

Chapter 7: Selecting studies and collecting data

Editors: Julian PT Higgins and Jonathan J Deeks.

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 7.9. 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.

Key points

- Assessment of eligibility of studies, and extraction of data from study reports, should be done by at least two people, independently.
- Cochrane Intervention reviews have studies, rather than reports, as the unit of interest, and so multiple reports of the same study need to be linked together.
- Data collection forms are invaluable. They should be designed carefully to target the objectives of the review, and should be piloted for each new review (or review team).
- Tips are available for helping with the design and use of data collection forms.
- Data may be reported in diverse formats, but can often be converted into a format suitable for meta-analysis.

7.1 Introduction

The findings of a systematic review depend critically on decisions relating to which studies are included, and on decisions relating to which data from these studies are presented and analysed. Methods used for these decisions must be transparent, and they should be chosen to minimize biases and human error. Here we describe approaches that should be used in Cochrane reviews for selecting studies and deciding which of their data to present.

7.2 Selecting studies

7.2.1 Studies (not reports) as the unit of interest

A Cochrane review is a review of *studies* that meet pre-specified criteria for inclusion in the review. Since each study may have been reported in several articles, abstracts or other reports, a comprehensive search for studies for the review may identify many *reports* from potentially relevant studies. Two distinct processes are therefore required to determine which studies can be included in the review. One is to link together multiple reports of the same study; and the other is to use the information available in the various reports to determine which studies are eligible for inclusion. Although sometimes there is a single report for each study, it should never be assumed that this is the case.

7.2.2 Identifying multiple reports from the same study

Duplicate publication can introduce substantial biases if studies are inadvertently included more than once in a meta-analysis (Tramèr 1997). Duplicate publication can take various forms, ranging from identical manuscripts to reports describing different numbers of participants and different outcomes (von Elm 2004). It can be difficult to detect duplicate publication, and some ‘detective work’ by the review authors may be required.

Some of the most useful criteria for comparing reports are:

- author names (most duplicate reports have authors in common, although it is not always the case);
- location and setting (particularly if institutions, such as hospitals, are named);
- specific details of the interventions (e.g. dose, frequency);
- numbers of participants and baseline data; and
- date and duration of the study (which can also clarify whether different sample sizes are due to different periods of recruitment).

Where uncertainties remain after considering these and other factors, it may be necessary to correspond with the authors of the reports.

7.2.3 A typical process for selecting studies

A typical process for selecting studies for inclusion in a review is as follows (the process should be detailed in the protocol for the review).

1. Merge search results using reference management software, and remove duplicate records of the same report (see Chapter 6, Section 6.5).
2. **Examine titles and abstracts** to remove obviously irrelevant reports (authors should generally be over-inclusive at this stage).
3. Retrieve full text of the potentially relevant reports.

4. Link together multiple reports of the same study (see Section 7.2.2).
5. **Examine full-text reports** for compliance of studies with eligibility criteria.
6. Correspond with investigators, where appropriate, to clarify study eligibility (it may be appropriate to request further information, such as missing results, at the same time).
7. Make final decisions on study inclusion and proceed to data collection.

7.2.4 Implementation of the selection process

Decisions about which studies to include in a review are among the most influential decisions that are made in the review process. However, they involve judgement. To help ensure that these judgements are reproducible, it is desirable for more than one author to repeat parts of the process. In practice, the exact approach may vary from review to review, depending in part on the experience and expertise of the review authors.

Authors must first decide if more than one of them will assess the titles and abstracts of records retrieved from the search (step 2 in Section 7.2.3). Using at least two authors may reduce the possibility that relevant reports will be discarded (Edwards 2002). It is most important that the final selection of studies into the review is undertaken by more than one author (step 5 in Section 7.2.3).

Experts in a particular area frequently have pre-formed opinions that can bias their assessments of both the relevance and validity of articles (Cooper 1989, Oxman 1993). Thus, while it is important that at least one author is knowledgeable in the area under review, it may be an advantage to have a second author who is not a content expert. Some authors may decide that assessments of relevance should be made by people who are blind or masked to information about the article, such as the journal that published it, the authors, the institution, and the magnitude and direction of the results. They could attempt to do this by editing copies of the articles. However, this takes much time, and may not be warranted given the resources required and the uncertain benefit in terms of protecting against bias (Berlin 1997).

Disagreements about whether a study should be included can generally be resolved by discussion. Often the cause of disagreement is a simple oversight on the part of one of the review authors. When the disagreement is due to a difference in interpretation, this may require arbitration by another person. Occasionally, it will not be possible to resolve disagreements about whether to include a study without additional information. In these cases, authors may choose to categorize the study in their review as one that is awaiting assessment until the additional information is obtained from the study authors.

In summary, the methods section of both the protocol and the review should detail:

- whether more than one author examines each title and abstract to exclude obviously irrelevant reports;
- whether those who examine each full-text report to determine eligibility will do so independently (this should be done by at least two people);
- whether the decisions on the above are made by content area experts, methodologists, or both;
- whether the people assessing the relevance of studies know the names of the authors, institutions, journal of publication and results when they apply the eligibility criteria; and
- how disagreements are handled.

A single failed eligibility criterion is sufficient for a study to be excluded from a review. In practice, therefore, eligibility criteria for each study should be assessed in order of importance, so that the first

‘no’ response can be used as the primary reason for exclusion of the study, and the remaining criteria need not be assessed.

For most reviews it will be worthwhile to pilot test the eligibility criteria on a sample of reports (say ten to twelve papers, including ones that are thought to be definitely eligible, definitely not eligible and doubtful). The pilot test can be used to refine and clarify the eligibility criteria, train the people who will be applying them and ensure that the criteria can be applied consistently by more than one person.

7.2.5 Selecting ‘excluded studies’

A Cochrane review includes a list of excluded studies, detailing any studies that a reader might plausibly expect to see among the included studies. This covers all studies that may on the surface appear to meet the eligibility criteria but on further inspection do not, and also those that do not meet all of the criteria but are well known and likely to be thought relevant by some readers. By listing such studies as excluded and giving the primary reason for exclusion, the review authors can show that consideration has been given to these studies. The list of excluded studies should be as brief as possible. It should not list all of the reports that were identified by a comprehensive search. It should not list studies that obviously do not fulfil the entry criteria for the review as listed under ‘Types of studies’, ‘Types of participants’, and ‘Types of interventions’, and in particular should not list studies that are obviously not randomized if the review includes only randomized trials.

7.2.6 Measuring agreement

Formal measures of agreement are available to describe the extent to which assessments by multiple authors were the same (Orwin 1994). We describe in Section 7.2.6.1 how a kappa statistic may be calculated for measuring agreement between two authors making simple inclusion/exclusion decisions. Values of kappa between 0.40 and 0.59 have been considered to reflect fair agreement, between 0.60 and 0.74 to reflect good agreement and 0.75 or more to reflect excellent agreement (Orwin 1994).

It is not recommended that kappa statistics are calculated as standard in Cochrane reviews, although they can reveal problems, especially in the early stages of piloting. Comparison of a value of kappa with arbitrary cut-points is unlikely to convey the real impact of any disagreements on the review. For example, disagreement about the eligibility of a large, well conducted, study will have more substantial implications for the review than disagreement about a small study with risks of bias. The reasons for any disagreement should be explored. They may reveal the need to revisit eligibility criteria or coding schemes for data collection, and any resulting changes should be reported.

7.2.6.1 Calculations for a simple kappa statistic

Suppose the K studies are distributed according to numbers a to i as in Table 7.2.a. Then

$$\text{kappa} = \frac{P_o - P_e}{1 - P_e},$$

where

$$P_o = \frac{a + e + i}{K}$$

is the proportion of studies for which there was agreement, and

$$P_e = \frac{I_1 \times I_2 + E_1 \times E_2 + U_1 \times U_2}{K^2}$$

is the proportion of studies in which one would expect there to be agreement by chance alone. As an example, from the data in [Table 7.2.b](#),

$$P_o = \frac{5+7+3}{25} = 0.6,$$

$$P_e = \frac{12 \times 5 + 10 \times 10 + 3 \times 10}{25^2} = 0.304,$$

and so

$$\text{kappa} = \frac{0.6 - 0.304}{1 - 0.304} = 0.43.$$

Table 7.2.a: Data for calculation of a simple kappa statistic

		Review author 2			Total
		Include	Exclude	Unsure	
Review author 1	Include	<i>a</i>	<i>b</i>	<i>c</i>	I ₁
	Exclude	<i>d</i>	<i>e</i>	<i>f</i>	E ₁
	Unsure	<i>g</i>	<i>h</i>	<i>i</i>	U ₁
	Total	I ₂	E ₂	U ₂	K

Table 7.2.b: Example data for calculation of a simple kappa statistic

		Review author 2			Total
		Include	Exclude	Unsure	
Review author 1	Include	5	3	4	12
	Exclude	0	7	3	10
	Unsure	0	0	3	3
	Total	5	10	10	25

7.3 What data to collect

7.3.1 What are data?

For the purposes of this chapter, we define ‘data’ to be any information about (or deriving from) a study, including details of methods, participants, setting, context, interventions, outcomes, results, publications and investigators. Review authors should plan in advance what data will be required for their systematic review, and develop a strategy for obtaining them. The following sections review the types of information that should be sought, and these are summarized in [Table 7.3.a](#). Section [7.4](#) reviews the main sources of the data.

Table 7.3.a: Checklist of items to consider in data collection or data extraction

Items not in square brackets should normally be collected in all reviews; items in square brackets may be relevant to some reviews and not others.

<p>Source</p> <ul style="list-style-type: none"> • Study ID (created by review author); • Report ID (created by review author); • Review author ID (created by review author); • Citation and contact details; <p>Eligibility</p> <ul style="list-style-type: none"> • Confirm eligibility for review; • Reason for exclusion; <p>Methods</p> <ul style="list-style-type: none"> • Study design; • Total study duration; • Sequence generation*; • Allocation sequence concealment*; • Blinding*; • Other concerns about bias*; <p>Participants</p> <ul style="list-style-type: none"> • Total number; • Setting; • Diagnostic criteria; • Age; • Sex; • Country; • [Co-morbidity]; • [Socio-demographics]; • [Ethnicity]; • [Date of study]; <p>Interventions</p> <ul style="list-style-type: none"> • Total number of intervention groups; <p><i>For each intervention and comparison group of interest:</i></p> <ul style="list-style-type: none"> • Specific intervention; • Intervention details (sufficient for replication, if feasible); • [Integrity of intervention]; 	<p>Outcomes</p> <ul style="list-style-type: none"> • Outcomes and time points (i) collected; (ii) reported*; <p><i>For each outcome of interest:</i></p> <ul style="list-style-type: none"> • Outcome definition (with diagnostic criteria if relevant); • Unit of measurement (if relevant); • For scales: upper and lower limits, and whether high or low score is good; <p>Results</p> <ul style="list-style-type: none"> • Number of participants allocated to each intervention group; <p><i>For each outcome of interest:</i></p> <ul style="list-style-type: none"> • Sample size; • Missing participants*; • Summary data for each intervention group (e.g. 2×2 table for dichotomous data; means and SDs for continuous data); • [Estimate of effect with confidence interval; P value]; • [Subgroup analyses]; <p>Miscellaneous</p> <ul style="list-style-type: none"> • Funding source; • Key conclusions of the study authors; • Miscellaneous comments from the study authors; • References to other relevant studies; • Correspondence required; • Miscellaneous comments by the review authors.
--	---

*Full description required for standard items in the ‘Risk of bias’ tool (see Chapter 8, Section 8.5).

7.3.2 Methods and potential sources of bias

Different research methods can influence study outcomes by introducing different biases into results. Basic study design characteristics should be collected for presentation in the table of ‘Characteristics of included studies’, including whether the study is randomized, whether the study has a cluster or cross-over design, and the duration of the study. If the review includes non-randomized studies, appropriate features of the studies should be described (see Chapter 13, Section 13.4).

Information should also be collected to facilitate assessments of the risk of bias in each included study using the tool described in Chapter 8 (Section 8.5). The tool covers issues such as sequence generation, allocation sequence concealment, blinding, incomplete outcome data and selective outcome reporting. For each item in the tool, a description of what happened in the study is required, which may include verbatim quotes from study reports. Information for assessment of incomplete outcome data and selective outcome reporting may be most conveniently collected alongside information on outcomes and results. Chapter 8 (Section 8.3.4) discusses some issues in the collection of information for assessments of risk of bias.

7.3.3 Participants and setting

Details of participants and setting are collected primarily for presentation in the table of ‘Characteristics of included studies’. Some Cochrane Review Groups have developed standards regarding which characteristics should be collected. Typically, aspects that should be collected are those that could (or are believed to) affect presence or magnitude of an intervention effect and those that could help users assess applicability. For example, if the review authors suspect important differences in intervention effect between different socio-economic groups (examples of this are rare), this information should be collected. If intervention effects are thought constant over such groups, and if such information would not be useful to help apply results, it should not be collected.

Participant characteristics that are often useful for assessing applicability include age and sex, and summary information about these should always be collected if they are not obvious from the context. These are likely to be presented in different formats (e.g. ages as means or medians, with SDs or ranges; sex as percentages or counts; and either of these for the whole study or for each intervention group separately). Review authors should seek consistent quantities where possible, and decide whether it is more relevant to summarize characteristics for the study as a whole or broken down, for example, by intervention group. Other characteristics that are sometimes important include ethnicity, socio-demographic details (e.g. education level) and the presence of co-morbid conditions.

If the settings of studies may influence intervention effects or applicability, then information on these should be collected. Typical settings of healthcare intervention studies include acute care hospitals, emergency facilities, general practice, extended care facilities such as nursing homes, offices, schools and communities. Sometimes studies are conducted in different geographical regions with important differences in cultural characteristics that could affect delivery of an intervention and its outcomes. Timing of the study may be associated with important technology differences or trends over time. If such information is important for the interpretation of the review, it should be collected.

Diagnostic criteria that were used to define the condition of interest can be a particularly important source of diversity across studies and should be collected. For example, in a review of drug therapy for congestive heart failure, it is important to know how the definition and severity of heart failure was determined in each study (e.g. systolic or diastolic dysfunction, severe systolic dysfunction with ejection fractions below 20%). Similarly, in a review of antihypertensive therapy, it is important to describe baseline levels of blood pressure of participants.

7.3.4 Interventions

Details of all experimental and comparison interventions of relevance to the review should be collected, primarily for presentation in the ‘Characteristics of included studies’ table. Again, details are required for aspects that could affect presence or magnitude of effect, or that could help users assess applicability. Where feasible, information should be sought (and presented in the review) that is sufficient for replication of the interventions under study, including any co-interventions administered as part of the study.

For many clinical trials of many non-complex interventions such as drugs or physical interventions, routes of delivery (e.g., oral or intravenous delivery, surgical technique used), doses (e.g. amount or intensity of each treatment, frequency of delivery), timing (e.g. within 24 hours of diagnosis) and length of treatment may be relevant. For complex interventions, such as those that evaluate psychotherapy, behavioural and educational approaches or healthcare delivery strategies, it is important to collect information about the contents of the intervention, who delivered it, and the format and timing of delivery.

7.3.4.1 Integrity of interventions

The degree to which specified procedures or components of the intervention are implemented as planned can have important consequences for the findings from a study. We will describe this as intervention integrity; related terms include compliance and fidelity. The verification of intervention integrity may be particularly important in reviews of preventive interventions and complex interventions, which are often implemented in conditions that present numerous obstacles to idealized delivery (Dane 1998). Information about integrity can help determine whether unpromising results are due to a poorly conceptualized intervention or to an incomplete delivery of the prescribed components. Assessment of the implementation of the intervention also reveals important information about the feasibility of an intervention in real life settings, and in particular how likely it is that the intervention can and will be implemented as planned. If it is difficult to achieve full implementation in practice, the program will have low feasibility (Dusenbury 2003).

The following five aspects of integrity of preventive programs are described by Dane and Schneider (Dane 1998):

1. The extent to which specified intervention components were delivered as prescribed (*adherence*);
2. Number, length and frequency of implementation of intervention components (*exposure*);
3. Qualitative aspects of intervention delivery that are not directly related to the implementation of prescribed content, such as implementer enthusiasm, training of implementers, global estimates of session effectiveness, and leader attitude towards the intervention (*quality of delivery*);
4. Measures of participant response to the intervention, which may include indicators such as levels of participation and enthusiasm (*participant responsiveness*);
5. Safeguard checks against the diffusion of treatments, that is, to ensure that the subjects in each experimental group received only the planned interventions (*program differentiation*).

The integrity of an intervention may be monitored during a study using process measures, and feedback from such an evaluation may lead to evolution of the intervention itself. Process evaluation studies are characterized by a flexible approach to data collection and the use of numerous methods generating a range of different types of data. They may encompass both quantitative and qualitative methods. Process evaluations may be published separately from the outcome evaluation of the intervention. When it is considered important, review authors should aim to address whether the trial accounted for, or measured, key process factors and whether the trials that thoroughly addressed integrity showed a greater impact. Process evaluations can be a useful source of factors that potentially

influence the effectiveness of an intervention. Note, however, that measures of the success of blinding (e.g. in a placebo-controlled drug trial) may not be valuable (see Chapter 8, Section 8.11.1).

An example of a Cochrane review evaluating intervention integrity is provided by a review of smoking cessation in pregnancy (Lumley 2004). The authors found that process evaluation of the intervention occurred in only some trials, and in others the implementation was less than ideal (including some of the largest trials). The review highlighted how the transfer of an intervention from one setting to another may reduce its effectiveness if elements are changed or aspects of the materials are culturally inappropriate.

7.3.5 Outcome measures

Review authors should decide in advance whether they will collect information about all outcomes measured in a study, or about only those outcomes of (pre-specified) interest in the review. Because we recommend in Section 7.3.6 that results should only be collected for pre-specified outcomes, we also suggest that only the outcomes listed in the protocol be described in detail. However, a complete list of the names of all outcomes measured allows a more detailed assessment of the risk of bias due to selective outcome reporting (see Chapter 8, Section 8.13).

Information about outcomes that is likely to be important includes:

- definition (diagnostic method, name of scale, definition of threshold, type of behaviour);
- timing;
- unit of measurement (if relevant); and
- for scales: upper and lower limits, and whether a high or low score is favourable.

It may be useful to collect details of cited reports associated with scales, since these may contain further information about upper and lower limits, direction of benefit, typical averages and standard deviations, minimally important effect magnitudes, and information about validation.

Further considerations for economics outcomes are discussed in Chapter 15 (Section 15.4.2), and for patient-reported outcomes in Chapter 17.

7.3.5.1 Adverse outcomes

Collection of adverse effect outcomes can pose particular difficulties, discussed in detail in Chapter 14. Information falling under any of the terms ‘adverse effect’, ‘adverse drug reaction’, ‘side effect’, ‘toxic effect’, ‘adverse event’ and ‘complication’ may be considered as being potentially suitable for data extraction when evaluating the harmful effects of an intervention. Furthermore, it may be unclear whether an outcome should be classified as an adverse outcome (and the same outcome may be considered to be an adverse effect in some studies and not in others). No mention of adverse effects does not necessarily mean that no adverse effects occurred. It is usually safest to assume that they were not ascertained or not recorded. Quality of life measures are usually general measures that do not look specifically at particular adverse effects of the intervention. While quality of life scales can be used to gauge the overall well-being, they should not be regarded as substitutes for a detailed evaluation of safety and tolerability.

Precise definitions of adverse effect outcomes and their intensity should be recorded, since they may vary between studies. For example, in a review of aspirin and gastrointestinal haemorrhage, some trials simply reported gastrointestinal bleeds, while others reported specific categories of bleeding, such as haematemesis, melaena, and proctorrhagia (Derry 2000). The definition and reporting of severity of the haemorrhages (for example, major, severe, requiring hospital admission) also varied considerably among the trials (Zanchetti 1999). Moreover, a particular adverse effect may be

described or measured in different ways among the studies. For example, the terms ‘tiredness’, ‘fatigue’ or ‘lethargy’ might all be used in reporting of adverse effects. Study authors may also use different thresholds for ‘abnormal’ results (for example, hypokalaemia diagnosed at a serum potassium concentration of 3.0 mmol/l or 3.5 mmol/l).

7.3.6 Results

Results should be collected only for the outcomes specified to be of interest in the protocol. Results for other outcomes should not be extracted unless the protocol is modified to add them, and this modification should be reported in the review. However, review authors should be alert to the possibility of important, unexpected findings, particularly serious adverse effects.

Reports of studies often include several results for the same outcome. For example, different measurement scales might be used, results may be presented separately for different subgroups, and outcomes may have been measured at different points in time. Variation in the results can be very large, depending on which data are selected (Gøtzsche 2007), and protocols should be as specific as possible about which outcome measures, time-points and summary statistics (e.g. final values versus change from baseline) are to be collected. Refinements to the protocol may be needed to facilitate decisions on which results should be extracted.

Section 7.7 describes the numbers that will be required in order to perform meta-analysis. The unit of analysis (e.g. participant, cluster, body part, treatment period) should be recorded for each result if it is not obvious (see Chapter 9, Section 9.3). The type of outcome data determines the nature of the numbers that will be sought for each outcome. For example, for a dichotomous (‘yes’ or ‘no’) outcome, the number of participants and the number who experienced the outcome will be sought for each group. It is important to collect the sample size relevant to each result, although this is not always obvious. Drawing a flow diagram as recommended in the CONSORT Statement (Moher 2001) can help to determine the flow of participants through a study if one is not available in a published report (available from www.consort-statement.org).

The numbers required for meta-analysis are not always available, and sometimes other statistics can be collected and converted into the required format. For example, for a continuous outcome, it is usually most convenient to seek the number of participants, the mean and the standard deviation for each intervention group. These are often not available directly, especially the standard deviation, and alternative statistics enable calculation or estimation of the missing standard deviation (such as a standard error, a confidence interval, a test statistic (e.g. from a t-test or F-test) or a P value). Details are provided in Section 7.7. Further considerations for dealing with missing data are discussed in Chapter 16 (Section 16.1).

7.3.7 Other information to collect

Other information will be required from each report of a study, including the citation, contact details for the authors of the study and any other details of sources of additional information about it (for example an identifier for the study that would allow it to be found in a register of trials). Of particular importance in many areas is the funding source of the study, or potential conflicts of interest of the study authors. Some review authors will wish to collect information on study characteristics that bear on the quality of the study’s conduct but that are unlikely to lead directly to a risk of bias, such as whether ethical approval was obtained and whether a sample size calculation was performed.

We recommend that review authors collect the key conclusions of the included study as reported by its authors. It is not necessary to report these conclusions in the review, but they should be used to verify

results of analyses undertaken by the review authors, particularly in relation to the direction of effect. Further comments by the study authors, for example any explanations they provide for unexpected findings, might be noted. References to other studies that are cited in the study report may be useful, although review authors should be aware of the possibility of citation bias (see Chapter 10, Section 10.2.2.3).

7.4 Sources of data

7.4.1 Reports

Most Cochrane reviews obtain the majority of their data from study reports. Study reports include journal articles, books, dissertations, conference abstracts and web sites. Note, however, that these are highly variable in their reliability as well as their level of detail. For example, conference abstracts may present preliminary findings and confirmation of final results may be required. It is strongly recommended that a data collection form is used for extracting data from study reports (see Section 7.6).

7.4.2 Correspondence with investigators

Review authors will often find that they are unable to extract all of the information they seek from available reports, with regard to both the details of the study and the numerical results. In such circumstances, authors are recommended to contact the original investigators. Review authors will need to consider whether they will contact study authors with a request that is open-ended, seeks specific pieces of information, includes a data collection form (either uncompleted or partially completed), or seeks data at the level of individual participants. Contact details of study authors, if not available from the study reports, can often be obtained from an alternative recent publication, from university staff listings, or by a general search of the world wide web.

7.4.3 Individual patient data

Rather than extracting data from study publications, the original research data may be sought directly from the researchers responsible for each study. Individual patient data (IPD) reviews, in which data are provided on each of the participants in each of the studies, are the gold standard in terms of availability of data. IPD can be re-analysed centrally and, if appropriate, combined in meta-analyses. IPD reviews are addressed in detail in Chapter 18.

7.5 Data collection forms

7.5.1 Rationale for data collection forms

The data collection form is a bridge between what is reported by the original investigators (e.g in journal articles, abstracts, personal correspondence) and what is ultimately reported by the review authors. The data collection form serves several important functions (Meade 1997). First, the form is linked directly to the review question and criteria for assessing eligibility of studies, and provides a clear summary of these that can be applied to identified study reports. Second, the data collection form is the historical record of the multitude of decisions (and changes to decisions) that occur throughout the review process. Third, the form is the source of data for inclusion in an analysis.

Given the important functions of data collection forms, ample time and thought should be invested in their design. Because each review is different, data collection forms will vary across reviews. However, there are many similarities in the types of information that are important, and forms can be adapted from one review to the next. Although we use the term ‘data collection form’ in the singular, in practice it may be a series of forms used for different purposes: for example, a separate form for

assessing eligibility of studies for inclusion in the review to facilitate the quick determination of studies that should be excluded.

7.5.2 Electronic versus paper data collection forms

The decision between data collection using paper forms and data collection using electronic forms is largely down to review authors' preferences. Potential advantages of paper forms include:

- convenience or preference;
- data extraction can be undertaken almost anywhere;
- easier to create and implement (no need for computer programming or specialist software);
- provides a permanent record of all manipulations and modifications (providing these manipulations and modifications are not erased);
- simple comparison of forms completed by different review authors.

Potential advantages of electronic forms include:

- convenience or preference;
- combines data extraction and data entry into one step;
- forms may be programmed (e.g. using Microsoft Access) to 'lead' the author through the data collection process, for example, by posing questions that depend on answers to previous questions;
- data from reviews involving large numbers of studies are more easily stored, sorted and retrieved;
- allows simple conversions at the time of data extraction (e.g. standard deviations from standard errors; pounds to kilograms);
- rapid comparison of forms completed by different review authors; and
- environmental considerations.

Electronic systems have been developed that offer most of the advantages of both approaches (including the commercial SRS software: see www.trialstat.com). If review authors plan to develop their own electronic forms using spreadsheet or database programs, we recommend that (i) a paper form is designed first, and piloted using more than one author and several study reports; (ii) the data entry is structured in a logical manner with coding of responses as consistent and straightforward as possible; (iii) compatibility of output with RevMan is checked; and (iv) mechanisms are considered for recording, assessing and correcting data entry errors.

7.5.3 Design of a data collection form

When adapting or designing a data collection form, review authors should first consider how much information should be collected. Collecting too much information can lead to forms that are longer than original study reports, and can be very wasteful of time. Collection of too little information, or omission of key data, can lead to the need to return to study reports later in the review process.

Here are some tips for designing a data collection form, based on the informal collation of experiences from numerous review authors. The checklist in [Table 7.3.a](#) should also be consulted.

- Include the title of the review or a unique identifier. Data collection forms are adaptable across reviews and some authors participate in multiple reviews.
- Include a revision date or version number for the data collection form. Forms occasionally have to be revised, and this reduces the chances of using an outdated form by mistake.

- Record the name (or ID) of the person who is completing the form.
- Leave space for notes near the beginning of the form. This avoids placing notes, questions or reminders on the last page of the form where they are least likely to be noticed. Important notes may be entered into RevMan in the ‘Notes’ column of the ‘Characteristics of included studies’ table, or in the text of the review.
- Include a unique study ID as well as a unique report ID. This provides a link between multiple reports of the same study. Each included study must be given a study identifier that is used in RevMan (usually comprising the last name of first author and the year of the primary reference for the study).
- Include assessment (or verification) of eligibility of the study for the review near the beginning of the form. Then the early sections of the form can be used for the process of assessing eligibility. Reasons for exclusion of a study can readily be deduced from such assessments. For example, if only truly randomized trials are eligible, a query on the data collection form might be: ‘Randomized? Yes, No, Unclear’. If a study used alternate allocation, the answer to the query is ‘No’, and this information may be entered into the ‘Characteristics of excluded studies’ table as the reason for exclusion.
- Record the source of each key piece of information collected, including where it was found in a report (this can be done by highlighting the data in hard copy, for example) or if information was obtained from unpublished sources or personal communications. Any unpublished information that is used should be coded in the same way as published information.
- Use tick boxes or coded responses to save time.
- Include ‘not reported’ or ‘unclear’ options alongside any ‘yes’ or ‘no’ responses.
- Consider formatting sections for collecting results to match RevMan data tables. However, data collection forms should incorporate sufficient flexibility to allow for variation in how data are reported. It is strongly recommended that outcome data be collected in the format in which they were reported (and then transformed in a subsequent step).
- Always collect sample sizes when collecting outcome data, in addition to collecting initial (e.g. randomized) numbers. There may be different sample sizes for different outcomes because of attrition or exclusions.
- Leave plenty of space for notes.

7.5.4 Coding and explanations

It is important to provide detailed instructions to all authors who will use the data collection form (Stock 1994). These might be inserted adjacent or near to the data field on the form, directly in the cell that contains the data (e.g. as a comment in Microsoft Excel) or, if they are lengthy, might be provided on a separate page. Use of coding schemes is efficient and facilitates a systematic presentation of study characteristics in the review. Accurate coding is important, and the coding should not be so complicated that the data collector is easily confused or likely to make poor classifications. Checks should be made that coding schemes are being used consistently by different review authors.

7.6 Extracting data from reports

7.6.1 Introduction

In most Cochrane reviews, the primary source of information about each study is published reports of studies, usually in the form of journal articles. One of the most important and time-consuming parts of a systematic review is extracting data from such reports. The data collection form will usually be designed with data extraction in mind.

Electronic searches for text can provide a useful aid to locating information within a report, for example using search facilities in PDF viewers, internet browsers and word processing software. Text searching should not be considered a replacement for reading the report, however, since information may be presented using variable terminology.

7.6.2 Who should extract data?

It is strongly recommended that more than one person extract data from every report to minimize errors and reduce potential biases being introduced by review authors. As a minimum, information that involves subjective interpretation and information that is critical to the interpretation of results (e.g. outcome data) should be extracted independently by at least two people. In common with implementation of the selection process (Section 7.2.4), it is preferable that data extractors are from complementary disciplines, for example a methodologist and a topic area specialist. It is important that everyone involved in data extraction has practice using the form and, if the form was designed by someone else, receives appropriate training.

Evidence in support of duplicate data extraction comes from several indirect sources. One study observed that independent data extraction by two authors resulted in fewer errors than a data extraction by a single author followed by verification by a second (Buscemi 2006). A high prevalence of data extraction errors (errors in 20 out of 34 reviews) has been observed (Jones 2005). A further study of data extraction to compute standardized mean differences found that a minimum of seven out of 27 reviews had substantial errors (Gøtzsche 2007).

7.6.3 Preparing for data extraction

All forms should be pilot tested using a representative sample of the studies to be reviewed. This testing may identify data that are missing from the form, or likely to be superfluous. It is wise to draft entries for the 'Characteristics of included studies' table (Chapter 11, Section 11.2) and the 'Risk of bias' table (Chapter 8, Section 8.5) using these pilot reports. Users of the form may provide feedback that certain coding instructions are confusing or incomplete (e.g. a list of options may not cover all situations). A consensus between review authors may be required before the form is modified to avoid any misunderstandings or later disagreements. It might be necessary to repeat the pilot testing on a new set of reports if major changes are needed after the first testing.

Problems with the data collection form will occasionally surface after pilot testing has been completed and the form may need to be revised after data extraction has started. In fact, it is common for a data collection form to require modifications after it has been piloted. When changes are made to the form or coding instructions, it may be necessary to return to reports that have already undergone data extraction. In some situations, it may only be necessary to clarify coding instructions without modifying the actual data collection form.

Some have proposed that some information in a report, such as its authors, be blinded to the review author prior to data extraction and assessment of risk of bias (Jadad 1996); see also Chapter 9 (Section 8.3.4). However, blinding of review authors to aspects of study reports is not generally recommended for Cochrane reviews (Berlin 1997).

7.6.4 Extracting data from multiple reports of the same study

Studies are frequently reported in more than one publication (Tramèr 1997, von Elm 2004). However, the unit of interest in a Cochrane Intervention review is the study and not the report. Thus, information from multiple reports needs to be collated. It is not appropriate to discard any report of an included

study, since it may contain valuable information not included in the primary report. Review authors will need to decide between two strategies:

- Extract data from each report separately, then combine information across multiple data collection forms.
- Extract data from all reports directly into a single data collection form.

The choice of which strategy to use will depend on the nature of the reports and may vary across studies and across reports. For example, if a full journal article and multiple conference abstracts are available, it is likely that the majority of information will be obtained from the journal article, and completing a new data collection form for each conference abstract may be a waste of time.

Conversely, if there are two or more detailed journal articles, perhaps relating to different periods of follow-up, then it is likely to be easier to perform data extraction separately for these articles and collate information from the data collection forms afterwards.

Drawing flow diagrams for participants in a study, such as those recommended in the CONSORT Statement (Moher 2001), can be particularly helpful when collating information from multiple reports.

7.6.5 Reliability and reaching consensus

When more than one author extracts data from the same reports, there is potential for disagreement. An explicit procedure or decision rule should be identified in the protocol for identifying and resolving disagreements. Most often, the source of the disagreement is an error by one of the extractors and is easily resolved. Thus, discussion among the authors is a sensible first step. More rarely, a disagreement may require arbitration by another person. Any disagreements that cannot be resolved should be addressed by contacting the study authors; if this is unsuccessful, the disagreement should be reported in the review.

The presence and resolution of disagreements should be carefully recorded. Maintaining a copy of the data ‘as extracted’ (in addition to the consensus data) allows assessment of reliability of coding. Examples of ways in which this can be achieved include:

- Use one author’s (paper) data collection form and record changes after consensus in a different ink colour.
- Use a separate (paper) form for consensus data.
- Enter consensus data onto an electronic form.

Agreement of coded items can be quantified, for example using kappa statistics (Orwin 1994), although this is not routinely done in Cochrane reviews. A simple calculation for agreement between two authors is described in Section 7.2.6. If agreement is assessed, this should be done only for the most important data (e.g. key risk of bias assessments, or availability of key outcomes).

Informal consideration of the reliability of data extraction should be borne in mind throughout the review process, however. For example, if after reaching consensus on the first few studies, the authors note a frequent disagreement for specific data, then coding instructions may need modification. Furthermore, an author’s coding strategy may change over time, as the coding rules are forgotten, indicating a need for re-training and, possibly, some re-coding.

7.6.6 Summary

In summary, the methods section of both the protocol and the review should detail:

- the data categories that are to be collected;
- how verification of extracted data from each report will be verified (e.g. extraction by two review authors, independently);
- whether data extraction is undertaken by content area experts, methodologists, or both;
- piloting, training and existence of coding instructions for the data collection form;
- how data are extracted from multiple reports of the same study; and
- how disagreements are handled if more than one author extracts data from each report.

7.7 Extracting study results and converting to the desired format

7.7.1 Introduction

We now outline the data that need to be collected from each study for analyses of dichotomous outcomes, continuous outcomes and other types of outcome data. These types of data are discussed in Chapter 9 (Section 9.2). It is usually desirable to collect summary data separately for each intervention group and to enter these into RevMan, where effect estimates can be calculated. Sometimes the required data may be obtained only indirectly, and the relevant results may not be obvious. This section provides some useful tips and techniques to deal with some of these situations. If summary data cannot be obtained from each intervention group, effect estimates may be presented directly. In Section 7.7.7 we describe how standard errors of such effect estimates can be obtained from confidence intervals and P values.

7.7.2 Data extraction for dichotomous outcomes

Dichotomous data are described in Chapter 9, Section 9.2.2, and their meta-analysis is described in Chapter 9, Section 9.4.4. The only data required for a dichotomous outcome are the numbers in each of the two outcome categories in each of the intervention groups (the numbers needed to fill in the four boxes S_E , F_E , S_C , F_C in Chapter 9, Box 9.2.a). These are entered into RevMan as the numbers with the outcomes and the total sample sizes for the two groups. It is most reliable to collect dichotomous outcome data as the numbers who specifically did, and specifically did not, experience the outcome in each group. Although in theory this is equivalent to collecting the total numbers and the numbers experiencing the outcome, it is not always clear whether the reported total numbers are those on whom the outcome was measured. Occasionally the numbers incurring the event need to be derived from percentages (although it is not always clear which denominator to use, and rounded percentages may be compatible with more than one numerator).

Sometimes the numbers of participants and numbers of events are not available, but an effect estimate such as an odds ratio or risk ratio may be reported, for example in a conference abstract. Such data may be included in meta-analyses using the generic inverse variance method only if they are accompanied by measures of uncertainty such as a standard error, 95% confidence interval or an exact P value: see Section 7.7.7.

7.7.3 Data extraction for continuous outcomes

Continuous data are described in Chapter 9, Section 9.2.3, and their meta-analysis is discussed in Chapter 9, Section 9.4.5. To perform a meta-analysis of continuous data using either mean differences or standardized mean differences review authors should seek:

- mean value of the outcome measurements in each intervention group (M_E and M_C);
- standard deviation of the outcome measurements in each intervention group (SD_E and SD_C);

- number of participants on whom the outcome was measured in each intervention group (N_E and N_C).

Due to poor and variable reporting it may be difficult or impossible to obtain the necessary information from the data summaries presented. Studies vary in the statistics they use to summarize the average (sometimes using medians rather than means) and variation (sometimes using standard errors, confidence intervals, interquartile ranges and ranges rather than standard deviations). They also vary in the scale chosen to analyse the data (e.g. post-intervention measurements versus change from baseline; raw scale versus logarithmic scale).

A particularly misleading error is to misinterpret a standard error as a standard deviation. Unfortunately it is not always clear what is being reported and some intelligent reasoning, and comparison with other studies, may be required. Standard deviations and standard errors are occasionally confused in the reports of studies, and the terminology is used inconsistently.

When needed, missing information and clarification about the statistics presented should always be sought from the authors. However, for several of the measures of variation there is an approximate or direct algebraic relationship with the standard deviation, so it may be possible to obtain the required statistic even if it is not published in the paper, as explained in Sections 7.7.3.2 to 7.7.3.7. More details and examples are available elsewhere (Deeks 1997a, Deeks 1997b). Chapter 16 (Section 16.1.3) discusses options if standard deviations remain missing after attempts to obtain them.

Sometimes the numbers of participants, means and standard deviations are not available, but an effect estimate such as a mean difference or standardized mean difference may be reported, for example in a conference abstract. Such data may be included in meta-analyses using the generic inverse variance method only if they are accompanied by measures of uncertainty such as a standard error, 95% confidence interval or an exact P value. A suitable standard error from a confidence interval for a mean difference should be obtained using the early steps of the process described in Section 7.7.3.3. For standardized mean differences, see Section 7.7.7.

7.7.3.1 Post-intervention versus change from baseline

A common feature of continuous data is that a measurement used to assess the outcome of each participant is also measured at baseline, that is before interventions are administered. This gives rise to the possibility of using differences in **changes from baseline** (also called a **change score**) as the primary outcome. Review authors are advised not to focus on change from baseline unless this method of analysis was used in some of the study reports.

When addressing change from baseline, a single measurement is created for each participant, obtained either by subtracting the final measurement from the baseline measurement or by subtracting the baseline measurement from the final measurement. Analyses then proceed as for any other type of continuous outcome variable using the changes rather than the final measurements.

Commonly, studies in a review will have used a mixture of changes from baseline and final values. Some studies will report both; others will report only change scores or only final values. As explained in Chapter 9 (Section 9.4.5.2), both final values and change scores can sometimes be combined in the same analysis so this is not necessarily a problem. Authors may wish to extract data on both change from baseline and final value outcomes if the required means and standard deviations are available. A key problem associated with the choice of which analysis to use is the possibility of selective reporting

of the one with the more exaggerated results, and review authors should seek evidence of whether this may be the case (see Chapter 8, Section 8.13).

A final problem with extracting information on change from baseline measures is that often baseline and final measurements will be reported for different numbers of participants due to missed visits and study withdrawals. It may be difficult to identify the subset of participants who report both baseline and final value measurements for whom change scores can be computed.

7.7.3.2 Obtaining standard deviations from standard errors and confidence intervals for group means

A standard deviation can be obtained from the standard error of a mean by multiplying by the square root of the sample size:

$$SD = SE \times \sqrt{N}$$

When making this transformation, standard errors must be of means calculated from within an intervention group and not standard errors of the difference in means computed between intervention groups.

Confidence intervals for means can also be used to calculate standard deviations. Again, the following applies to confidence intervals for mean values calculated within an intervention group and not for estimates of differences between interventions (for these, see Section 7.7.3.3). Most confidence intervals are 95% confidence intervals. If the sample size is large (say bigger than 100 in each group), the 95% confidence interval is 3.92 standard errors wide ($3.92 = 2 \times 1.96$). The standard deviation for each group is obtained by dividing the length of the confidence interval by 3.92, and then multiplying by the square root of the sample size:

$$SD = \sqrt{N} \times (\text{upper limit} - \text{lower limit}) / 3.92$$

For 90% confidence intervals 3.92 should be replaced by 3.29, and for 99% confidence intervals it should be replaced by 5.15.

If the sample size is small (say less than 60 in each group) then confidence intervals should have been calculated using a value from a t distribution. The numbers 3.92, 3.29 and 5.15 need to be replaced with slightly larger numbers specific to the t distribution, which can be obtained from tables of the t distribution with degrees of freedom equal to the group sample size minus 1. Relevant details of the t distribution are available as appendices of many statistical textbooks, or using standard computer spreadsheet packages. For example the t value for a 95% confidence interval from a sample size of 25 can be obtained by typing `=tinv(1-0.95,25-1)` in a cell in a Microsoft Excel spreadsheet (the result is 2.0639). The divisor, 3.92, in the formula above would be replaced by $2 \times 2.0639 = 4.128$.

For moderate sample sizes (say between 60 and 100 in each group), either a t distribution or a standard normal distribution may have been used. Review authors should look for evidence of which one, and might use a t distribution if in doubt.

As an example, consider data presented as follows:

Group	Sample size	Mean	95% CI
Experimental intervention	25	32.1	(30.0, 34.2)

The confidence intervals should have been based on t distributions with 24 and 21 degrees of freedom respectively. The divisor for the experimental intervention group is 4.128, from above. The standard deviation for this group is $\sqrt{25 \times (34.2 - 30.0)/4.128} = 5.09$. Calculations for the control group are performed in a similar way.

It is important to check that the confidence interval is symmetrical about the mean (the distance between the lower limit and the mean is the same as the distance between the mean and the upper limit). If this is not the case, the confidence interval may have been calculated on transformed values (see Section 7.7.3.4).

7.7.3.3 Obtaining standard deviations from standard errors, confidence intervals, t values and P values for differences in means

Standard deviations can be obtained from standard errors, confidence intervals, t values or P values that relate to the differences between means in two groups. The difference in means itself (MD) is required in the calculations from the t value or the P value. An assumption that the standard deviations of outcome measurements are the same in both groups is required in all cases, and the standard deviation would then be used for both intervention groups. We describe first how a t value can be obtained from a P value, then how a standard error can be obtained from a t value or a confidence interval, and finally how a standard deviation is obtained from the standard error. Review authors may select the appropriate steps in this process according to what results are available to them. Related methods can be used to derive standard deviations from certain F statistics, since taking the square root of an F value may produce the same t value. Care is often required to ensure that an appropriate F value is used, and advice of a knowledgeable statistician is recommended.

From P value to t value

Where actual P values obtained from t tests are quoted, the corresponding t value may be obtained from a table of the t distribution. The degrees of freedom are given by $N_E + N_C - 2$, where N_E and N_C are the sample sizes in the experimental and control groups. We will illustrate with an example. Consider a trial of an experimental intervention ($N_E = 25$) versus a control intervention ($N_C = 22$), where the difference in means was $MD = 3.8$. It is noted that the P value for the comparison was $P = 0.008$, obtained using a two-sample t-test.

The t value that corresponds with a P value of 0.008 and $25+22-2=45$ degrees of freedom is $t = 2.78$. This can be obtained from a table of the t distribution with 45 degrees of freedom or a computer (for example, by entering `=tinv(0.008, 45)` into any cell in a Microsoft Excel spreadsheet).

Difficulties are encountered when levels of significance are reported (such as $P < 0.05$ or even $P = NS$ which usually implies $P > 0.05$) rather than exact P values. A conservative approach would be to take the P value at the upper limit (e.g. for $P < 0.05$ take $P = 0.05$, for $P < 0.01$ take $P = 0.01$ and for $P < 0.001$ take $P = 0.001$). However, this is not a solution for results which are reported as $P = NS$: see Section 7.7.3.7.

From t value to standard error

The t value is the ratio of the difference in means to the standard error of the difference in means. The standard error of the difference in means can therefore be obtained by dividing the difference in means (MD) by the t value:

$$SE = \frac{MD}{t} .$$

In the example, the standard error of the difference in means is obtained by dividing 3.8 by 2.78, which gives 1.37.

From confidence interval to standard error

If a 95% confidence interval is available for the difference in means, then the same standard error can be calculated as:

$$SE = (\text{upper limit} - \text{lower limit})/3.92$$

as long as the trial is large. For 90% confidence intervals 3.92 should be replaced by 3.29, and for 99% confidence intervals it should be replaced by 5.15. If the sample size is small then confidence intervals should have been calculated using a t distribution. The numbers 3.92, 3.29 and 5.15 need to be replaced with larger numbers specific to both the t distribution and the sample size, and can be obtained from tables of the t distribution with degrees of freedom equal to $N_E + N_C - 2$, where N_E and N_C are the sample sizes in the two groups. Relevant details of the t distribution are available as appendices of many statistical textbooks, or using standard computer spreadsheet packages. For example, the t value for a 95% confidence interval from a comparison of a sample size of 25 with a sample size of 22 can be obtained by typing `=tinv(1-0.95,25+22-2)` in a cell in a Microsoft Excel spreadsheet.

From standard error to standard deviation

The within-group standard deviation can be obtained from the standard error of the difference in means using the following formula:

$$SD = \frac{SE}{\sqrt{\frac{1}{N_E} + \frac{1}{N_C}}}$$

In the example,

$$SD = \frac{1.37}{\sqrt{\frac{1}{25} + \frac{1}{22}}} = 4.69 .$$

Note that this standard deviation is the average of the standard deviations of the experimental and control arms, and should be entered into RevMan twice (once for each intervention group).

7.7.3.4 Transformations and skewed data

Summary statistics may be presented after a transformation has been applied to the raw data. For example, means and standard deviations of logarithmic values may be available (or, equivalently, a geometric mean and its confidence interval). Such results should be collected, as they may be included in meta-analyses, or – with certain assumptions – may be transformed back to the raw scale

For example, a trial reported meningococcal antibody responses 12 months after vaccination with meningitis C vaccine and a control vaccine (MacLennan 2000), as geometric mean titres of 24 and 4.2 with 95% confidence intervals of 17 to 34 and 3.9 to 4.6 respectively. These summaries were obtained by finding the means and confidence intervals of the natural logs of the antibody responses (for vaccine 3.18: 95%CI (2.83 to 3.53), and control 1.44 (1.36 to 1.53)), and taking their exponentials (anti-logs). A meta-analysis may be performed on the scale of these natural log antibody responses.

Standard deviations of the log-transformed data may be derived from the latter pair of confidence intervals using methods described in Section 7.7.3.2. For further discussion of meta-analysis with skewed data, see Chapter 9 (Section 9.4.5.3).

7.7.3.5 Medians and interquartile ranges

The median is very similar to the mean when the distribution of the data is symmetrical, and so occasionally can be used directly in meta-analyses. However, means and medians can be very different from each other if the data are skewed, and medians are often reported *because* the data are skewed (see Chapter 9, Section 9.4.5.3).

Interquartile ranges describe where the central 50% of participants' outcomes lie. When sample sizes are large and the distribution of the outcome is similar to the normal distribution, the width of the interquartile range will be approximately 1.35 standard deviations. In other situations, and especially when the outcomes distribution is skewed, it is not possible to estimate a standard deviation from an interquartile range. Note that the use of interquartile ranges rather than standard deviations can often be taken as an indicator that the outcomes distribution is skewed.

7.7.3.6 Ranges

Ranges are very unstable and, unlike other measures of variation, increase when the sample size increases. They describe the extremes of observed outcomes rather than the average variation. Ranges should not be used to estimate standard deviations. One common approach has been to make use of the fact that, with normally distributed data, 95% of values will lie within $2 \times \text{SD}$ either side of the mean. The SD may therefore be estimated to be approximately one quarter of the typical range of data values. This method is not robust and we recommend that it should not be used.

7.7.3.7 No information on variability

If none of the above methods allow calculation of the standard deviations from the trial report (and the information is not available from the trialists) then, in order to perform a meta-analysis, an author may be forced to impute ('fill in') the missing data or to exclude the study from the meta-analysis: see Chapter 16 (Section 16.1.3). A narrative approach to synthesis may also be used. It is valuable to tabulate available results for all studies included in the systematic review, even if they cannot be included in a formal meta-analysis.

7.7.3.8 Combining groups

Sometimes it is desirable to combine two reported subgroups into a single group. This might be the case, for example, if a study presents sample sizes, means and standard deviations separately for males and females in each of the intervention groups. The formulae in Table 7.7.a can be used to combine numbers into a single sample size, mean and standard deviation for each intervention group (i.e. combining across males and females in this example). Note that the rather complex-looking formula for the SD produces the SD of outcome measurements *as if the combined group had never been divided into two*. An approximation to this standard deviation is obtained by using the usual pooled standard deviation, which provides a slight underestimate of the desired standard deviation.

These formulae are also appropriate for use in studies that compare more than two interventions, to combine two intervention groups into a single intervention group (see Chapter 16, Section 16.5). For example, 'Group 1' and 'Group 2' might refer to two alternative variants of an intervention to which participants were randomized.

If there are more than two groups to combine, the simplest strategy is to apply the above formula sequentially (i.e. combine group 1 and group 2 to create group ‘1+2’, then combine group ‘1+2’ and group 3 to create group ‘1+2+3’, and so on).

Table 7.7.a: Formulae for combining groups

	Group 1 (e.g. males)	Group 2 (e.g. females)	Combined groups
Sample size	N_1	N_2	$N_1 + N_2$
Mean	M_1	M_2	$\frac{N_1M_1 + N_2M_2}{N_1 + N_2}$
SD	SD_1	SD_2	$\sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2 + \frac{N_1N_2}{N_1 + N_2}(M_1^2 + M_2^2 - 2M_1M_2)}{N_1 + N_2 - 1}}$

7.7.4 Data extraction for ordinal outcomes

Ordinal data, when outcomes are categorized into several, ordered, categories, are described in Chapter 9, Section 9.2.4, and their meta-analysis is discussed in Chapter 9, Section 9.4.7. The data that need to be extracted for ordinal outcomes depend on whether the ordinal scale will be dichotomized for analysis (see Section 7.7.2), treated as a continuous outcome (see Section 7.7.3) or analysed directly as ordinal data. This decision, in turn, will be influenced by the way in which authors of the studies analysed their data. Thus it may be impossible to pre-specify whether data extraction will involve calculation of numbers of participants above and below a defined threshold, or mean values and standard deviations. In practice, it is wise to extract data in all forms in which they are given as it will not be clear which is the most common until all studies have been reviewed, and in some circumstances more than one form of analysis may justifiably be included in a review.

Where ordinal data are being dichotomized and there are several options for selecting a cut-point (or the choice of cut-point is arbitrary) it is sensible to plan from the outset to investigate the impact of choice of cut-point in a sensitivity analysis (see Chapter 9, Section 9.7). To do this it is necessary to collect the data that would be used for each alternative dichotomization. Hence it is preferable to record the numbers in each category of short ordinal scales to avoid having to extract data from a paper more than once. This approach of recording all categorizations is also sensible when studies use slightly different short ordinal scales, and it is not clear whether there will be a cut-point that is common across all the studies which can be used for dichotomization.

It is also necessary to record the numbers in each category of the ordinal scale for each intervention group if the proportional odds ratio method will be used (see Chapter 9, Section 9.2.4).

7.7.5 Data extraction for counts

Counts are described in Chapter 9, Section 9.2.5, and their meta-analysis is discussed in Chapter 9, Section 9.4.8. Data that are inherently counts may be analysed in several ways. The essential decision is whether to make the outcome of interest dichotomous, continuous, time-to-event or a rate. A common error is to treat counts directly as dichotomous data, using as sample sizes either the total

number of participants or the total number of, say, person-years of follow-up. Neither of these approaches is appropriate for an event that may occur more than once for each participant. This becomes obvious when the total number of events exceeds the sample size, leading to nonsensical results. Although it is preferable to decide how count data will be analysed in advance, the choice is often determined by the format of the available data, and thus cannot be decided until the majority of studies have been reviewed. Review authors should generally, therefore, extract count data in the form in which they are reported.

Sometimes detailed data on events and person-years at risk are not available, but results calculated from them are. For example, an estimate of a rate ratio or rate difference may be presented in a conference abstract. Such data may be included in meta-analyses only if they are accompanied by measures of uncertainty such as a 95% confidence interval: see Section 7.7.7. From this a standard error can be obtained and the generic inverse variance method used for meta-analysis.

7.7.5.1 Extracting counts as dichotomous data

To consider the outcome as a dichotomous outcome, the author must determine the number of participants in each intervention group, and the number of participants in each intervention group who experience *at least one event* (or some other appropriate criterion which classified all participants into one of two possible groups). Any time element in the data is lost through this approach, though it may be possible to create a series of dichotomous outcomes, for example at least one stroke during the first year of follow-up, at least one stroke during the first two years of follow-up, and so on. It may be difficult to derive such data from published reports.

7.7.5.2 Extracting counts as continuous data

To extract counts as continuous data (i.e. average number of events per patient), guidance in Section 7.7.3 should be followed, although particular attention should be paid to the likelihood that the data will be highly skewed.

7.7.5.3 Extracting counts as time-to-event data

For rare events that can happen more than once, an author may be faced with studies that treat the data as time-to-*first*-event. To extract counts as time-to-event data, guidance in Section 7.7.6 should be followed.

7.7.5.4 Extracting counts as rate data

If it is possible to extract the total number of events in each group, and the total amount of person-time at risk in each group, then count data can be analysed as rates (see Chapter 9, Section 9.4.8). Note that the total number of participants is not required for an analysis of rate data but should be recorded as part of the description of the study.

7.7.6 Data extraction for time-to-event outcomes

Time-to-event outcomes are described in Chapter 9, Section 9.2.6, and their meta-analysis is discussed in Chapter 9, Section 9.4.9. Meta-analysis of time-to-event data commonly involves obtaining individual patient data from the original investigators, re-analysing the data to obtain estimates of the log hazard ratio and its standard error, and then performing a meta-analysis (see Chapter 18). Conducting a meta-analysis using summary information from published papers or trial reports is often problematic as the most appropriate summary statistics are typically not presented. Two approaches can be used to obtain estimates of log hazard ratios and their standard errors, for inclusion in a meta-analysis using the generic inverse variance methods, regardless of whether individual patient data or

aggregate data are being used. For practical guidance, review authors should consult Tierney et al (Tierney 2007).

In the first approach an estimate of the log hazard ratio can be obtained from statistics computed during a log-rank analysis. Collaboration with a knowledgeable statistician is advised if this approach is followed. The log hazard ratio (experimental relative to control) is estimated by $(O - E)/V$, which has standard error $1/\sqrt{V}$, where O is the observed number of events on the experimental intervention, E is the log-rank expected number of events on the experimental intervention, $O - E$ is the log-rank statistic and V is the variance of the log-rank statistic. It is therefore necessary to obtain values of $O - E$ and V for each study.

These statistics are easily computed if individual patient data are available, and can sometimes be extracted from quoted statistics and survival curves (Parmar 1998, Williamson 2002). Alternatively, use can sometimes be made of aggregated data for each intervention group in each trial. For example, suppose that the data comprise the number of participants who have the event during the first year, second year, etc., and the number of participants who are event free and still being followed up at the end of each year. A log-rank analysis can be performed on these data, to provide the $O - E$ and V values, although careful thought needs to be given to the handling of censored times. Because of the coarse grouping the log hazard ratio is estimated only approximately, and in some reviews it has been referred to as a log odds ratio (Early Breast Cancer Trialists' Collaborative Group 1990). If the time intervals are large, a more appropriate approach is one based on interval-censored survival (Collett 1994).

The second approach can be used if trialists have analysed the data using a Cox proportional hazards model, or if a Cox model is fitted to individual patient data. Cox models produce direct estimates of the log hazard ratio and its standard error (so that a generic inverse variance meta-analysis can be performed). If the hazard ratio is quoted in a report together with a confidence interval or P value, estimates of standard error can be obtained as described in Section 7.7.7.

7.7.7 Data extraction for estimates of effects

7.7.7.1 Effect estimates and generic inverse variance meta-analysis

In some reviews, an overall estimate of effect will be sought from each study rather than summary data for each intervention group. This may be the case, for example, for non-randomized studies, cross-over trials, cluster-randomized trials, or studies with time-to-event outcomes. Meta-analysis can be applied to such effect estimates if their standard errors are available, using the generic inverse variance outcome type in RevMan (see Chapter 9, Section 9.4.3). When extracting data from non-randomized studies, and from some randomized studies, adjusted effect estimates may be available (e.g. adjusted odds ratios from logistic regression analyses, or adjusted rate ratios from Poisson regression analyses). The process of data extraction, and analysis using the generic inverse variance method, is the same as for unadjusted estimates, although the variables that have been adjusted for should be recorded (see Chapter 13, Section 13.6.2).

On occasion, summary data for each intervention group (for example, numbers of events and participants, or means and standard deviations) may be sought, but cannot be extracted. In such situations it may still be possible to include the study in a meta-analysis using the generic inverse variance method. A limitation of this approach is that estimates and standard errors of the same effect measure must be calculated for all the other studies in the same meta-analysis, even if they provide the summary data by intervention group. For example, if numbers in each outcome category by intervention group are known for some studies, but only odds ratios (ORs) are available for other

studies, then ORs would need to be calculated for the first set of studies and entered into RevMan under the generic inverse variance outcome type to enable meta-analysis with the second set of studies. RevMan may be used to calculate these ORs (entering them as dichotomous data), and the confidence intervals that RevMan presents may be transformed to standard errors using the methods that follow.

Estimates of an effect measure of interest may be presented along with a confidence interval or a P value. It is usually desirable to obtain a standard error from these numbers, so that the generic inverse variance outcome type in RevMan can be used to perform a meta-analysis. The procedure for obtaining a standard error depends on whether the effect measure is an absolute measure (e.g. mean difference, standardized mean difference, risk difference) or a ratio measure (e.g. odds ratio, risk ratio, hazard ratio, rate ratio). We describe these procedures in Section 7.7.7.2 and Section 7.7.7.3, respectively. However, for continuous outcome measures, the special cases of extracting results for a mean from one intervention arm, and extracting results for the difference between two means, are addressed in Section 7.7.3.

7.7.7.2 Obtaining standard errors from confidence intervals and P values: absolute (difference) measures

If a 95% confidence interval is available for an absolute measure of intervention effect (e.g. SMD, risk difference, rate difference), then the standard error can be calculated as

$$SE = (\text{upper limit} - \text{lower limit}) / 3.92.$$

For 90% confidence intervals divide by 3.29 rather than 3.92; for 99% confidence intervals divide by 5.15.

Where exact P values are quoted alongside estimates of intervention effect, it is possible to estimate standard errors. While all tests of statistical significance produce P values, different tests use different mathematical approaches to obtain a P value. The method here assumes P values have been obtained through a particularly simple approach of dividing the effect estimate by its standard error and comparing the result (denoted Z) with a standard normal distribution (statisticians often refer to this as a Wald test). Where significance tests have used other mathematical approaches the estimated standard errors may not coincide exactly with the true standard errors.

The first step is to obtain the Z value corresponding to the reported P value from a table of the standard normal distribution. A standard error may then be calculated as

$$SE = \text{intervention effect estimate} / Z.$$

As an example, suppose a conference abstract presents an estimate of a risk difference of 0.03 (P = 0.008). The Z value that corresponds to a P value of 0.008 is Z = 2.652. This can be obtained from a table of the standard normal distribution or a computer (for example, by entering =**abs(normsinv(0.008/2))** into any cell in a Microsoft Excel spreadsheet). The standard error of the risk difference is obtained by dividing the risk difference (0.03) by the Z value (2.652), which gives 0.011.

7.7.7.3 Obtaining standard errors from confidence intervals and P values: ratio measures

The process of obtaining standard errors for ratio measures is similar to that for absolute measures, but with an additional first step. Analyses of ratio measures are performed on the natural log scale (see Chapter 9, Section 9.2.7). For a ratio measure, such as a risk ratio, odds ratio or hazard ratio (which we will denote generically as RR here), first calculate

$$\text{lower limit} = \ln(\text{lower confidence limit given for RR})$$

upper limit = $\ln(\text{upper confidence limit given for RR})$

intervention effect estimate = $\ln\text{RR}$

Then the formulae in Section 7.7.7.2 can be used. Note that the standard error refers to the log of the ratio measure. When using the generic inverse variance method in RevMan, the data should be entered on the natural log scale, that is as $\ln\text{RR}$ and the standard error of $\ln\text{RR}$, as calculated here (see Chapter 9, Section 9.4.3).

7.8 Managing data

It is possible to collect data on paper data collection forms and to enter them directly into RevMan. Often, however, there will be a need or desire to manage data in intermediate computer software before entry into RevMan. A variety of software and data management programs may be helpful for this, including spreadsheet software (e.g. Microsoft Excel) and database programs (e.g. Microsoft Access). For example, tabulation of extracted information about studies in a spreadsheet can facilitate classifying of studies into comparisons and subgroups. Furthermore, statistical conversions, for example from standard errors to standard deviations, should ideally be undertaken with a computer rather than using a hand calculator, since it allows a permanent record to be kept of the original and calculated numbers as well as the actual calculations used.

7.9 Chapter information

Editors: Julian PT Higgins and Jonathan J Deeks.

This chapter should be cited as: Higgins JPT, Deeks JJ (editors). Chapter 7: Selecting studies and collecting data. 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.

Acknowledgements: This section builds on earlier versions of the *Handbook*. For details of previous authors and editors of the *Handbook*, see Chapter 1 (Section 1.4). Andrew Herxheimer, Nicki Jackson, Yoon Loke, Deirdre Price and Helen Thomas contributed text. Stephanie Taylor and Sonja Hood contributed suggestions for designing data collection forms. We are grateful to Judith Anzures, Mike Clarke, Miranda Cumpston and Peter Gøtzsche for helpful comments.

7.10 References

Berlin 1997

Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. *The Lancet* 1997; 350: 185-186.

Buscemi 2006

Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of Clinical Epidemiology* 2006; 59: 697-703.

Collett 1994

Collett D. *Modelling Survival Data in Medical Research*. London (UK): Chapman & Hall, 1994.

Cooper 1989

Cooper H, Ribble RG. Influences on the outcome of literature searches for integrative research reviews. *Knowledge* 1989; 10: 179-201.

Dane 1998

Dane AV, Schneider BH. Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review* 1998; 18: 23-45.

Deeks 1997a

Deeks J. Are you sure that's a standard deviation? (part 1). *Cochrane News* 1997; Issue No. 10: 11-12. (Available from www.cochrane.org/newslett/ccnewsbi.htm).

Deeks 1997b

Deeks J. Are you sure that's a standard deviation? (part 2). *Cochrane News* 1997; Issue No. 11: 11-12. (Available from www.cochrane.org/newslett/ccnewsbi.htm).

Derry 2000

Derry S, Loke YK. Risk of gastrointestinal haemorrhage with long term use of aspirin: meta-analysis. *BMJ* 2000; 321: 1183-1187.

Dusenbury 2003

Dusenbury L, Brannigan R, Falco M, Hansen WB. A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research* 2003; 18: 237-256.

Early Breast Cancer Trialists' Collaborative Group 1990

Early Breast Cancer Trialists' Collaborative Group. *Treatment of Early Breast Cancer. Volume 1: Worldwide Evidence 1985-1990*. Oxford (UK): Oxford University Press, 1990. (Available from www.ctsu.ox.ac.uk).

Edwards 2002

Edwards P, Clarke M, DiGuseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in Medicine* 2002; 21: 1635-1640.

Gøtzsche 2007

Gøtzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. *JAMA* 2007; 298: 430-437.

Jadad 1996

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay H. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials* 1996; 17: 1-12.

Jones 2005

Jones AP, Remington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *Journal of Clinical Epidemiology* 2005; 58: 741-742.

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.

MacLennan 2000

MacLennan JM, Shackley F, Heath PT, Deeks JJ, Flamank C, Herbert M, Griffiths H, Hatzmann E, Goilav C, Moxon ER. Safety, immunogenicity, and induction of immunologic memory by a serogroup C meningococcal conjugate vaccine in infants: A randomized controlled trial. *JAMA* 2000; 283: 2795-2801.

Meade 1997

Meade MO, Richardson WS. Selecting and appraising studies for a systematic review. *Annals of Internal Medicine* 1997; 127: 531-537.

Moher 2001

Moher D, Schulz KF, Altman DG. The CONSORT Statement: revised recommendations for

improving the quality of reports of parallel-group randomised trials. *The Lancet* 2001; 357: 1191-1194. (Available from www.consort-statement.org).

Orwin 1994

Orwin RG. Evaluating coding decisions. In: Cooper H, Hedges LV (editors). *The Handbook of Research Synthesis*. New York (NY): Russell Sage Foundation, 1994.

Oxman 1993

Oxman AD, Guyatt GH. The science of reviewing research. *Annals of the New York Academy of Sciences* 1993; 703: 125-133.

Parmar 1998

Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine* 1998; 17: 2815-2834.

Stock 1994

Stock WA. Systematic coding for research synthesis. In: Cooper H, Hedges LV (editors). *The Handbook of Research Synthesis*. New York (NY): Russell Sage Foundation, 1994.

Tierney 2007

Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007; 16.

Tramèr 1997

Tramèr MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997; 315: 635-640.

von Elm 2004

von Elm E, Poggia G, Walder B, Tramèr MR. Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *JAMA* 2004; 291: 974-980.

Williamson 2002

Williamson PR, Smith CT, Hutton JL, Marson AG. Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine* 2002; 21: 3337-3351.

Zanchetti 1999

Zanchetti A, Hansson L. Risk of major gastrointestinal bleeding with aspirin (Authors' reply). *The Lancet* 1999; 353: 149-150.