining the best course of action. Furthermore, individuals within societies have widely varying values and preferences regarding health states, and use of societal resources to achieve particular health states. Even in the face of the same values and preferences, people may interpret the same research evidence differently. For all these reasons, different people will often make different decisions based on the same evidence.

Thus, the purpose of the review should be to present information and aid interpretation rather than to offer recommendations. The discussion and conclusions should help people understand the implications of the evidence in relation to practical decisions and apply the results to their specific situation. Authors should avoid specific recommendations that depend on assumptions about available resources and values. Authors can, however, aid decision-making by laying out different scenarios that describe certain value structures.

In this chapter we address first one of the key aspects of interpreting findings that is also fundamental in completing a 'Summary of findings' table: the quality of evidence related to each of the outcomes. We then provide a more detailed consideration of issues around applicability and around interpretation of numerical results, and provide suggestions for presenting authors' conclusions.

## 12.2 Assessing the quality of a body of evidence

### 12.2.1 The GRADE approach

The Grades of Recommendation, Assessment, Development and Evaluation Working Group (GRADE Working Group) has developed a system for grading the quality of evidence (GRADE Working Group 2004, Schünemann 2006b, Guyatt 2008a, Guyatt 2008b). Over 20 organizations including the World Health Organization (WHO), the American College of Physicians, the American College of Chest Physicians (ACCP), the American Endocrine Society, the American Thoracic Society (ATS), the Canadian Agency for Drugs and Technology in Health (CADTH), BMJ Clinical Evidence, the National Institutes of Health and Clinical Excellence (NICE) in the UK, and UpToDate® have adopted the GRADE system in its original format or with minor modifications (Schünemann 2006b, Guyatt 2006a, Guyatt 2006b). The BMJ encourages authors of clinical guidelines to use the GRADE system ([www.bmj.com/advice/sections.shtml](http://www.bmj.com/advice/sections.shtml)). The Cochrane Collaboration has adopted the principles of the GRADE system for evaluating the quality of evidence for outcomes reported in systematic reviews. This assessment is being phased in together with the introduction of the ‘Summary of findings’ table (see Chapter 11, Section 11.5).

For purposes of systematic reviews, the GRADE approach defines the quality of a body of evidence as the extent to which one can be confident that an estimate of effect or association is close to the quantity of specific interest. Quality of a body of evidence involves consideration of within-study risk of bias (methodological quality), directness of evidence, heterogeneity, precision of effect estimates and risk of publication bias, as described in Section 12.2.2. The GRADE system entails an assessment of the quality of a body of evidence for each individual outcome.

The GRADE approach specifies four levels of quality (Table 12.2.a). The highest quality rating is for randomized trial evidence. Review authors can, however, downgrade randomized trial evidence to moderate, low, or even very low quality evidence, depending on the presence of the five factors in Table 12.2.b. Usually, quality rating will fall by one level for each factor, up to a maximum of three levels for all factors. If there are very severe problems for any one factor (e.g. when assessing limitations in design and implementation, all studies were unconcealed, unblinded, and lost over 50% of their patients to follow-up), randomized trial evidence may fall by two levels due to that factor alone.

Review authors will generally grade evidence from sound observational studies as low quality. If, however, such studies yield large effects and there is no obvious bias explaining those effects, review authors may rate the evidence as moderate or – if the effect is large enough – even high quality (Table 12.2.c). The very low quality level includes, but is not limited to, studies with critical problems and unsystematic clinical observations (e.g. case series or case reports).

**Table 12.2.a: Levels of quality of a body of evidence in the GRADE approach**

| <b>Underlying methodology</b>  | <b>Quality rating</b> |
|--|-----------------------|
| Randomized trials; or double-upgraded observational studies.   | High                  |
| Downgraded randomized trials; or upgraded observational studies.                                       | Moderate              |
| Double-downgraded randomized trials; or observational studies.   | Low                   |
| Triple-downgraded randomized trials; or downgraded observational studies; or case series/case reports. | Very low              |

**Table 12.2.b: Factors that may decrease the quality level of a body of evidence**

1. Limitations in the design and implementation of available studies suggesting high likelihood of bias.
2. Indirectness of evidence (indirect population, intervention, control, outcomes).
3. Unexplained heterogeneity or inconsistency of results (including problems with subgroup analyses).
4. Imprecision of results (wide confidence intervals).
5. High probability of publication bias.

**Table 12.2.c: Factors that may increase the quality level of a body of evidence**

1. Large magnitude of effect.
2. All plausible confounding would reduce a demonstrated effect or suggest a spurious effect when results show no effect.
3. Dose-response gradient.

## 12.2.2 Factors that decrease the quality level of a body of evidence

We now describe in more detail the five reasons for downgrading the quality of a body of evidence for a specific outcome (Table 12.2.b). In each case, if a reason is found for downgrading the evidence, it should be classified as ‘serious’ (downgrading the quality rating by one level) or ‘very serious’ (downgrading the quality grade by two levels).

1. **Limitations in the design and implementation:** Our confidence in an estimate of effect decreases if studies suffer from major limitations that are likely to result in a biased assessment of the intervention effect. For randomized trials, these methodological limitations include lack of allocation concealment, lack of blinding (particularly with subjective outcomes highly susceptible to biased assessment), a large loss to follow-up, randomized trials stopped early for benefit or selective reporting of outcomes. Chapter 8 provides a detailed discussion of study-level assessments of risk of bias in the context of a Cochrane review, and proposes an approach to assessing the risk of bias for an outcome across studies as ‘low risk of bias’, ‘unclear risk of bias’ and ‘high risk of bias’ (Chapter 8, Section 8.7). These assessments should feed directly into this factor. In particular, ‘low risk of bias’ would indicate ‘no limitation’; ‘unclear risk of bias’ would indicate either ‘no limitation’ or ‘serious limitation’; and ‘high risk of bias’ would indicate either ‘serious limitation’ or ‘very serious limitation’. Authors must use their judgement to decide between alternative categories, depending on the likely magnitude of the potential biases.

Every study addressing a particular outcome will differ, to some degree, in the risk of bias. Review authors must make an overall judgement on whether the quality of evidence for an outcome warrants downgrading on the basis of study limitations. The assessment of study limitations should apply to the studies contributing to the results in the ‘Summary of findings’ table, rather than to all studies that could potentially be included in the analysis. We have argued in Chapter 8 (Section 8.8.3) that the primary analysis should be restricted to studies at low (or low and unclear) risk of bias.

Table 12.2.d presents the judgements that must be made in going from assessments of the risk of bias to judgements about study limitations for each outcome included in a ‘Summary of findings’

table. A rating of high quality evidence can be achieved only when most evidence comes from studies that met the criteria for low risk of bias. For example, of the 22 trials addressing the impact of beta blockers on mortality in patients with heart failure, most probably or certainly used concealed allocation, all blinded at least some key groups and follow-up of randomized patients was almost complete (Brophy 2001). The quality of evidence might be downgraded by one level when most of the evidence comes from individual studies either with a crucial limitation for one criterion, or with some limitations for multiple criteria. For example, we cannot be confident that, in patients with falciparum malaria, amodiaquine and sulfadoxine-pyrimethamine together reduce treatment failures compared with sulfadoxine-pyrimethamine, because the apparent advantage of sulfadoxine-pyrimethamine was sensitive to assumptions regarding the event rate in those lost to follow-up (>20% loss to follow-up in two of three studies) (McIntosh 2005). An example of very serious limitations, warranting downgrading by two levels, is provided by evidence on surgery versus conservative treatment in the management of patients with lumbar disc prolapse (Gibson 2007). We are uncertain of the benefit of surgery in reducing symptoms after one year or longer, because the one trial included in the analysis had inadequate concealment of allocation and the outcome was assessed using a crude rating by the surgeon without blinding.

2. **Indirectness of evidence.** Two types of indirectness are relevant. First, a review comparing the effectiveness of alternative interventions (say A and B) may find that randomized trials are available, but they have compared A with placebo and B with placebo. Thus, the evidence is restricted to indirect comparisons between A and B. Second, a review may find randomized trials that meet eligibility criteria but which address a restricted version of the main review question in terms of population, intervention, comparator or outcomes. For example, suppose that in a review addressing an intervention for secondary prevention of coronary heart disease, the majority of identified studies happened to be in people who also had diabetes. Then the evidence may be regarded as indirect in relation to the broader question of interest because the population is restricted to people with diabetes. The opposite scenario can equally apply: a review addressing the effect of a preventative strategy for coronary heart disease in people with diabetes may consider trials in people without diabetes to provide relevant, albeit indirect, evidence. This would be particularly likely if investigators had conducted few if any randomized trials in the target population (e.g. people with diabetes). Other sources of indirectness may arise from interventions studied (e.g. if in all included studies a technical intervention was implemented by expert, highly trained specialists in specialist centres, then evidence on the effects of the intervention outside these centres may be indirect), comparators used (e.g. if the control groups received an intervention that is less effective than standard treatment in most settings) and outcomes assessed (e.g. indirectness due to surrogate outcomes when data on patient-important outcomes are not available, or when investigators sought data on quality of life but only symptoms were reported). Review authors should make judgements transparent when they believe downgrading is justified based on differences in anticipated effects in the group of primary interest.
3. **Unexplained heterogeneity or inconsistency of results:** When studies yield widely differing estimates of effect (heterogeneity or variability in results), investigators should look for robust explanations for that heterogeneity. For instance, drugs may have larger relative effects in sicker populations or when given in larger doses. A detailed discussion of heterogeneity and its investigation is provided in Chapter 9 (Sections 9.5 and 9.6). If an important modifier exists, with strong evidence that important outcomes are different in different subgroups (which would ideally be pre-specified), then a separate 'Summary of findings' table may be considered for a separate population. For instance, a separate 'Summary of findings' table would be used for carotid endarterectomy in symptomatic patients with high grade stenosis in which the intervention is, in the hands of the right surgeons, beneficial (Cina 2000), and another (if they considered it worth it) for asymptomatic patients with moderate grade stenosis in which surgery is not beneficial (Chambers 2005). When heterogeneity exists and affects the interpretation of results, but authors fail to identify a plausible explanation, the quality of evidence decreases.
4. **Imprecision of results:** When studies include few participants and few events and thus have wide confidence intervals, authors can lower their rating of the quality of the evidence. The confidence

intervals included in the ‘Summary of findings’ table will provide readers with information that allows them to make, to some extent, their own rating of precision.

5. **High probability of publication bias:** The quality of evidence level may be downgraded if investigators fail to report studies (typically those that show no effect: publication bias) or outcomes (typically those that may be harmful or for which no effect was observed: selective outcome reporting bias) on the basis of results. Selective reporting of outcomes is assessed at the study level as part of the assessment of risk of bias (see Chapter 8, Section 8.13), so for the studies contributing to the outcome in the ‘Summary of findings’ table this is addressed by factor 1 above (limitations in the design and implementation). If a large number of studies included in the review do not contribute to an outcome, or if there is evidence of publication bias, the quality of the evidence may be downgraded. Chapter 10 provides a detailed discussion of reporting biases, including publication bias, and how it may be tackled in a Cochrane review. A prototypical situation that may elicit suspicion of publication bias is when published evidence includes a number of small trials, all of which are industry funded (Bhandari 2004). For example, 14 trials of flavanoids in patients with haemorrhoids have shown apparent large benefits, but enrolled a total of only 1432 patients (that is, each trial enrolled relatively few patients) (Alonso-Coello 2006). The heavy involvement of sponsors in most of these trials raises questions of whether unpublished trials suggesting no benefit exist.

A particular body of evidence can suffer from problems associated with more than one of the five factors above, and the greater the problems, the lower the quality of evidence rating that should result. One could imagine a situation in which randomized trials were available, but all or virtually all of these limitations would be present, and in serious form. A very low quality of evidence rating would result.

**Table 12.2.d: Further guidelines for factor 1 (of 5) in a GRADE assessment: Going from assessments of risk of bias to judgements about study limitations for main outcomes**

| <b>Risk of bias</b>   | <b>Across studies</b>  | <b>Interpretation</b>  | <b>Considerations</b>  | <b>GRADE assessment of study limitations</b> |
|-----------------------|--|--|--|--|
| Low risk of bias.     | Most information is from studies at low risk of bias.  | Plausible bias unlikely to seriously alter the results.          | No apparent limitations.   | No serious limitations, do not downgrade.    |
| Unclear risk of bias. | Most information is from studies at low or unclear risk of bias.   | Plausible bias that raises some doubt about the results.         | Potential limitations are unlikely to lower confidence in the estimate of effect.  | No serious limitations, do not downgrade.    |
|                       |  |  | Potential limitations are likely to lower confidence in the estimate of effect.  | Serious limitations, downgrade one level.    |
| High risk of bias.    | The proportion of information from studies at high risk of bias is sufficient to affect the interpretation | Plausible bias that seriously weakens confidence in the results. | Crucial limitation for one criterion, or some limitations for multiple criteria, sufficient to lower confidence in the estimate of effect. | Serious limitations, downgrade one level.    |

|  |             |  |   |   |
|--|-------------|--|---|---|
|  | of results. |  | Crucial limitation for one or more criteria sufficient to substantially lower confidence in the estimate of effect. | Very serious limitations, downgrade two levels. |
|--|-------------|--|---|---|

### 12.2.3 Factors that increase the quality level of a body of evidence

Although observational studies and downgraded randomized trials will generally yield a low rating for quality of evidence, there will be unusual circumstances in which authors could ‘upgrade’ such evidence to moderate or even high quality (Table 12.2.c).

1. On rare occasions when methodologically well-done observational studies yield large, consistent and precise estimates of the magnitude of an intervention effect, one may be particularly confident in the results. A large effect (e.g.  $RR > 2$  or  $RR < 0.5$ ) in the absence of plausible confounders, or a very large effect (e.g.  $RR > 5$  or  $RR < 0.2$ ) in studies with no major threats to validity, might qualify for this. In these situations, while the observational studies are likely to have provided an overestimate of the true effect, the weak study design may not explain all of the apparent observed benefit. Thus, despite reservations based on the observational study design, authors are confident that the effect exists. The magnitude of the effect in these studies may move the assigned quality of evidence from low to moderate (if the effect is large in the absence of other methodological limitations). For example, a meta-analysis of observational studies showed that bicycle helmets reduce the risk of head injuries in cyclists by a large margin (odds ratio [OR] 0.31, 95%CI 0.26 to 0.37) (Thompson 2000). This large effect, in the absence of obvious bias that could create the association, suggests a rating of moderate-quality evidence.
2. On occasion, all plausible biases from observational or randomized studies may be working to underestimate an apparent intervention effect. For example, if only sicker patients receive an experimental intervention or exposure, yet they still fare better, it is likely that the actual intervention or exposure effect is larger than the data suggest. For instance, a rigorous systematic review of observational studies including a total of 38 million patients demonstrated higher death rates in private for-profit versus private not-for-profit hospitals (Devereaux 2004). One possible bias relates to different disease severity in patients in the two hospital types. It is likely, however, that patients in the not-for-profit hospitals were sicker than those in the for-profit hospitals. Thus, to the extent that residual confounding existed, it would bias results against the not-for-profit hospitals. The second likely bias was the possibility that higher numbers of patients with excellent private insurance coverage could lead to a hospital having more resources and a spill-over effect that would benefit those without such coverage. Since for-profit hospitals are likely to admit a larger proportion of such well-insured patients than not-for-profit hospitals, the bias is once again against the not-for-profit hospitals. Because the plausible biases would all diminish the demonstrated intervention effect, one might consider the evidence from these observational studies as moderate rather than low quality. A parallel situation exists when observational studies have failed to demonstrate an association but all plausible biases would have increased an intervention effect. This situation will usually arise in the exploration of apparent harmful effects. For example, because the hypoglycaemic drug phenformin causes lactic acidosis, the related agent metformin is under suspicion for the same toxicity. Nevertheless, very large observational studies have failed to demonstrate an association (Salpeter 2007). Given the likelihood that clinicians would be more alert to lactic acidosis in the presence of the agent and over-report its occurrence, one might consider this moderate, or even high quality, evidence refuting a causal relationship between typical therapeutic doses of metformin and lactic acidosis.
3. The presence of a dose-response gradient may also increase our confidence in the findings of observational studies and thereby enhance the assigned quality of evidence. For example, our confidence in the result of observational studies that show an increased risk of bleeding in patients who have supratherapeutic anticoagulation levels is increased by the observation that there is a

dose-response gradient between higher levels of the international normalized ratio (INR) and the increased risk of bleeding (Levine 2004).

## 12.3 Issues in applicability

### 12.3.1 The role of the review author

“A leap of faith is always required when applying any study findings to the population at large” or to a specific person. “In making that jump, one must always strike a balance between making justifiable broad generalizations and being too conservative in one’s conclusions” (Friedman 1985).

To address adequately the extent to which a review is relevant for the purpose to which it is being put (‘directness’), there are certain things the review author must do, and certain things the user of the review must do. We discuss here what the review author can do to help the user. Cochrane review authors must be extremely clear on the population, intervention, and outcomes that they are intending to address. Chapter 11 (Section 11.5.2) emphasizes a crucial step that has not traditionally been part of Cochrane reviews: the specification of all patient-important outcomes relevant to the intervention strategies under comparison.

With respect to participant and intervention factors, review authors need to make *a priori* hypotheses about possible effect modifiers, and then examine those hypotheses. If they find apparent subgroup effects, they must ultimately decide whether or not these effects are credible (Oxman 2002). Differences between subgroups, particularly those that correspond to differences between studies, need to be interpreted cautiously. Some chance variation between subgroups is inevitable, so unless there is strong evidence of an interaction authors should not assume that the subgroup effect exists. If, despite due caution, review authors judge subgroup effects as credible, they should conduct separate meta-analyses for the relevant subgroups, and produce separate ‘Summary of findings’ tables for those subgroups.

The user of the review will be challenged with ‘individualization’ of the findings. For example, even if relative effects are similar across subgroups, absolute effects will differ according to baseline risk. Review authors can help provide this information by identifying identifiable groups of people with varying risks in the ‘Summary of findings’ tables, as discussed in Chapter 11 (Section 11.5.5). Users can then identify the patients before them as belonging to a particular risk group, and assess their likely magnitude of benefit or harm accordingly.

Another decision users must make is whether the patients before them are so different from those included in the studies that they cannot use the results of the systematic review and meta-analysis at all. Review authors can point out that, rather than rigidly applying the inclusion and exclusion criteria of studies, it is better to ask whether there are compelling reasons why the evidence should not be applied to a particular patient (Guyatt 1994). Authors can sometimes help clinical decision makers by identifying important variation where divergence might limit the applicability of results (Schünemann 2006a), including: biologic and cultural variation, and variation in adherence to an intervention.

In addressing these issues, authors cannot be aware of, or address, the myriad of differences in circumstances around the world. They can, however, address differences of known importance to many people and, importantly, they should avoid assuming that other people’s circumstances are the same as their own in discussing the results and drawing conclusions.

### **12.3.2 Biologic variation**

Issues of biologic variation that authors should consider include divergence in pathophysiology (e.g. biologic differences between women and men that are likely to affect responsiveness to a treatment) and divergence in a causative agent (e.g. for infectious diseases such as malaria).

### **12.3.3 Variation in context and culture**

Some interventions, particularly non-pharmacological interventions, may work in some contexts but not in others; the situation has been described as program by context interaction (Hawe 2004). Context factors might pertain to the host organization in which an intervention is offered, such as the expertise, experience and morale of the staff expected to carry out the intervention, the competing priorities for the staff's attention, the local resources such as service and facilities made available to the program and the status or importance given to the program by the host organization. Broader context issues might include aspects of the system within which the host organization operates, such as the fee or payment structure for healthcare providers. Context factors may also pertain to the characteristics of the target group or population services (such aspects include the cultural and linguistic diversity, socioeconomic position, rural/urban setting), which may mean that a particular style of care or relationship evolves between service providers and consumers that may or may not match the values and technology of the program. For many years these aspects have been acknowledged (but not clearly specified) when decision makers have argued that results of evidence reviews from other countries do not apply in their own country.

Whilst some programs/interventions have been transferred from one context to another and benefits have been observed, others have not (Resnicow 1993, Lumley 2004). Authors should take caution when making generalizations from one context to another. Authors should report on the presence (or otherwise) of context-related information in intervention studies, where this information is available (Hawe 2004).

### **12.3.4 Variation in adherence**

Variation in the adherence of the recipients and providers of care can limit the applicability of results. Predictable differences in adherence can be due to divergence in economic conditions or attitudes that make some forms of care not accessible or not feasible in some settings, such as in developing countries (Dans 2007). It should not be assumed that high levels of adherence in closely monitored randomized trials will translate into similar levels of adherence in normal practice.

### **12.3.5 Variation in values and preferences**

Management decisions involve trading off benefits and downsides of proposed management strategies. The right choice may differ for people with different values and preferences, and it is up to the clinician to ensure that decisions are consistent with patients' values and preferences. We describe how the review author can help this process in Section [12.7](#).

## **12.4 Interpreting results of statistical analyses**

### **12.4.1 Confidence intervals**

Results for both individual studies and meta-analyses are reported with a point estimate together with an associated confidence interval. For example, "The odds ratio was 0.75 with a 95% confidence interval of 0.70 to 0.80". The point estimate (0.75) is the best guess of the magnitude and direction of the experimental intervention's effect compared with the control intervention. The confidence interval describes the uncertainty inherent in this estimate, and describes a range of values within which we

can be reasonably sure that the true effect actually lies. If the confidence interval is relatively narrow (e.g. 0.70 to 0.80), the effect size is known precisely. If the interval is wider (e.g. 0.60 to 0.93) the uncertainty is greater, although there may still be enough precision to make decisions about the utility of the intervention. Intervals that are very wide (e.g. 0.50 to 1.10) indicate that we have little knowledge about the effect, and that further information is needed.

A 95% confidence interval is often interpreted as indicating a range within which we can be 95% certain that the true effect lies. This statement is a loose interpretation, but is useful as a rough guide. The strictly-correct interpretation of a confidence interval is based on the hypothetical notion of considering the results that would be obtained if the study were repeated many times. If a study were repeated infinitely often, and on each occasion a 95% confidence interval calculated, then 95% of these intervals would contain the true effect.

The width of the confidence interval for an individual study depends to a large extent on the sample size. Larger studies tend to give more precise estimates of effects (and hence have narrower confidence intervals) than smaller studies. For continuous outcomes, precision depends also on the variability in the outcome measurements (the standard deviation of measurements across individuals); for dichotomous outcomes it depends on the risk of the event, and for time-to-event outcomes it depends on the number of events observed. All these quantities are used in computation of the standard errors of effect estimates from which the confidence interval is derived.

The width of a confidence interval for a meta-analysis depends on the precision of the individual study estimates and on the number of studies combined. In addition, for random-effects models, precision will decrease with increasing heterogeneity and confidence intervals will widen correspondingly (see Chapter 9, Section 9.5.4). As more studies are added to a meta-analysis the width of the confidence interval usually decreases. However, if the additional studies increase the heterogeneity in the meta-analysis and a random-effects model is used, it is possible that the confidence interval width will increase.

Confidence intervals and point estimates have different interpretations in fixed-effect and random-effects models. While the fixed-effect estimate and its confidence interval address the question 'what is the best (single) estimate of the effect?', the random-effects estimate assumes there to be a distribution of effects, and the estimate and its confidence interval address the question 'what is the best estimate of the average effect?'

A confidence interval may be reported for any level of confidence (although they are most commonly reported for 95%, and sometimes 90% or 99%). For example, the odds ratio of 0.80 could be reported with an 80% confidence interval of 0.73 to 0.88; a 90% interval of 0.72 to 0.89; and a 95% interval of 0.70 to 0.92. As the confidence level increases, the confidence interval widens.

There is logical correspondence between the confidence interval and the P value (see Section 12.4.2). The 95% confidence interval for an effect will exclude the null value (such as an odds ratio of 1.0 or a risk difference of 0) if and only if the test of significance yields a P value of less than 0.05. If the P value is exactly 0.05, then either the upper or lower limit of the 95% confidence interval will be at the null value. Similarly, the 99% confidence interval will exclude the null if and only if the test of significance yields a P value of less than 0.01.

Together, the point estimate and confidence interval provide information to assess the clinical usefulness of the intervention. For example, suppose that we are evaluating a treatment that reduces the risk of an event and we decide that it would be useful only if it reduced the risk of an event from

30% by at least 5 percentage points to 25% (these values will depend on the specific clinical scenario and outcome). If the meta-analysis yielded an effect estimate of a reduction of 10 percentage points with a tight 95% confidence interval, say, from 7% to 13%, we would be able to conclude that the treatment was useful since both the point estimate and the entire range of the interval exceed our criterion of a reduction of 5% for clinical usefulness. However, if the meta-analysis reported the same risk reduction of 10% but with a wider interval, say, from 2% to 18%, although we would still conclude that our best estimate of the effect of treatment is that it is useful, we could not be so confident as we have not excluded the possibility that the effect could be between 2% and 5%. If the confidence interval was wider still, and included the null value of a difference of 0%, we will not have excluded the possibility that the treatment has any effect whatsoever, and would need to be even more sceptical in our conclusions.

Confidence intervals with different levels of confidence can demonstrate that there is differential evidence for different degrees of benefit or harm. For example, it might be possible to report the same analysis results (i) with 95% confidence that the intervention does not cause harm; (ii) with 90% confidence that it has some effect; and (iii) with 80% confidence that it has a patient important benefit. These elements may suggest both usefulness of the intervention and the need for additional research.

Review authors may use the same general approach to conclude that an intervention is *not* useful. Continuing with the above example where the criterion for a minimal patient-important difference is a 5% risk difference, an effect estimate of 2% with a confidence interval of 1% to 4% suggests that the intervention is not useful.

## 12.4.2 P values and statistical significance

A P value is the probability of obtaining the observed effect (or larger) under a ‘null hypothesis’, which in the context of Cochrane reviews is either an assumption of ‘no effect of the intervention’ or ‘no differences in the effect of intervention between studies’ (no heterogeneity). Thus, a P value that is very small indicates that the observed effect is very unlikely to have arisen purely by chance, and therefore provides evidence against the null hypothesis. It has been common practice to interpret a P value by examining whether it is smaller than particular threshold values. In particular, P values less than 0.05 are often reported as “statistically significant”, and interpreted as being small enough to justify rejection of the null hypothesis. However, the 0.05 threshold is an arbitrary one that became commonly used in medical and psychological research largely because P values were determined by comparing the test statistic against tabulations of specific percentage points of statistical distributions. RevMan, like other statistical packages, reports precise P values. If review authors decide to present a P value with the results of a meta-analysis, they should report a precise P value, together with the 95% confidence interval.

In RevMan, two P values are provided. One relates to the summary effect in a meta-analysis and is from a Z test of the null hypothesis that there is no effect (or no effect on average in a random-effects meta-analysis). The other relates to heterogeneity between studies and is from a chi-squared test of the null hypothesis that there is no heterogeneity (see Chapter 9, Section 9.5.2).

For tests of a summary effect, the computation of P involves both the effect estimate and the sample size (or, more strictly, the precision of the effect estimate). As sample size increases, the range of plausible effects that could occur by chance is reduced. Correspondingly, the statistical significance of an effect of a particular magnitude will be greater (the P value will be smaller) in a larger study than in a smaller study.

P values are commonly misinterpreted in two ways. First, a moderate or large P value (e.g. greater than 0.05) may be misinterpreted as evidence that “the intervention has no effect”. There is an important difference between this statement and the correct interpretation that “there is not strong evidence that the intervention has an effect”. To avoid such a misinterpretation, review authors should always examine the effect estimate and its 95% confidence interval, together with the P value. In small studies or small meta-analyses it is common for the range of effects contained in the confidence interval to include both no intervention effect and a substantial effect. Review authors are advised not to describe results as ‘not statistically significant’ or ‘non-significant’.

The second misinterpretation is to assume that a result with a small P value for the summary effect estimate implies that an intervention has an important benefit. Such a misinterpretation is more likely to occur in large studies, such as meta-analyses that accumulate data over dozens of studies and thousands of participants. The P value addresses the question of whether the intervention effect is precisely nil; it does not examine whether the effect is of a magnitude of importance to potential recipients of the intervention. In a large study, a small P value may represent the detection of a trivial effect. Again, inspection of the point estimate and confidence interval helps correct interpretations (see Section 12.4.1).

## 12.5 Interpreting results from dichotomous outcomes (including numbers needed to treat)

### 12.5.1 Relative and absolute risk reductions

Clinicians may be more inclined to prescribe an intervention that reduces the risk of death by 25% than one that reduces the risk of death by 1 percentage point, although both presentations of the evidence may relate to the same benefit (i.e. a reduction in risk from 4% to 3%). The former refers to the *relative* reduction in risk and the latter to the *absolute* reduction in risk. As described in Chapter 9 (Section 9.2.2), there are several measures for comparing dichotomous outcomes in two groups. Meta-analyses are usually undertaken using risk ratios (RR), odds ratios (OR) or risk differences (RD), but there are several alternative ways of expressing results.

**Relative risk reduction** (RRR) is a convenient way of re-expressing a risk ratio as a percentage reduction:

$$\text{RRR} = 100\% \times (1 - \text{RR}).$$

For example, a risk ratio of 0.75 translates to a relative risk reduction of 25%, as in the example above.

The risk difference is often referred to as the **absolute risk reduction** (ARR), and may be presented as a percentage (for example, 1%), as a decimal (for example, 0.01), or as counts, (for example, 10 out of 1000). A simple transformation of the risk difference known as the number needed to treat (NNT) is a common alternative way of presenting the same information. We discuss NNTs in Section 12.5.2, and consider different choices for presenting absolute effects in Section 12.5.3. We then describe computations for obtaining these numbers from the results of individual studies and of meta-analyses.

### 12.5.2 More about the number needed to treat (NNT)

The **number needed to treat** (NNT) is defined as the expected number of people who need to receive the experimental rather than the comparator intervention for one additional person to either incur or avoid an event in a given time frame. Thus, for example, an NNT of 10 can be interpreted as ‘it is expected that one additional (or less) person will incur an event for every 10 participants receiving the experimental intervention rather than control over a given time frame’. It is important to be clear that:

1. since the NNT is derived from the risk difference, it is still a *comparative* measure of effect (experimental versus a certain control) and not a general property of a single intervention; and
2. the NNT gives an ‘expected value’. For example,  $NNT = 10$  does not imply that one additional event *will* occur in each and every group of ten people.

NNTs can be computed for both beneficial and detrimental events, and for interventions that cause both improvements and deteriorations in outcomes. In all instances NNTs are expressed as positive whole numbers, all decimals being rounded up. Some authors use the term ‘number needed to harm’ (NNH) when an intervention leads to a deterioration rather than improvement in outcome. However, this phrase is unpleasant, misleading and inaccurate (most notably, it can easily be read to imply the number of people who will experience a harmful outcome if given the intervention), and it is strongly recommended that ‘number needed to harm’ and ‘NNH’ are avoided. The preferred alternative is to use phrases such as ‘number needed to treat for an additional beneficial outcome’ (NNTB) and ‘number needed to treat for an additional harmful outcome’ (NNTH) to indicate direction of effect.

As NNTs refer to events, their interpretation needs to be worded carefully when the binary outcome is a dichotomization of a scale-based outcome. For example, if the outcome is pain measured on a ‘none, mild, moderate or severe’ scale it may have been dichotomized as ‘none or mild’ versus ‘moderate or severe’. It would be inappropriate for an NNT from these data to be referred to as an ‘NNT for pain’. It is an ‘NNT for moderate or severe pain’.

### 12.5.3 Expressing absolute risk reductions

Users of reviews are liable to be influenced by the choice of statistical presentations of the evidence. Hoffrage et al. suggest that physicians’ inferences about statistical outcomes are more appropriate when they deal with ‘natural frequencies’ – whole numbers of people, both treated and untreated – (e.g. treatment results in a drop from 20 out of 1000 to 10 out of 1000 women having breast cancer), than when effects are presented as percentages (e.g. 1% absolute reduction in breast cancer risk) (Hoffrage 2000). Probabilities may be more difficult to understand than frequencies, particularly when events are rare. While standardization may be important in improving the presentation of research evidence (and participation in healthcare decisions), current evidence suggests that the presentation of natural frequencies for expressing differences in absolute risk is best understood by consumers of healthcare information. This evidence provides the rationale for presenting absolute risks in ‘Summary of findings’ tables as numbers of people with events per 1000 people receiving the intervention.

Risk ratios and relative risk reductions remain crucial because relative effect tends to be substantially more stable across risk groups than does absolute benefit. Review authors can use their own data to study this consistency (Cates 1999, Smeeth 1999). Risk differences are least likely to be consistent across baseline event rates; thus, they are rarely appropriate for computing numbers needed to treat in systematic reviews. If a relative effect measure (OR or RR) is chosen for meta-analysis, then a control group risk needs to be specified as part of the calculation of an ARR or NNT. It is crucial to express absolute benefit for each clinically identifiable risk group, clarifying the time period to which this applies. Studies in patients with differing severity of disease, or studies with different lengths of follow-up will almost certainly have different control group risks. In these cases, different control group risks lead to different ARRs and NNTs (except when the intervention has no effect). A recommended approach is to re-express an odds ratio or a risk ratio as a variety of NNTs across a range of assumed control risks (ACRs) (McQuay 1997, Smeeth 1999, Sackett 2000). Review authors should bear these considerations in mind not only when constructing their ‘Summary of findings’ table, but also in the text of their review.

For example a review of oral anticoagulants to prevent stroke presented information to users by describing absolute benefits for various baseline risks (Aguilar 2005). They presented their principal findings as “The inherent risk of stroke should be considered in the decision to use oral anticoagulants in atrial fibrillation patients, selecting those who stand to benefit most for this therapy” (Aguilar 2005). Among high-risk atrial fibrillation patients with prior stroke or transient ischaemic attack who have stroke rates of about 12% (120 per 1000) per year, warfarin prevents about 70 strokes yearly per 1000 patients, whereas for low-risk atrial fibrillation patients (with a stroke rate of about 2% per year or 20 per 1000), warfarin prevents only 12 strokes. This presentation helps users to understand the important impact that typical baseline risks have on the absolute benefit that they can expect.

## 12.5.4 Computations

Direct computation of an absolute risk reduction (ARR) or a number needed to treat (NNT) depends on the summary statistic (odds ratio, risk ratio or risk differences) available from the study or meta-analysis. When expressing results of meta-analyses, authors should use, in the computations, whatever statistic they determined to be the most appropriate summary for pooling (see Chapter 9, Section 9.4.4.4). Here we present calculations to obtain ARR as a reduction in the number of participants per 1000. For example, a risk difference of  $-0.133$  corresponds to 133 *fewer* participants with the event per 1000.

ARRs and NNTs should not be computed from the aggregated total numbers of participants and events across the trials. This approach ignores the randomization within studies, and may produce seriously misleading results if there is unbalanced randomization in any of the studies.

When computing NNTs, the values obtained are by convention always rounded up to the next whole number.

### 12.5.4.1 Computing NNT from a risk difference (RD)

NNTs can be calculated for single studies as follows. Note that this approach, although applicable, should only very rarely be used for the results of a meta-analysis of risk differences, because meta-analyses should usually be undertaken using a relative measure of effect (RR or OR).

A NNT may be computed from a risk difference as

$$\text{NNT} = \frac{1}{\text{absolute value of risk difference}} = \frac{1}{|\text{RD}|},$$

where the vertical bars (‘absolute value of’) in the denominator indicate that any minus sign should be ignored. It is convention to round the NNT up to the nearest whole number. For example, if the risk difference is  $-0.12$  the NNT is 9; if the risk difference is  $-0.22$  the NNT is 5.

### 12.5.4.2 Computing absolute risk reduction or NNT from a risk ratio (RR)

To aid interpretation, review authors may wish to compute an absolute risk reduction or NNT from the results of a meta-analysis of risk ratios. In order to do this, an assumed control risk (ACR) is required. It will usually be appropriate to do this for a range of different ACRs. The computation proceeds as follows:

$$\text{number fewer per 1000} = 1000 \times \text{ACR} \times (1 - \text{RR}),$$

$$\text{NNT} = \left\lceil \frac{1}{\text{ACR} \times (1 - \text{RR})} \right\rceil$$

As an example, suppose the risk ratio is  $RR = 0.92$ , and an assumed control risk of  $ACR = 0.3$  (300 per 1000) is assumed. Then the effect on risk is 24 fewer per 1000:

$$\text{number fewer per 1000} = 1000 \times 0.3 \times (1 - 0.92) = 24$$

The NNT is 42:

$$NNT = \left| \frac{1}{0.3 \times (1 - 0.92)} \right| = \left| \frac{1}{0.3 \times 0.08} \right| = 41.67$$

#### 12.5.4.3 Computing absolute risk reduction or NNT from an odds ratio (OR)

Review authors may wish to compute an absolute risk reduction or NNT from the results of a meta-analysis of odds ratios. In order to do this, an assumed control risk (ACR) is required. It will usually be appropriate to do this for a range of different ACRs. The computation proceeds as follows:

$$\begin{aligned} \text{number fewer per 1000} &= 1000 \times \left( ACR - \frac{OR \times ACR}{1 - ACR + OR \times ACR} \right) \\ NNT &= \frac{1}{\left| ACR - \frac{OR \times ACR}{1 - ACR + OR \times ACR} \right|} \end{aligned}$$

As an example, suppose the odds ratio is  $OR = 0.73$ , and a control risk of  $ACR = 0.3$  is assumed. Then the effect on risk is 62 fewer per 1000:

$$\begin{aligned} \text{number fewer per 1000} &= 1000 \times \left( 0.3 - \frac{0.73 \times 0.3}{1 - 0.3 + 0.73 \times 0.3} \right) \\ &= 1000 \times \left( 0.3 - \frac{0.219}{1 - 0.3 + 0.219} \right) = 1000 \times (0.3 - 0.238) = 61.7 \end{aligned}$$

The NNT is 17:

$$NNT = \frac{1}{\left| \left( 0.3 - \frac{0.73 \times 0.3}{1 - 0.3 + 0.73 \times 0.3} \right) \right|} = \frac{1}{\left| 0.3 - \frac{0.219}{1 - 0.3 + 0.219} \right|} = \frac{1}{|0.3 - 0.238|} = 16.2$$

#### 12.5.4.4 Computing risk ratio from an odds ratio (OR)

Because risk ratios are easier to interpret than odds ratios, but odds ratios have favourable mathematical properties, a review author may decide to undertake a meta-analysis based on odds ratios, but to express the result as a summary risk ratio (or relative risk reduction). This requires an assumed control risk (ACR). Then

$$RR = \frac{OR}{1 - ACR \times (1 - OR)}$$

It will often be reasonable to perform this transformation using the median control group risk from the studies in the meta-analysis.

### **12.5.4.5 Computing confidence limits**

Confidence limits for ARR and NNT may be calculated by applying the above formulae to the upper and lower confidence limits for the summary statistic (RD, RR or OR) (Altman 1998). Note that this confidence interval does not incorporate uncertainty around the control group risk (CGR).

In the case of what conventionally are considered non-statistically significant results (for example, the 95% confidence interval of OR or RR includes the value 1) one of the confidence limits will indicate benefit and the other harm. Thus, appropriate use of the words ‘fewer’ and ‘more’ is required for each limit when presenting results in terms of events. For NNTs, the two confidence limits should be labelled as NNTB and NNTH to indicate the direction of effect in each case. The confidence interval for the NNT will include a ‘discontinuity’: within the interval there will be an infinitely large NNTB, which will switch to an infinitely large NNTH.

## **12.6 Interpreting results from continuous outcomes (including standardized mean differences)**

### **12.6.1 Meta-analyses with continuous outcomes**

When outcomes are continuous, review authors have a number of options in presenting pooled results. If all studies have used the same units, a meta-analysis may generate a pooled estimate in those units, as a difference in mean response (see, for instance, the row summarizing results for oedema in Chapter 11, Figure 11.5.a). The units of such outcomes may be difficult to interpret, particularly when they relate to rating scales. ‘Summary of findings’ tables should include the minimum and maximum of the scale of measurement, and the direction (again, see the Oedema column of Chapter 11, Figure 11.5.a). Knowledge of the smallest change in instrument score that patients perceive is important – the minimal important difference – and can greatly facilitate the interpretation of results. Knowing the minimal important difference allows authors and users to place results in context, and authors should state the minimal important difference – if known – in the Comments column of their ‘Summary of findings’ table.

When studies have used different instruments to measure the same construct, a standardized mean difference (SMD) may be used in meta-analysis for combining continuous data (see Chapter 9, Section 9.2.3.2). For clinical interpretation, such an analysis may be less helpful than dichotomizing responses and presenting proportions of patients benefiting. Methods are available for creating dichotomous data out of reported means and standard deviations, but require assumptions that may not be met (Suijsa 1991, Walter 2001).

The SMD expresses the intervention effect in standard units rather than the original units of measurement. The SMD is the difference in mean effects in the experimental and control groups divided by the pooled standard deviation of participants’ outcomes (see Chapter 9, Section 9.2.3.2). The value of a SMD thus depends on both the size of the effect (the difference between means) and the standard deviation of the outcomes (the inherent variability among participants).

Without guidance, clinicians and patients may have little idea how to interpret results presented as SMDs. There are several possibilities for re-expressing such results in more helpful ways, as follows.

### **12.6.2 Re-expressing SMDs using rules of thumb for effect sizes**

Rules of thumb exist for interpreting SMDs (or ‘effect sizes’), which have arisen mainly from researchers in the social sciences. One example is as follows: 0.2 represents a small effect, 0.5 a moderate effect, and 0.8 a large effect (Cohen 1988). Variations exist (for example, <0.41 = small,

0.40 to 0.70 = moderate, >0.70 = large). Review authors might consider including a rule of thumb in the Comments column of a ‘Summary of findings’ table. However, some methodologists believe that such interpretations are problematic because *patient* importance of a finding is context-dependent and not amenable to generic statements.

### 12.6.3 Re-expressing SMDs by transformation to odds ratio

A transformation of a SMD to a (log) odds ratio is available, based on the assumption that an underlying continuous variable has a logistic distribution with equal standard deviation in the two intervention groups (Furukawa 1999, Chinn 2000). The assumption is unlikely to hold exactly and the results must be regarded as an approximation. The log odds ratio is estimated as

$$\ln\text{OR} = \frac{\pi}{\sqrt{3}}\text{SMD},$$

(or approximately 1.81×SMD) The resulting odds ratio can then be combined with an assumed control group risk to obtain an absolute risk reduction as in Section 12.5.4.3. These control group risks refer to proportions of people who have improved by some (unspecified) amount in the continuous outcome (‘responders’). Table 12.6.a shows some illustrative results from this method. These NNTs may be converted to people per thousand by using the formula 1000/NNT.

**Table 12.6.a: NNTs equivalent to specific SMDs for various given ‘proportions improved’ in the control group**

| Control group proportion improved | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|-----------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SMD = 0.1                         | 57  | 33  | 26  | 23  | 23  | 24  | 28  | 37  | 66  |
| SMD = 0.2                         | 27  | 16  | 13  | 12  | 12  | 13  | 15  | 20  | 36  |
| SMD = 0.5                         | 9   | 6   | 5   | 5   | 5   | 6   | 7   | 10  | 18  |
| SMD = 0.8                         | 5   | 4   | 3   | 3   | 4   | 4   | 5   | 7   | 14  |
| SMD = 1.0                         | 4   | 3   | 3   | 3   | 3   | 4   | 5   | 7   | 13  |

### 12.6.4 Re-expressing SMDs using a familiar instrument

The final possibility for interpreting the SMD is to express it in the units of one or more of the specific measurement instruments. Multiplying a SMD by a typical among-person standard deviation for a particular scale yields an estimate of the difference in mean outcome scores (experimental versus control) on that scale. The standard deviation could be obtained as the pooled standard deviation of baseline scores in one of the studies. To better reflect among-person variation in practice, it may be preferable to use a standard deviation from a representative observational study. The pooled effect is thus re-expressed in the original units of that particular instrument and the clinical relevance and impact of the intervention effect can be interpreted. However, authors should be aware that such back-transformation of effect sizes can be misleading if it is applied to individual studies rather than for a summary measure of effect (Scholten 1999). Consider two studies that *did* use the same instrument and observed the same effect, but observed different among-participant variability (perhaps due to different inclusion criteria). Then back-transformations using the different standard deviations from these studies would yield different sizes of effect for *the same scale* and *the same effect*.

## 12.7 Drawing conclusions

### 12.7.1 Conclusions sections of a Cochrane review

Authors' conclusions from a Cochrane review are divided into implications for practice and implications for research. In deciding what these implications are, it is useful to consider four factors: the quality of evidence, the balance of benefits and harms, values and preferences and resource utilization (Eddy 1990). Considering these factors involves judgements and effort that go beyond the work of most review authors.

### 12.7.2 Implications for practice

Drawing conclusions about the practical usefulness of an intervention entails making trade-offs, either implicitly or explicitly, between the estimated benefits, harms and the estimated costs. Making such trade-offs, and thus making specific recommendations for an action, goes beyond a systematic review and requires additional information and informed judgements that are typically the domain of clinical practice guideline developers. Authors of Cochrane reviews should not make recommendations.

If authors feel compelled to lay out actions that clinicians and patients could take, they should – after describing the quality of evidence and the balance of benefits and harms – highlight different actions that might be consistent with particular patterns of values and preferences. Other factors that might influence a decision should also be highlighted, including any known factors that would be expected to modify the effects of the intervention, the baseline risk or status of the patient, costs and who bears those costs, and the availability of resources. Authors should ensure they consider all patient-important outcomes, including those for which limited data may be available. This process implies a high level of explicitness about judgements about values or preferences attached to different outcomes. The highest level of explicitness would involve a formal economic analysis with sensitivity analysis involving different assumptions about values and preferences; this is beyond the scope of most Cochrane reviews (although they might well be used for such analyses) (Mugford 1989, Mugford 1991); this is discussed in Chapter 15.

A review on the use of anticoagulation in cancer patients to increase survival (Akl 2007) provides an example for laying out clinical implications for situations where there are important trade-offs between desirable and undesirable effects of the intervention: “The decision for a patient with cancer to start heparin therapy for survival benefit should balance the benefits and downsides and integrate the patient’s values and preferences (Haynes 2002). Patients with a high preference for survival prolongation (even though that prolongation may be short) and limited aversion to bleeding who do not consider heparin therapy a burden may opt to use heparin, while those with aversion to bleeding and the related burden of heparin therapy may not.”

### 12.7.3 Implications for research

Review conclusions should help people make well-informed decisions about future healthcare research. The ‘Implications for research’ should comment on the need for further research, and the nature of the further research that would be most desirable. A format has been proposed for reporting research recommendations (‘EPICOT’), as follows (Brown 2006):

- E (Evidence): What is the current evidence?
- P (Population): Diagnosis, disease stage, co-morbidity, risk factor, sex, age, ethnic group, specific inclusion or exclusion criteria, clinical setting.
- I (Intervention): Type, frequency, dose, duration, prognostic factor.
- C (Comparison): Placebo, routine care, alternative treatment/management.

- O (Outcome): Which clinical or patient-related outcomes will the researcher need to measure, improve, influence or accomplish? Which methods of measurement should be used?
- T (Time stamp): Date of literature search or recommendation.

Other factors that might be considered in recommendations include the disease burden of the condition being addressed, the timeliness (e.g. length of follow-up, duration of intervention), and the study type that would best suit subsequent research (Brown 2006).

Cochrane review authors should ensure that they include the PICO aspects of this format. It is also helpful to note the study types, as well as any particular design features, that would best address the research question.

A review of compression stockings for prevention of deep vein thrombosis in airline passengers provides an example where there is some convincing evidence of a benefit of the intervention: “This review shows that the question of the effects on symptomless DVT of wearing versus not wearing compression stockings in the types of people studied in these trials should now be regarded as answered. Further research may be justified to investigate the relative effects of different strengths of stockings or of stockings compared to other preventative strategies. Further randomized trials to address the remaining uncertainty about the effects of wearing versus not wearing compression stockings on outcomes such as death, pulmonary embolus and symptomatic DVT would need to be large.” (Clarke 2006).

A review of therapeutic touch for anxiety disorder provides an example of the implications for research when no eligible studies had been found: “This review highlights the need for randomized controlled trials to evaluate the effectiveness of therapeutic touch in reducing anxiety symptoms in people diagnosed with anxiety disorders. Future trials need to be rigorous in design and delivery, with subsequent reporting to include high quality descriptions of all aspects of methodology to enable appraisal and interpretation of results.” (Robinson 2007).

#### **12.7.4 Common errors in reaching conclusions**

A common mistake when there is inconclusive evidence is to confuse ‘no evidence of an effect’ with ‘evidence of no effect’. When there is inconclusive evidence, it is wrong to claim that it shows that an intervention has ‘no effect’ or is ‘no different’ from the control intervention. It is safer to report the data, with a confidence interval, as being compatible with either a reduction or an increase in the outcome. When there is a ‘positive’ but statistically non-significant trend authors commonly describe this as ‘promising’, whereas a ‘negative’ effect of the same magnitude is not commonly described as a ‘warning sign’; such language may be harmful.

Another mistake is to frame the conclusion in wishful terms. For example, authors might write “the included studies were too small to detect a reduction in mortality” when the included studies showed a reduction or even increase in mortality that failed to reach conventional levels of statistical significance. One way of avoiding errors such as these is to consider the results blinded; i.e. consider how the results would be presented and framed in the conclusions had the direction of the results been reversed. If the confidence interval for the estimate of the difference in the effects of the interventions overlaps the null value, the analysis is compatible with both a true beneficial effect and a true harmful effect. If one of the possibilities is mentioned in the conclusion, the other possibility should be mentioned as well.

Another common mistake is to reach conclusions that go beyond the evidence. Often this is done implicitly, without referring to the additional information or judgements that are used in reaching

conclusions about the implications of a review for practice. Even when additional information and explicit judgements support conclusions about the implications of a review for practice, review authors rarely conduct systematic reviews of the additional information. Furthermore, implications for practice are often dependent on specific circumstances and values that must be taken into consideration. As we have noted, authors should always be cautious when drawing conclusions about implications for practice and they should not make recommendations.

## 12.8 Chapter information

**Authors:** Holger J Schünemann, Andrew D Oxman, Gunn E Vist, Julian PT Higgins, Jonathan J Deeks, Paul Glasziou and Gordon H Guyatt on behalf of the Cochrane Applicability and Recommendations Methods Group.

**This chapter should be cited as:** Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou P, Guyatt GH. Chapter 12: Interpreting results and drawing conclusions. In: Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from [www.cochrane-handbook.org](http://www.cochrane-handbook.org).

**Acknowledgements:** Jonathan Sterne, Michael Borenstein and Rob JM Scholten contributed text.

**Declarations of interest:** Holger Schünemann, Andrew Oxman, Gunn Vist, Paul Glasziou and Gordon Guyatt have, to varying degrees, taken leadership roles in the GRADE Working Group from which many of the ideas in this chapter have arisen.

### Box 12.8.a: The Cochrane Applicability and Recommendations Methods Group

We anticipate continued evolution of the methodologies described in this chapter. The main arenas in which relevant discussions will take place are the Applicability and Recommendations Methods Group (ARMG) and the GRADE Working Group. Both discussion groups welcome new participants with an eagerness to learn more and to contribute to further developments in rating quality of evidence, and in framing issues in the application of Cochrane reviews.

The Applicability and Recommendations Methods Group (ARMG) is comprised of individuals with interest and expertise in the interpretation, applicability and transferability of the results of systematic reviews to individuals and groups. The ARMG's objective is to explore the process of going from evidence to healthcare recommendations. The ultimate goals are to make this process as rigorous as possible.

Specific areas currently considered important include:

- evaluating the quality of evidence ([www.gradeworkinggroup.org](http://www.gradeworkinggroup.org));
- variation of effect with baseline risk;
- prediction of benefit from the patient's expected event rate or severity;
- consideration of how the strength of evidence and the magnitude and precision of the effects bear on the implications; and
- consideration of how people's values bear on the implications when weighing benefits and harms based on individual clinical features.

## 12.9 References

### **Aguilar 2005**

Aguilar MI, Hart R. Oral anticoagulants for preventing stroke in patients with non-valvular atrial fibrillation and no previous history of stroke or transient ischemic attacks. *Cochrane Database of Systematic Reviews* 2005, Issue 3. Art No: CD001927.

### **Akl 2007**

Akl EA, Kamath G, Kim SY, Yosunico V, Barba M, Terrenato I, Sperati F, Schünemann HJ. Oral anticoagulation for prolonging survival in patients with cancer. *Cochrane Database of Systematic Reviews* 2007, Issue 2. Art No: CD006466.

### **Alonso-Coello 2006**

Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, Mills E, Heels-Ansdell D, Johanson JF, Guyatt G. Meta-analysis of flavonoids for the treatment of haemorrhoids. *British Journal of Surgery* 2006; 93: 909-920.

### **Altman 1998**

Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998; 317: 1309-1312.

### **Bhandari 2004**

Bhandari M, Busse JW, Jackowski D, Montori VM, Schünemann H, Sprague S, Mears D, Schemitsch EH, Heels-Ansdell D, Devereaux PJ. Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. *Canadian Medical Association Journal* 2004; 170: 477-480.

### **Brophy 2001**

Brophy JM, Joseph L, Rouleau JL. Beta-blockers in congestive heart failure. A Bayesian meta-analysis. *Annals of Internal Medicine* 2001; 134: 550-560.

### **Brown 2006**

Brown P, Brunnhuber K, Chalkidou K, Chalmers I, Clarke M, Fenton M, Forbes C, Glanville J, Hicks NJ, Moody J, Twaddle S, Timimi H, Young P. How to formulate research recommendations. *BMJ* 2006; 333: 804-806.

### **Cates 1999**

Cates C. Confidence intervals for the number needed to treat: Pooling numbers needed to treat may not be reliable. *BMJ* 1999; 318: 1764-1765.

### **Chambers 2005**

Chambers BR, Donnan GA. Carotid endarterectomy for asymptomatic carotid stenosis. *Cochrane Database of Systematic Reviews* 2005, Issue 4. Art No: CD001923.

### **Chinn 2000**

Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine* 2000; 19: 3127-3131.

### **Cina 2000**

Cina CS, Clase CM, Haynes RB. Carotid endarterectomy for symptomatic carotid stenosis. *Cochrane Database of Systematic Reviews* 2000, Issue 2. Art No: CD001081.

### **Clarke 2006**

Clarke M, Hopewell S, Juszczak E, Eisinga A, Kjeldstrøm M. Compression stockings for preventing deep vein thrombosis in airline passengers. *Cochrane Database of Systematic Reviews* 2006, Issue 2. Art No: CD004002.

### **Cohen 1988**

Cohen J. *Statistical Power Analysis in the Behavioral Sciences* (2nd edition). Hillsdale (NJ): Lawrence Erlbaum Associates, Inc., 1988.

**Dans 2007**

Dans AM, Dans L, Oxman AD, Robinson V, Acuin J, Tugwell P, Dennis R, Kang D. Assessing equity in clinical practice guidelines. *Journal of Clinical Epidemiology* 2007; 60: 540-546.

**Devereaux 2004**

Devereaux PJ, Choi PT, El-Dika S, Bhandari M, Montori VM, Schünemann HJ, Garg AX, Busse JW, Heels-Ansdell D, Ghali WA, Manns BJ, Guyatt GH. An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. *Journal of Clinical Epidemiology* 2004; 57: 1232-1236.

**Eddy 1990**

Eddy DM. Clinical decision making: from theory to practice. Anatomy of a decision. *JAMA* 1990; 263: 441-443.

**Friedman 1985**

Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials* (2nd edition). Littleton (MA): John Wright PSG, Inc., 1985.

**Furukawa 1999**

Furukawa TA. From effect size into number needed to treat. *The Lancet* 1999; 353: 1680.

**Gibson 2007**

Gibson JN, Waddell G. Surgical interventions for lumbar disc prolapse. *Cochrane Database of Systematic Reviews* 2007, Issue 2. Art No: CD001350.

**GRADE Working Group 2004**

GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004; 328: 1490-1494.

**Guyatt 1994**

Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA* 1994; 271: 59-63.

**Guyatt 2006a**

Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schünemann H. Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians Task Force. *Chest* 2006; 129: 174-181.

**Guyatt 2006b**

Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schünemann H. An emerging consensus on grading recommendations? *ACP Journal Club* 2006; 144: A8-A9.

**Guyatt 2008a**

Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is 'quality of evidence' and why is it important to clinicians? *BMJ* 2008; 336: 995-998.

**Guyatt 2008b**

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336: 924-926.

**Hawe 2004**

Hawe P, Shiell A, Riley T, Gold L. Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. *Journal of Epidemiology and Community Health* 2004; 58: 788-793.

**Haynes 2002**

Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *ACP Journal Club* 2002; 136: A11-A14.

**Hoffrage 2000**

Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Medicine. Communicating statistical information. *Science* 2000; 290: 2261-2262.

**Levine 2004**

Levine MN, Raskob G, Beyth RJ, Kearon C, Schulman S. Hemorrhagic complications of anticoagulant treatment: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 2004; 126: 287S-310S.

**Lumley 2004**

Lumley J, Oliver SS, Chamberlain C, Oakley L. Interventions for promoting smoking cessation during pregnancy. *Cochrane Database of Systematic Reviews* 2004, Issue 4. Art No: CD001055.

**McIntosh 2005**

McIntosh HM, Jones KL. Chloroquine or amodiaquine combined with sulfadoxine-pyrimethamine for treating uncomplicated malaria. *Cochrane Database of Systematic Reviews* 2005, Issue 4. Art No: CD000386.

**McQuay 1997**

McQuay HJ, Moore A. Using numerical results from systematic reviews in clinical practice. *Annals of Internal Medicine* 1997; 126: 712-720.

**Mugford 1989**

Mugford M, Kingston J, Chalmers I. Reducing the incidence of infection after caesarean section: implications of prophylaxis with antibiotics for hospital resources. *BMJ* 1989; 299: 1003-1006.

**Mugford 1991**

Mugford M, Piercy J, Chalmers I. Cost implications of different approaches to the prevention of respiratory distress syndrome. *Archives of Disease in Childhood* 1991; 66: 757-764.

**Oxman 2002**

Oxman A, Guyatt G. When to believe a subgroup analysis. In: Guyatt G, Rennie D (editors). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago (IL): AMA Press, 2002.

**Resnicow 1993**

Resnicow K, Cross D, Wynder E. The Know Your Body program: a review of evaluation studies. *Bulletin of the New York Academy of Medicine* 1993; 70: 188-207.

**Robinson 2007**

Robinson J, Biley FC, Dolk H. Therapeutic touch for anxiety disorders. *Cochrane Database of Systematic Reviews* 2007, Issue 3. Art No: CD006240.

**Sackett 2000**

Sackett DL, Richardson WS, Rosenberg W, Haynes BR. *Evidence-Based Medicine: How to Practice and Teach EBM* (2nd edition). Edinburgh (UK): Churchill Livingstone, 2000.

**Salpeter 2007**

Salpeter S, Greyber E, Pasternak G, Salpeter E. Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. *Cochrane Database of Systematic Reviews* 2007, Issue 4. Art No: CD002967.

**Scholten 1999**

Scholten RJPM. From effect size into number needed to treat [letter]. *The Lancet* 1999; 453: 598.

**Schünemann 2006a**

Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 13. Applicability, transferability and adaptation. *Health Research Policy and Systems* 2006; 4: 25.

**Schünemann 2006b**

Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, Fahy BF, Gould MK, Horan

KL, Krishnan JA, Manthous CA, Maurer JR, McNicholas WT, Oxman AD, Rubenfeld G, Turino GM, Guyatt G. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *American Journal of Respiratory and Critical Care Medicine* 2006; 174: 605-614.

**Smeeth 1999**

Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses - sometimes informative, usually misleading. *BMJ* 1999; 318: 1548-1551.

**Suissa 1991**

Suissa S. Binary methods for continuous outcomes: a parametric alternative. *Journal of Clinical Epidemiology* 1991; 44: 241-248.

**Thompson 2000**

Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. *Cochrane Database of Systematic Reviews* 2000, Issue 2. Art No: CD001855.

**Walter 2001**

Walter SD. Number needed to treat (NNT): estimation of a measure of clinical benefit. *Statistics in Medicine* 2001; 20: 3947-3962.