

Chapter 13: Including non-randomized studies

Authors: Barnaby C Reeves, Jonathan J Deeks, Julian PT Higgins and George A Wells on behalf of the Cochrane Non-Randomised Studies 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 13.8. 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.

Key points

- For some Cochrane reviews, the question of interest cannot be answered by randomized trials, and review authors may be justified in including non-randomized studies.
- Potential biases are likely to be greater for non-randomized studies compared with randomized trials, so results should always be interpreted with caution when they are included in reviews and meta-analyses. Particular concerns arise with respect to differences between people in different intervention groups (selection bias) and studies that do not explicitly report having had a protocol (reporting bias).
- We recommend that eligibility criteria, data collection and critical assessment of included studies place an emphasis on specific features of study design (e.g. which parts of the study were prospectively designed) rather than ‘labels’ for study designs (such as case-control versus cohort).
- Risk of bias in non-randomized studies can be assessed in a similar manner to that used for randomized trials, although more attention must be paid to the possibility of selection bias.
- Meta-analyses of non-randomized studies must consider how potential confounders are addressed, and consider the likelihood of increased heterogeneity resulting from residual confounding and from other biases that vary across studies.

13.1 Introduction

13.1.1 What this chapter is about

This chapter has been prepared by the Non-Randomised Studies Methods Group (NRSMG) of The Cochrane Collaboration (see [Box 13.8.a](#)). It is intended to support review authors who are considering including non-randomized studies in Cochrane reviews. **Non-randomized studies** (NRS) are defined

here as any quantitative study estimating the effectiveness of an intervention (harm or benefit) that does not use randomization to allocate units to comparison groups. This includes studies where allocation occurs in the course of usual treatment decisions or peoples' choices, i.e. studies usually called *observational*. There are many types of non-randomized intervention study, including cohort studies, case-control studies, controlled before-and-after studies, interrupted-time-series studies and controlled trials that use inappropriate randomization strategies (sometimes called quasi-randomized studies). [Box 13.1.a](#) summarizes some commonly-used study design labels for non-randomized studies. We explain in [Section 13.5.1](#) why we do not necessarily advise that these labels are used in Cochrane reviews.

This chapter aims to describe the particular challenges that arise if NRS are included in a Cochrane review, and is informed by theoretical or epidemiological considerations, empirical research, and discussions among members of the NRSMG. The chapter makes recommendations about what to do when it is possible to support the recommendations on the basis of evidence or established theory. When it is not possible to make any recommendations, the chapter aims to set out the pros and cons of alternative actions and to identify questions for further methodological research.

Review authors who are considering including NRS in a Cochrane review should not start with this chapter unless they are already familiar with the process of preparing a systematic review of randomized trials. The format and basic steps of a Cochrane review should be the same whether it includes only randomized trials or includes NRS. The reader is referred to Part 1 of the Handbook for a detailed description of these steps. Every step in carrying out a systematic review is more difficult when NRS are included and a review author should seek to include expert epidemiologists and methodologists in the review team. As an example of such collaboration, a review of NRS included nine authors, five of whom were methodologists (Siegfried 2003).

Box 13.1.a: Some types of NRS design used for evaluating the effects of interventions

Designs are distinguished below by labels in common use and descriptions are intentionally non-specific because the labels are interpreted in different ways with respect to details. The NRSMG does not advocate using these labels for reasons explained in [Section 13.5.1](#).

Non-randomized controlled trial.	An experimental study in which people are allocated to different interventions using methods that are not random.
Controlled before-and-after study.	A study in which observations are made before and after the implementation of an intervention, both in a group that receives the intervention and in a control group that does not.
Interrupted-time-series study.	A study that uses observations at multiple time points before and after an intervention (the 'interruption'). The design attempts to detect whether the intervention has had an effect significantly greater than any underlying trend over time.
Historically controlled study.	A study that compares a group of participants receiving an intervention with a similar group from the past who did not.
Cohort study.	A study in which a defined group of people (the cohort) is followed over time, to examine associations between different interventions received and subsequent outcomes. A 'prospective' cohort study recruits participants before any intervention and follows them into the future. A 'retrospective' cohort study identifies subjects from past records describing the interventions received and follows them from the time of those records.

Case-control study.	A study that compares people with a specific outcome of interest ('cases') with people from the same source population but without that outcome ('controls'), to examine the association between the outcome and prior exposure (e.g. having an intervention). This design is particularly useful when the outcome is rare.
Cross-sectional study.	A study that collects information on interventions (past or present) and current health outcomes, i.e. restricted to health states, for a group of people at a particular point in time, to examine associations between the outcomes and exposure to interventions.
Case series (uncontrolled longitudinal study).	Observations are made on a series of individuals, usually all receiving the same intervention, before and after an intervention but with no control group.

13.1.2 Why consider non-randomized studies?

The Cochrane Collaboration focuses particularly on systematic reviews of randomized trials because they are more likely to provide unbiased information than other study designs about the differential effects of alternative forms of health care. Reviews of NRS are only likely to be undertaken when the question of interest cannot be answered by a review of randomized trials. The NRSMG believes that review authors may be justified in including NRS which are moderately susceptible to bias. Broadly, the NRSMG considers that there are three main reasons for including NRS in a Cochrane review:

- a) To examine the case for undertaking a randomized trial by providing an explicit evaluation of the weaknesses of available NRS. The findings of a review of NRS may also be useful to inform the design of a subsequent randomized trial, e.g. through the identification of relevant subgroups.
- b) To provide evidence of the effects (benefit or harm) of interventions that cannot be randomized, or which are extremely unlikely to be studied in randomized trials. In these contexts, a disinterested (free from bias and partiality) review that systematically reports the findings and limitations of available NRS can be useful.
- c) To provide evidence of effects (benefit or harm) that cannot be adequately studied in randomized trials, such as long-term and rare outcomes, or outcomes that were not known to be important when existing, major randomized trials were conducted.

Three other reasons are often cited in support of systematic reviews of NRS but are poor justifications:

- d) Studying effects in patient groups not recruited to randomized trials (such as children, pregnant women, the elderly). Although it is important to consider whether the results of trials can be generalized to people who are excluded from them, it is not clear that this can be achieved by consideration of non-randomized studies. Regardless of whether estimates from NRS agree or disagree with those of randomized trials, there is always potential for bias in the results of the NRS, such that misleading conclusions are drawn.
- e) To supplement existing randomized trial evidence. Adding non-randomized to randomized evidence may change an imprecise but unbiased estimate into a precise but biased estimate, i.e. an exchange of undesirable uncertainty for unacceptable error.
- f) When an intervention effect is really large. Implicitly, this is a result-driven or *post hoc* justification, since the review (or some other synthesis of the evidence) needs to be undertaken to observe the likely size of the effects. Whilst it is easier to argue that large effects are less likely to be completely explained by bias than small effects (Glasziou 2007), for the practice of health care it is still important to obtain unbiased estimates of the magnitude of large effects to make clinical and economic decisions (Reeves 2006). Thus randomized trials are still needed for large effects (and they need not be large if the effects are truly large). There may be ethical opposition to

randomized trials of interventions already suspected to be associated with a large benefit as a result of a systematic review of NRS, making it difficult to randomize participants, and interventions postulated to have large effects may also be difficult to randomize for other reasons (e.g. surgery vs. no surgery). However, the justification for a systematic review of NRS in these circumstances should be classified as (b), i.e. interventions that are unlikely to be randomized, rather than as (f).

13.1.3 Key issues about the inclusion of non-randomized studies in a Cochrane review

Randomized trials are the preferred design for studying the effects of healthcare interventions because, in most circumstances, the randomized trial is the study design that is least likely to be biased. Any Cochrane review must consider the risk of bias in individual primary studies, including both the likely direction and magnitude of bias (see Chapter 8). A review that includes NRS also requires review authors to do this. The principle of considering risk of bias is exactly the same. However, potential biases are likely to be greater for NRS compared with randomized trials. Review authors need to consider (a) the weaknesses of the designs that have been used (such as noting their potential to ascertain causality), (b) the execution of the studies through a careful assessment of their risk of bias, especially (c) the potential for selection bias and confounding to which all NRS are suspect and (d) the potential for reporting biases, including selective reporting of outcomes.

Susceptibility to selection bias (understood in this *Handbook* to mean differences in the baseline characteristics of individuals in different intervention groups, rather than whether the selected sample is representative of the population) is widely regarded as the principal difference between randomized trials and NRS. Randomization with adequate allocation sequence concealment reduces the possibility of systematic selection bias in randomized trials so that differences in characteristics between groups can be attributed to chance. In NRS, allocation to groups depends on other factors, often unknown. Confounding occurs when selection bias gives rise to imbalances between intervention and control groups (or case and control groups in case-control studies) on prognostic factors, i.e. the distributions of the factors differ between groups *and* the factors are associated with outcome. Confounding can have two effects in a meta-analysis: (a) shifting the estimate of the intervention effect (systematic bias) and (b) increasing the variability of the observed effects, introducing excessive heterogeneity among studies (Deeks 2003). It is important to consider both of these possible effects (see Section 13.6.1). Section 13.5 provides a more detailed discussion of susceptibility to bias in NRS.

13.1.4 The importance of a protocol for a Cochrane review that includes non-randomized studies

Chapter 2 establishes the importance of writing a protocol for a Cochrane review before carrying out the review. As the methodological choices made during a review of NRS are complex and may affect the review findings, a protocol is even more important for a review that includes NRS. The rationale for doing a review that includes NRS (see Section 13.1.2) should be documented in the protocol. The protocol should include much more detail than for a review of randomized trials, pre-specifying key methodological decisions about the methods to be used and the analyses that are planned. The protocol needs to specify details that are not relevant for randomized trials (e.g. the methods planned to identify potential confounding factors and to assess the susceptibility of primary studies to confounding), as well as providing more detail about standard steps in the review process that are more difficult when including NRS (e.g. specification of eligibility criteria and the search strategy for identifying eligible studies).

The NRSMG recognizes that it may not be possible to pre-specify all decisions about the methods used in a review. Nevertheless, review authors should aim to make all decisions about the methods for

the review without reference to the findings of primary studies, and report methodological decisions that had to be made or modified after collecting data about the study findings.

13.1.5 Structure of subsequent sections in the chapter

Each of the sections in this chapter, which focus in turn on different steps of the review process, is structured in the same way. First, for a particular step, we summarize what is different when NRS (compared with randomized trials) are included in Cochrane reviews and, where applicable, describe conceptual issues that need to be considered. This first part includes relevant evidence, where there is some. Second, we summarize our guidance and, where available, describe existing resources that are available to support review authors.

13.2 Developing criteria for including non-randomized studies

13.2.1 What is different when including non-randomized studies?

13.2.1.1 Including both randomized and non-randomized studies

Review authors may want to include NRS in a review because only a small number of randomized trials can be identified, or because of perceived limitations of the randomized trials. In this chapter, we strongly recommend that review authors should not make any attempt to combine evidence from randomized trials and NRS. This recommendation means that criteria for included study designs should generally specify randomized or non-randomized studies when trying to evaluate the effect of an intervention on a particular outcome. (However, a single review might consist of ‘component’ reviews that include different study designs for different outcomes, for example, randomized trials for evaluating benefits and NRS to evaluate harms; see Chapter 14.) Alternatively, where randomized trial evidence is desired but unlikely to be available, eligibility criteria could reasonably be structured to say that NRS would only be included where randomized trials are found not to be available. In time, as such a review is updated, the NRS may be dropped when randomized trials become available. Where both randomized trials and NRS of an intervention exist and, for one or more of the reasons given in Section 13.1.2, both are included in the review, these should be presented separately; alternatively, if there is an adequate number of randomized trials, comments about relevant NRS can be included in the Discussion section of a review although this is rarely particularly helpful.

13.2.1.2 Evaluating benefits and harms

Cochrane reviews aim to quantify the effects of healthcare interventions, both beneficial and harmful, and both expected and unexpected. Most reviews estimate the expected benefits of an intervention that are assessed in randomized trials. Randomized trials may report some of the harms of an intervention, either those which were expected and which the trial was designed to assess, or those which were not expected but which were collected in the trial as part of standard monitoring of safety. However, many serious harms of an intervention are too rare or do not appear during the follow-up period of randomized trials, and therefore will not be reported. Therefore, one of the most important roles for reviews of NRS is to assess potential unexpected or rare harms of interventions (reason (c) in Section 13.1.1). Criteria for selecting important and relevant studies for evaluating rare or long-term adverse and unexpected effects are difficult to set. Although the relative strengths and weaknesses of different study designs are the same as for beneficial outcomes, the choice of study designs to include may depend on both the frequency of an outcome and its importance. For example, for some rare adverse outcomes only case series or case-control studies may be available. Study designs that are more susceptible to bias may be acceptable for evaluation of serious events in the absence of better evidence.

Confounding may be less of a threat to the validity of a review when researching rare harms or unexpected effects of interventions than when researching expected effects, since it is argued that ‘confounding by indication’ mainly influences treatment decisions with respect to outcomes about which the clinicians are primarily concerned. However, confounding can never be ruled out because the same features that are confounders for the expected effects may also be direct confounders for the unexpected effects, or be correlated with features that are confounders.

A related issue is the need to distinguish between *quantifying* and *detecting* an effect of an intervention. Quantifying the intended benefits of an intervention – maximizing the precision of the estimate and minimizing susceptibility to bias – is critical when weighing up the relative merits of alternative interventions for the same condition. A review should also try to quantify the harms of an intervention, minimizing susceptibility to bias as far as possible. However, if a review can establish beyond reasonable doubt that an intervention causes a particular harm, the precision and susceptibility to bias of the estimated effect may not be critical. In other words, the seriousness of the harm may outweigh any benefit from the intervention. This situation is more likely to occur when there are competing interventions for a condition.

13.2.1.3 Determining which types of non-randomized study to include

A randomized trial is a prospective, experimental study design specifically involving random allocation of participants to interventions. Although there are variations in randomized trial design (including random allocation of individuals, clusters or body parts; multi-arm trials, factorial trials and cross-over trials) they constitute a distinctive study category. By contrast, NRS cover a number of fundamentally different designs, several of which were originally conceived in the context of aetiological epidemiology. Some of these are summarized in [Box 13.1.a](#), although this is not an exhaustive list, and many studies combine ideas from different basic designs. As we discuss in [13.2.2](#) these labels are not consistently applied. The diversity of NRS designs raises two related questions. First, should all NRS designs of a particular effectiveness question be included in a review? Second, if review authors do not include all NRS designs, what criteria should be used to decide which study designs to include and which to exclude?

It is generally accepted that criteria should be set to limit the kinds of evidence included in a systematic review. The primary reason is that the risk of bias varies across studies. For this reason, many Cochrane reviews only include randomized trials (when available). For the same reason, it is argued that review authors should only include NRS that are least likely to be biased. It is not helpful to include primary studies in a review when the results of the studies are likely to be biased, even if there is no better evidence. This is because a misleading effect estimate may be more harmful to future patients than no estimate at all, particularly if the people using the evidence to make decisions are unaware of its limitations (Doll 1993, Peto 1995).

There is no agreement about the study design criteria that should be used to limit the inclusion of NRS in a Cochrane review. One strategy is to include only those study designs that will give reasonably valid effect estimates. Another strategy is to include the best available study designs which have been used to answer a question. The first strategy would mean that reviews are consistent and include the same types of NRS, but that some reviews include no studies at all. The second strategy leads to different reviews including different study designs according to what was available. For example, it might be entirely appropriate to use different criteria for inclusion when reviewing the harms, compared with the benefits, of an intervention. This approach is already evident in the *Cochrane Database of Systematic Reviews (CDSR)*, with editors of some Cochrane Review Groups (CRGs) restricting reviews to randomized trials only and other CRG editors allowing specific types of NRS to be included in reviews (typically in healthcare areas where randomized trials are infrequent).

Whichever point of view is adopted, criteria can only be chosen with respect to a hierarchy of primary study designs, ranked in order of risk of bias according to study design features. Existing ‘evidence hierarchies’ for studies of effectiveness (Eccles 1996, National Health and Medical Research Council 1999, Oxford Centre for Evidence-based Medicine 2001) appear to have arisen largely by applying hierarchies for aetiological research questions to effectiveness questions. For example, cohort studies are conventionally regarded as providing better evidence than case-control studies. It is not clear that this is always appropriate since aetiological hierarchies place more emphasis on establishing causality (e.g. dose-response relationship, exposure preceding outcome) than on valid quantification of the effect size. Also, study designs used for studying the effects of interventions can be very much more diverse and complex (Shadish 2002) and may not be easily assimilated into existing evidence hierarchies (see the array of designs in [Box 13.1.a](#), for example). Different designs are susceptible to different biases, and it is often unclear which biases have the greatest impact and how they vary between clinical situations.

13.2.1.4 Distinguishing between aetiology and effectiveness research questions

Including NRS in a Cochrane review allows, in principle, the inclusion of truly observational studies where the use of an intervention has occurred in the course of usual health care or daily life. For interventions that are not restricted to a medical setting, this may mean interventions that a study participant chooses to take, e.g. over-the-counter preparations. Including observational studies in a review also allows exposures to be studied that are not obviously ‘interventions’, e.g. nutritional choices, and other behaviours that may affect health. This introduces a ‘grey area’ between evidence about effectiveness and aetiology. It is important to distinguish carefully between different aetiological and effectiveness research questions related to a particular exposure. For example, nutritionists may be interested in the health-related effects of a diet that includes a minimum of five portions of fruit or vegetables per day (‘five-a-day’), an aetiological question. On the other hand, public health professionals may be interested in the health-related effects of interventions to promote a change in diet to include ‘five-a-day’, an effectiveness question. Because of other differences between studies relevant to these two kinds of question (e.g. duration of follow-up and outcomes investigated), studies addressing the former type of question are often perceived as being ‘better’ or ‘more relevant’ without acknowledging or realizing that they are addressing different research questions. In other instances the health intervention being evaluated in the NRS will have been undertaken for a purpose other than improving health. For example, a review of circumcision for preventing transmission of HIV included NRS where circumcision had been undertaken for cultural or religious reasons (Siegfried 2003), and it was unclear whether using the intervention for health purposes would have the same effect.

13.2.2 Guidance and resources available to support review authors

Review authors should first check with the editors of the CRG under which they propose to register their protocol whether there is a CRG-specific policy in place about the inclusion of NRS in a review. Authors should also discuss with the editors the extent of methodological advice available in the CRG since they are likely to require more support than with a review that includes randomized trials only, and attempt to recruit informed methodologists to their review team. Regrettably, the NRSMSG is not currently in a position to collaborate with authors on particular reviews, but encourages authors who include NRS in their reviews to feedback their experiences to the NRSMSG, particularly where their experiences support, or contradict, the experiences described in this chapter.

Review authors intending to review the adverse effects (harms) of an intervention should read Chapter 14, which has been prepared by the Adverse Effects Methods Group.

We recommend that review authors use explicit study design features (NB: not study design labels) when deciding which types of NRS to include in a review. Members of the NRSMSG have developed two lists that can be used for this purpose, although experience using them is limited. [Table 13.2.a](#) and

Table 13.2.b describe separate lists for individually-allocated and cluster-allocated studies. Sixteen (or fifteen) items are grouped under four headings:

1. Was there a comparison?
2. How were groups created?
3. Which parts of the study were prospective?
4. On which variables was comparability [between groups receiving different interventions] assessed?

The items are designed to characterize key features of studies which, on the basis of the experiences of NRSMG members and ‘first principles’ (rather than evidence), are suspected to define the major study design categories or to be associated with susceptibility to bias. The tables indicate which features are associated with different NRS designs, identified by labels that are more specific than those in Box 13.1.a. There is not total consensus about the use of these (column) labels. This disagreement does not mean that the row items are inappropriate or poorly described; the value of the lists depends on the agreement between review authors when classifying primary studies. We will also propose that these lists be used as checklists in the processes of data collection and as part of the critical assessment of the studies (Section 13.4.2 and Section 13.5.2). Instructions for using the items as checklists in Box 13.4.a provide further explanation of the terms.

A number of organizations are carrying out systematic reviews of NRS where there are no, or very few, randomized trials. Reviews are often commissioned on behalf of organizations responsible for issuing policy or guidance to healthcare professionals, e.g. the National Institute for Health and Clinical Excellence (NICE), the Canadian Agency for Drugs and Technologies in Health (CADTH), and carried out by teams of systematic reviewers in university departments of health sciences. In general, reviewers in these teams have sought to apply methods developed for systematic reviews of randomized trials to NRS. These groups include:

- Effective Practice and Organisation of Care (EPOC) Group (www.epoc.cochrane.org).
- The Centre for Reviews and Dissemination (www.york.ac.uk/inst/crd).
- EPPI centre, Institute of Education, University of London (eppi.ioe.ac.uk).
- The Effective Public Health Practice Project (EPHPP), Canadian Ministry of Health, Long-Term Care and the City of Hamilton, Public Health Services ([link to list of EPHPP reviews: old.hamilton.ca/phcs/ephpp](http://link.to.list.of.EPHPP.reviews:old.hamilton.ca/phcs/ephpp)).

CRGs and Cochrane review authors have tended to limit inclusion of NRS by study design or methodological quality, acknowledging that NRS design influences susceptibility to bias. For example, the EPOC CRG accepts protocols that include interrupted time series and controlled before-and-after studies, but not other NRS designs. Other reviews have limited inclusion to studies with ‘adequate methodological quality’ (Taggart 2001).

13.2.3 Summary

- Review authors should carefully justify their rationale for including NRS in their systematic review.
- Review authors should consult the editorial policy of the CRG under which they propose to register their protocol concerning inclusion of NRS. Authors should consider the extent of methodological advice available in the CRG and the methodological support they have in their team.
- Review authors should specify eligibility criteria based on what researchers did (i.e. important aspects of study design), as well as factors relating to the specific review question of interest (i.e.

intervention, population, health problem), to avoid ambiguity. We suggest that authors use the items in the NRSMG checklist, or a similar checklist, to do this.

- Review authors also need information about what researchers did in primary studies to categorize the studies identified. We suggest that authors use the NRSMG lists of study design features, or a similar tool, for these purposes, and record when important aspects of study design are unclear or not reported.
- Authors reviewing questions about the adverse effects (harms) of interventions should read Chapter 14.

Table 13.2.a: List of study design features (studies with allocation to interventions at the individual level)

	RCT	Q-RCT	NRCT	CBA	PCS	RCS	HCT	NCC	CC	XS	BA	CR/CS
<i>Was there a comparison:</i>												
Between two or more groups of participants receiving different interventions?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
Within the same group of participants over time?	P	P	N	Y	N	N	N	N	N	N	Y	N
<i>Were participants allocated to groups by:</i>												
Concealed randomization?	Y	N	N	N	N	N	N	N	N	N	na	na
Quasi-randomization?	N	Y	N	N	N	N	N	N	N	N	na	na
By other action of researchers?	N	N	Y	P	N	N	N	N	N	N	na	na
Time differences?	N	N	N	N	N	N	Y	N	N	N	na	na
Location differences?	N	N	P	P	P	P	P	na	na	na	na	na
Treatment decisions?	N	N	N	P	P	P	N	N	N	P	na	na
Participants' preferences?	N	N	N	P	P	P	N	N	N	P	na	na
On the basis of outcome?	N	N	N	N	N	N	N	Y	Y	P	na	na
Some other process? (specify)												
<i>Which parts of the study were prospective:</i>												
Identification of participants?	Y	Y	Y	P	Y	N	P*	Y	N	N	P	P
Assessment of baseline and allocation to intervention?	Y	Y	Y	P	Y	N	P*	Y	N	N	na	na
Assessment of outcomes?	Y	Y	Y	P	Y	P	P	Y	N	N	P	P
Generation of hypotheses?	Y	Y	Y	Y	Y	Y	Y	Y	P	P	P	na
<i>On what variables was comparability between groups assessed:</i>												
Potential confounders?	P	P	P	P	P	P	P	P	P	P	N	na
Baseline assessment of outcome variables?	P	P	P	Y	P	P	P	N	N	N	N	na

Y=Yes; P=Possibly; P*=Possible for one group only; N=No; na=not applicable. NB: Note that 'possibly' is used in the table to indicate cells where either 'Y' or 'N' may be the case. It should not be used as a response option when applying the checklist; if uncertain, the response should be 'can't tell' (see [Box 13.4.a](#)).

RCT=Randomized controlled trial; Q-RCT=Quasi-randomized controlled trial; NRCT=Non-randomized controlled trial; CBA=Controlled before-and-after study; PCS=Prospective cohort study; RCS=Retrospective cohort study; HCT=Historically controlled trial; NCC=Nested case-control study; CC=Case-control study; XS=Cross-sectional study; BA=Before-and-after comparison; CR/CS=Case report/Case series.

Table 13.2.b: List of study design features (studies with allocation to interventions at the group level)

	CIRCT	CIQ-RCT	CINRT	CITS	CChBA	ITS	ChBA	EcoXS
<i>Was there a comparison:</i>								
Between two or more groups of clusters receiving different interventions?	Y	Y	Y	Y	Y	Y	Y	Y
Within the same group of clusters over time?	P	P	N	Y	N	N	N	N
<i>Were clusters allocated to groups by:</i>								
Concealed randomization?	Y	N	N	N	N	N	N	N
Quasi-randomization?	N	Y	N	N	N	N	N	N
By other action of researchers?	N	N	Y	P	P	N	N	N
Time differences?	N	N	N	Y	Y	Y	Y	N
Location differences?	N	N	P	P	P	N	N	P
Policy/public health decisions?	Na	na	P	P	P	P	na	na
Cluster preferences?	Na	na	P	P	P	P	na	na
Some other process? (specify)								
<i>Which parts of the study were prospective:</i>								
Identification of participating clusters?	Y	Y	Y	P	P	P	P	N
Assessment of baseline and allocation to intervention?	Y	Y	Y	P	P	P	P	N
Assessment of outcomes?	Y	Y	Y	P	P	P	P	N
Generation of hypotheses?	Y	Y	Y	Y	Y	Y	Y	P
<i>On what variables was comparability between groups assessed:</i>								
Potential confounders?	P	P	P	P	P	P	P	P
Baseline assessment of outcome variables?	P	P	P	Y	Y	Y	Y	N

Note that ‘cluster’ refers to an entity (e.g. an organization), not necessarily to a group of participants; ‘group’ refers to one or more clusters; see Box 13.4.a.

Note that ‘possibly’ is used in the table to indicate cells where *either* ‘Y’ or ‘N’ may be the case. It should not be used as a response option when applying the checklist; if uncertain, ‘can’t tell’ should be used (see [Box 13.4.a](#)).

Y=Yes; P=Possibly; P*=Possible for one group only; N=No; NR=Not required. CIRCT=Cluster randomized controlled trial; CIQ-RCT=Cluster quasi-randomized controlled trial; CINRT=Cluster non-randomized controlled trial; CITS=Controlled interrupted time series (Shadish 2002); CChBA=Controlled cohort before-and-after study (Shadish 2002); ITS=Interrupted time series; ChBA=Cohort before-and-after study (Shadish 2002); EcoXS=Ecological cross-sectional study.

13.3 Searching for non-randomized studies

13.3.1 What is different when including non-randomized studies?

13.3.1.1 Comprehensiveness of search strategy

When a review aims to include randomized trials only, a key principle of searching for eligible studies is that review authors should try as hard as possible to identify all randomized trials of the review question that have ever been started. Therefore, review authors are recommended to search trial registers, conference abstracts, grey literature, etc, as well as standard bibliographic databases such as MEDLINE, PUBMED, EMBASE (see Chapter 6). It is argued that a systematic review needs to search comprehensively in order to avoid publication biases. It is easy to argue that authors of a review that includes NRS should do the same (Petticrew 2001). However, it is important to set out the premises underpinning the original rationale for a comprehensive search and to consider very carefully whether they apply to reviews of NRS. The premises are:

- a) A finite population exists of randomized trials that investigate the review question.
- b) All randomized trials in this population can be identified through a search that is sufficiently comprehensive because randomized trials are relatively easily identified, registers of them are available, and they are difficult to do without funding and ethics approval, which also create an 'audit trail' (Chan 2004).
- c) All randomized trials in this population, if well conducted, provide valuable information.
- d) Ease of access to information about these randomized trials is related to their findings, so that the most readily identified trials may be a biased subset. This is publication bias: studies with statistically significant and favourable findings are more likely to be published in accessible places (see Chapter 10, Section 10.2). Because smaller studies are less likely to produce such findings, failure to identify all studies may result in funnel plot asymmetry. An unbiased answer can in theory be reached by identifying all randomized trials, i.e. by a comprehensive search to uncover the small, non-significant or unfavourable studies. Smaller studies may also suffer differentially from other biases, giving rise to an alternative cause of funnel plot asymmetry. The risks of these biases are reasonably well understood and may be assessed (Chapter 10, Section 10.4).

It is not clear that these premises apply equally to NRS.

Section 13.2.1.3 points out that NRS include diverse designs, and that there is difficulty in categorizing them. Even if review authors are able to set specific study design criteria against which potential NRS should be assessed for inclusion, many of the potentially eligible NRS will report insufficient information to allow them to be classified.

There is a further problem in defining exactly when a NRS comes into existence. For example, is a cohort study that has collected data on the interventions and outcome of interest, but that has not examined their association, an eligible NRS? Is computer output in a filing cabinet that includes a calculated odds ratio for the relevant association an eligible NRS? Consequently, it is difficult to define a 'finite population of NRS' for a particular review question. Some NRS that have been done may not be traceable at all, i.e. they are not to be found even in the proverbial 'bottom drawer'.

Notwithstanding the problems in defining what constitutes an eligible NRS, the actual identification of NRS provides important challenges. This is not just to do with poor reporting but also to do with:

- the absence of registers of NRS;
- poor indexing of important study design characteristics, etc;
- NRS not always requiring ethical approval (at least in the past);
- NRS not always having a research sponsor or funder; and

- NRS not always having been executed according to a pre-specified protocol.

There is no evidence that reporting biases affect randomized trials and NRS differentially. However, it is difficult to believe that reporting biases could affect NRS *less* than randomized trials, given the increasing number of features associated with carrying out and reporting randomized trials that act to prevent reporting biases which are frequently absent in NRS (pre-specified protocol, ethical approval including progress and final reports, the CONSORT statement (Moher 2001), trial registers and indexing of publication type in bibliographic databases). Unlike the situation for randomized trials, the likely magnitude and determinants of publication bias are not known.

The benefits of comprehensive searching for NRS are unclear, and this is a topic that requires further research. It is possible that the studies which are the hardest to find may be the most biased, if being hard to identify relates to poor design and small size. With reviews of randomized trials, comprehensive searching offers potential protection against bias because a defined population of eligible studies exists, so small studies with non-significant findings should, ultimately, be identified. With reviews of NRS, even if a *theoretical* finite population of eligible studies can be defined, one does not have similar confidence that missing studies with non-significant findings can be identified.

13.3.1.2 Identifying NRS in searches

It is easy to design a search strategy that identifies all evidence about an intervention by creating search strings for the population and disease characteristics, the intervention, and possibly the comparator. When a review aims to include randomized trials only, various approaches are available to restrict the search strategy to randomized trials (see Chapter 6):

- a) Search for previous reviews of the review question.
- b) Use resources, such as CENTRAL or CRG-specific registers, that are 'rich' in randomized trials.
- c) Use methodological filters and indexing fields, such as publication type in MEDLINE, to limit searches to studies that are likely to be randomized trials.
- d) Search trial registers.

To restrict the search to particular non-randomized study designs is more difficult. Of the above approaches, only (a) and (b) are likely to be at all helpful. Review authors should certainly search CRG-specific registers for potentially relevant NRS. Some CRGs (e.g. the EPOC Group) include particular types of NRS in CRG-specific registers (authors should check with their CRG). The process of identifying studies for inclusion in CENTRAL means that some, but not all, NRS are included, so searches of this database will not be comprehensive, even for studies that use a particular design. There are no databases of NRS similar to CENTRAL.

As discussed in Section 13.2.1.3, study design labels are not used consistently by authors and are not indexed reliably by bibliographic databases. Strategy (c) is unlikely to be helpful because study design labels other than randomized trial are not reliably indexed by bibliographic databases and are often used inconsistently by authors of primary studies. Some review authors have tried to develop and 'validate' search strategies for NRS (Wieland 2005, Fraser 2006, Furlan 2006). Authors have also sought to optimize search strategies for adverse effects (see Chapter 14, Section 14.5) (Golder 2006b, Golder 2006c). Because of the time consuming nature of systematic reviews that include NRS, attempts to develop search strategies for NRS have not investigated large numbers of review questions. Therefore, review authors should be cautious about assuming that previous strategies can necessarily be applied to new topics.

13.3.1.3 Reviewing citations and abstracts

Randomized trials can usually be identified in search results simply from the titles and abstracts, particularly since the implementation of reporting standards. Unfortunately, the design details of NRS that are required to assess eligibility are often not described in titles or abstracts and require access to the full study report.

13.3.2 Guidance and resources available to support review authors

The NRSMG does not recommend limiting search strategies by index terms relating to study design. However, review authors may wish to contact researchers who have reported some success in developing efficient search strategies for NRS (see Section 13.3.1) and other review authors who have carried out Cochrane reviews (or other systematic reviews) of NRS for review questions similar to their own.

When searching for NRS, review authors are recommended to search for studies investigating all effects of an intervention and not to limit search strategies to specific outcomes (Chapter 6). When searching for NRS of specific rare or long-term (usually adverse or unintended) outcomes of an intervention, including free text and MeSH terms for specific outcomes in the search strategy may be justified. Members of the Adverse Effects Methods Group have experience of doing this (see Chapter 14, Section 14.5).

Review authors should check with their CRG editors whether the CRG-specific register includes studies with particular study design features and should seek the advice of information retrieval experts within the CRG and in the Information Retrieval Methods Group (see Chapter 6, Box 6.7.a).

13.3.3 Summary

- To identify studies of the expected beneficial effects of interventions, search strategies should include search strings for the intervention and the population and health problem of interest. Currently, there are no recommended methods for restricting search strategies by study design.
- Review authors searching for evidence relating to ‘suspected’ adverse effects may want to consider searching for specific outcomes (i.e. adverse effects) of interest. This approach obviously cannot be used for more general searches of possible adverse effects of an intervention (see Chapter 14, Section 14.5).
- Exhaustive searching, which is recommended for randomized trials, may not be justified when reviewing NRS. However, there is no research at present to guide authors about this important issue.

13.4 Selecting studies and collecting data

13.4.1 What is different when including non-randomized studies?

Search results often contain large numbers of irrelevant citations and abstracts often do not provide adequate detail about NRS design (which are likely to be required to judge eligibility). Therefore, unlike the situation when reviewing randomized trials, very many full reports of studies may need to be obtained and read in order to select eligible studies.

Review authors need to collect all of the data required for a systematic review of randomized trials (see Chapter 7) and also data to describe (a) the features of the design of a primary study (see Section 13.2.2), (b) confounding factors considered and the methods used to control for confounding (see

Section 13.1.3), (c) aspects of risk of bias specific for NRS (see Section 13.5.1) and (d) the results (see Section 13.6.1).

Review authors normally collect ‘raw’ information about the results when reviewing randomized trials, e.g. for a dichotomous outcome, the total number of participants and the number experiencing the outcome in each group. If participants are randomized to groups, a comparison of these raw data is assumed to be unbiased. For a NRS, a comparison of the same raw data is ‘unadjusted’ and susceptible to confounding. Authors usually also report an ‘adjusted’ comparison estimated from a regression model which cannot be summarized in the same way. Review authors should still record the sample size recruited to each group, and the number analysed and the number of events, but also need to document any adjusted effect estimates and their standard errors or confidence intervals. These data can be used to display adjusted effect estimates and their precision in forest plots and, if appropriate, to pool data across studies.

Anecdotally, the experience of review authors is that NRS are poorly reported so that the required information is difficult to find, and different review authors may extract different information from the same paper. Data collection forms may need to be customized to the research question being investigated. Because of the diversity of potentially eligible studies and the ways in which they are reported, developing the data collection form can require several iterations in the course of reviewing a sample of primary studies. It is almost impossible to finalize these forms in advance.

Results in NRS may be presented using different measures of effect and uncertainty or statistical significance depending on the reporting style and analyses undertaken. Expert statistical advice may assist review authors to transform or ‘work back’ from the information provided in a paper to obtain a consistent effect measure across studies. Data collection sheets need to be able to handle the different kinds of information about study findings that authors may encounter.

13.4.2 Guidance and resources available to support review authors

As well as providing information for deciding about eligibility, the questions in [Table 13.2.a](#) and [Table 13.2.b](#) represent a convenient checklist for collecting relevant data from NRS about study design features. In using this checklist to collect information about the studies and to decide on eligibility, the intention should be to document what researchers did in the primary studies, rather than what researchers called their studies or think they did. Items should be recorded as ‘Yes’, ‘No’ or ‘Can’t tell’. [Box 13.4.a](#) provides guidance on using these tables as checklists.

Data collection forms have been developed for use in NRSMG workshops to illustrate data extraction from NRS. These include: the study design checklist, templates for collecting information about confounding factors, their comparability at baseline, methods used to adjust for confounding, and effect estimates. These resources (available from the *Handbook* resource web site, www.cochrane.org/resources/handbook) can be used as a guide to the types of data collection forms that review authors will need. However, review authors will need to customize the forms carefully for the review question being studied.

Box 13.4.a: User guide for data collection/study assessment using checklist in [Table 13.2.a](#) or [Table 13.2.a](#)

Note: Users need to be very clear about the way in which the terms ‘group’ and ‘cluster’ are used in these tables. [Table 13.2.a](#) only refers to groups, which is used in its conventional sense to mean a number of individual participants. With the exception of allocation on the basis of outcome, ‘group’ can be interpreted synonymously with ‘intervention group’. [Table 13.2.b](#) refers to both clusters and

groups. In this table, 'clusters' are typically an organizational entity such as a family health practice, or administrative area, not an individual. As in Table 13.2.a, 'group' is synonymous with 'intervention group' and is used to describe a collection of allocated units, but in Table 13.2.b these units are clusters rather than individuals. Furthermore, although individuals are nested in clusters, a cluster does not necessarily represent a fixed collection of individuals. For instance, in cluster-allocated studies, clusters are often studied at two or more time-points (periods) with different collections of individuals contributing to the data collected at each time-point.

Was there a comparison?

Typically, researchers compare two or more groups that receive different interventions; the groups may be studied over the same time period, or over different time periods (see below). Sometimes researchers compare outcomes in just one group but at two time-points. It is also possible that researchers may have done both, i.e. studying two or more groups and measuring outcomes at more than one time-point.

Were participants/clusters allocated to groups by?

These items aim to describe how groups were formed. None will apply if the study does not compare two or more groups of subjects. The information is often not reported or is difficult to find in a paper. The items provided cover the main ways in which groups may be formed. More than one option may apply to a single study, although some options are mutually exclusive (i.e. a study is either randomized or not).

Randomization: Allocation was carried out on the basis of truly random sequence. Such studies are covered by the standard guidance elsewhere in this *Handbook*. Check carefully whether allocation was adequately concealed until subjects were definitively recruited.

Quasi-randomization: Allocation was done on the basis of a pseudo-random sequence, e.g. odd/even hospital number or date of birth, alternation. Note: when such methods are used, the problem is that allocation is rarely concealed. These studies are often included in systematic reviews that only include randomized trials, using assessment of the risk of bias to distinguish them from properly randomized trials.

By other action of researchers: This is a catch-all category and further details should be noted if the researchers report them. Allocation happened as the result of some decision or system applied by the researchers. For example, subjects managed in particular 'units' of provision (e.g. wards, general practices) were 'chosen' to receive the intervention and subjects managed in other units to receive the control intervention.

Time differences: Recruitment to groups did not occur contemporaneously. For example, in a historically controlled study subjects in the control group are typically recruited earlier in time than subjects in the intervention group; the intervention is then introduced and subjects receiving the intervention are recruited. Both groups are usually recruited in the same setting. If the design was under the control of the researchers, both this option and 'other action of researchers' must be ticked for a single study. If the design 'came about' by the introduction of a new intervention, both this option and 'treatment decisions' must be ticked for a single study.

Location differences: Two or more groups in different geographic areas were compared, and the choice of which area(s) received the intervention and control interventions was not made randomly. So, both this option and 'other action of researchers' could be ticked for a single study.

Treatment decisions: Intervention and control groups were formed by naturally occurring variation in treatment decisions. This option is intended to reflect treatment decisions taken mainly by the clinicians responsible; the following option is intended to reflect treatment decisions made mainly on the basis of subjects' preferences. If treatment preferences are uniform for particular provider 'units', or switch over time, both this option and 'location' or 'time' differences should be ticked.

Patient preferences: Intervention and control groups were formed by naturally occurring variation in patients' preferences. This option is intended to reflect treatment decisions made mainly on the basis of subjects' preferences; the previous option is intended to reflect treatment decisions taken mainly by the clinicians responsible.

On the basis of outcome: A group of people who experienced a particular outcome of interest were compared with a group of people who did not, i.e. a case-control study. Note: this option should be ticked for papers that report analyses of *multiple risk factors for a particular outcome* in a large series of subjects, i.e. in which the total study population is divided into those who experienced the outcome and those who did not. These studies are much closer to nested case-control studies than

cohort studies, even when longitudinal data are collected prospectively for consecutive patients.

Additional options for cluster-allocated studies

Location differences: see above.

Policy/public health decisions: Intervention and control groups were formed by decisions made by people with the responsibility for implementing policies about public health or service provision. Where such decisions are coincident with clusters, or where such people are the researchers themselves, this item overlaps with 'other action of researchers' and 'cluster preferences'.

Cluster preferences: Intervention and control groups were formed by naturally occurring variation in the preferences of clusters, e.g. preferences made collectively or individually at the level of the cluster entity.

Which parts of the study were prospective?

These items aim to describe which parts of the study were conducted prospectively. In a randomized controlled trial, all four of these items would be prospective. For NRS it is also possible that all four are prospective, although inadequate detail may be presented to discern this, particularly for generation of hypotheses. In some cohort studies, participants may be identified, and have been allocated to treatment retrospectively, but outcomes are ascertained prospectively.

On what variables was comparability of groups assessed?

These questions should identify 'before-and-after' studies. Baseline assessment of outcome variables is particularly useful when outcomes are measured on continuous scales, e.g. health status or quality of life.

Response options

Try to use only 'Yes', 'No' and 'Can't tell' response options. 'N/a' should be used if a study does not report a comparison between groups.

13.4.3 Summary

- Reviewing citations and abstracts identified by searching will be very time consuming, first because of the volume of citations identified and second because the information needed to judge eligibility may not be reported in the title or abstract.
- Collect data as for a randomized trial (i.e. details of study, study population, sample size recruited, sample size analysed, etc).
- Collect data about what researchers did (NRSMG checklist, or similar).
- Collect data about the confounding factors considered.
- Collect data about the comparability of groups on confounding factors considered.
- Collect data about the methods used to control for confounding.
- Collect data about multiple effect estimates (both unadjusted and adjusted estimates, if available).

13.5 Assessing risk of bias in non-randomized studies

13.5.1 What is different when including non-randomized studies?

13.5.1.1 Sources of bias in non-randomized studies

Bias may be present in findings from NRS in many of the same ways as in poorly designed or conducted randomized trials (see Chapter 8). For example, numbers of exclusions in NRS are frequently unclear, intervention and outcome assessment are often not conducted according to standardized protocols, and outcomes may not be assessed blind. The biases caused by these problems are likely to be similar to those that occur in randomized trials, and review authors should be familiar

with Chapter 8 that describes these issues. None of these problems are any less difficult to overcome in a well-planned non-randomized prospective study than in a randomized trial.

In NRS, use of allocation mechanisms other than concealed randomization means that groups are unlikely to be comparable. These potential systematic differences between characteristics of participants in different intervention ‘groups’ are likely to be the issue of key concern in most NRS, and we refer to this as selection bias. When selection bias produces imbalances in prognostic factors associated with the outcome of interest then ‘confounding’ is said to occur. Statistical methods are sometimes used to counter bias introduced from confounding by producing ‘adjusted’ estimates of intervention effects, and part of the assessment of study quality may involve making judgements about the appropriateness of the analysis as well as the design and execution of the study.

The variety of study designs classified as NRS, and their varying susceptibility to different biases, makes it difficult to produce a generic robust tool that can be used to evaluate risk of bias. Within a review that includes NRS of different designs, several tools for assessment of risk of bias may need to be created. Inclusion of a knowledgeable methodologist in the review team is essential to identify the key areas of weakness in the included study designs.

With randomized trials, assessment of the risk of bias focuses on systematic bias, which is usually assumed to be ‘optimistic’ in direction. The tendency for researchers to design, execute, analyse and report their primary studies to give the findings that are expected, consciously or subconsciously, is also likely to apply to NRS where researchers have control over key decisions (e.g. allocation to intervention, or selection of centres). In truly observational NRS, bias arising from ‘confounding by indication’ may not be so consistent; healthcare professionals may have differing opinions about the appropriateness of alternative interventions for their patients, contingent on the patients’ presenting severity of illness or co-morbidities. Differences in case-mix between locations that are being compared may be haphazard. Therefore, when reviewing NRS, the variability of biases and the between-study heterogeneity they induce is at least as important as systematic bias when reviewing NRS.

13.5.1.2 Evidence of risk of bias in non-randomized studies

Some insight into the risk of bias in non-randomized studies can be obtained by comparing randomized trials at low risk of bias with randomized trials at high risk of bias. Controlled trials that allocate participants by quasi-randomization, or that fail to conceal allocation during recruitment, are at risk of selection bias, just like a prospectively conducted, overtly non-randomized, trial or cohort study. Chapter 8 reviews evidence on several aspects of risk of bias in randomized trials, and points out that methodological limitations in randomized trials tend to exaggerate the beneficial effects of interventions.

Researchers have also compared the findings of separate meta-analyses of randomized trials and NRS of the same research question, assuming that such methodological systematic reviews provide a way to investigate the risk of bias in NRS. Some reviews of this kind have reported discrepancies by study design but fair comparisons are very difficult to make (MacLehose 2000). There are at least two reasons for this:

- Randomized trials and NRS of precisely the same question are rare; for example, studies of the same intervention using different study designs usually differ systematically with respect to the population, intervention or outcome.
- Randomized trials and NRS may differ systematically in several ways with respect to their risk of bias (reporting biases as well as selection, performance, detection and attrition biases), and NRS are frequently of relatively poor quality.

These reasons may explain the inconsistent conclusions from methodological systematic reviews that have compared findings from randomized trials and NRS of the same research question. Deeks et al. reviewed eight such reviews (Deeks 2003), and found that:

- 5/8 concluded that there were differences between effects estimated by randomized trials and NRS for many but not all interventions, with no consistent pattern;
- 1/8 concluded that NRS overestimated the effect [benefit] for all interventions studied;
- 2/8 concluded that the effects estimated by randomized trials and NRS were “remarkably similar”.

A similar methodological review compared the findings of randomized trials and patient preference studies (King 2005). The review concluded that there is little evidence that preferences “significantly affect validity”, such that preferences did not appear to confound intervention effects.

Some considerations in the interpretation of these sorts of empirical studies are relevant. First, both the publication of primary studies and the selection of primary studies by review authors may be biased. There is also the possibility of bias in their classification of the review findings. Deeks et al. found that the same comparison was sometimes classified as discrepant in one review and comparable in a second. This highlights the difficulty of defining what represents a ‘difference’.

Second, the observation that differences were not consistently optimistic remains an important one and is consistent with the principle that effect estimates from NRS are more heterogeneous than expected by chance (Greenland 2004). Some empirical evidence for this comes from innovative simulation studies (Deeks 2003). Deeks et al. pointed out that biases in NRS are highly variable, and may best be considered as introducing extra uncertainty in the results rather than an estimable systematic bias. This uncertainty acts over and above that accounted for in confidence intervals, and in large studies may easily be 5 to 10 times the magnitude of the 95% confidence interval.

Finally, methodological reviews are caught in a circular loop: they need to assume either that NRS are valid and hence differences between effect estimates from randomized trials and NRS are also valid and can be attributed to external factors, or that NRS are biased and hence differences between effect estimates from randomized trials and NRS can be explained by differential risk of bias. The truth may well lie somewhere in between these extremes, but the fact remains that methodological reviews cannot unequivocally partition discrepancies to different sources. Moreover, if multiple factors distinguish randomized trials and NRS and influence effect size, then observing no difference between the effect sizes estimated from randomized trials and NRS can also be explained as the consequence of effects of multiple factors influencing the effect of an intervention in different directions. It is not logical to assume that finding no difference means that NRS are valid and finding a difference means that NRS are not valid.

13.5.2 Guidance and resources available to support review authors

13.5.2.1 General considerations in assessing risk of bias in non-randomized studies

Reporting of randomized trials is relatively straightforward and, increasingly, guided by the CONSORT statement (Moher 2001). A similar consensus statement, STROBE, for the reporting of observational epidemiological studies has been developed, although much more recently (Vandenbroucke 2007, von Elm 2007). Therefore, the quality of reporting of information required to assess the risk of bias is likely to be less good for NRS. This is likely to hinder any assessment of risk of bias.

A protocol is a tool to protect against bias; when registered in advance of a study starting, it proves that aspects of study design and analysis were considered in advance of starting to recruit, and that data definitions and methods for standardizing data collection were defined. Because of the need for research ethics approval, all randomized trials must have a protocol, even if protocols vary in their quality and the items that they specify; many randomized trials, particularly those sponsored by industry, also have detailed study manuals. Historically, researchers have not had to obtain research ethics approval for many NRS, and primary NRS rarely report whether the methods are based on a protocol. Therefore, the protection offered by a protocol often does not exist for NRS. The implications of not having a protocol have not been researched. However, it means, for example, that there is no constraint on the tendency of researchers to ‘cherry-pick’ outcomes, subgroups and analyses to report, which happens to a greater or lesser extent even in randomized trials where protocols exist (Chan 2004).

In common with randomized trials, dimensions of bias to be assessed include selection bias (concerning comparability of groups, confounding and adjustment), performance bias (concerning the fidelity of the interventions, and quality of the information regarding who received what interventions, including blinding of participants and healthcare providers), detection bias (concerning unbiased and correct assessment of outcome, including blinding of assessors), attrition bias (concerning completeness of sample, follow-up and data) and reporting bias (concerning publication biases and selective reporting of results). Assessment of risk of bias in randomized trials has developed by identifying the design features which are used to prevent each of these dimensions, and noting whether each trial fulfils the requirements. Risk of bias assessments for NRS should proceed in the same way, with pre-specification of the features to be assessed in the protocol, recording what happened in the study, and a judgement of whether this was adequate, inadequate or unclear as a method to avoid risk of this particular bias. Determining these features is likely to require expert input from an epidemiologist, and will depend in part on the clinical question. Particular care should be given to the assessment of confounding (see Section 13.5.2.2).

The reason for careful attention to the design *features* of primary studies (such as how participants were allocated to groups, or which parts of the study were prospective) rather than design *labels* (such as ‘cohort’ or ‘cross-sectional’) is because it is hypothesized that the risk of bias is influenced by the specific features of a study rather than a broad categorization of the approach taken. Furthermore, terms such as ‘cohort’ and ‘cross-sectional’ are ambiguous and cover a diverse range of specific study designs. No empirically-derived list is available of study design features that are relevant to the risk of bias, although a shortlist can be constructed from evidence and theory about the risk of bias in aetiological studies and randomized trials (see Section 13.2.2 and 13.4.2).

Because of the diversity of NRS, different methods may be needed to assess NRS with different design features. One important distinction is between studies in which allocation to groups is by outcome (e.g. case-control studies) and studies in which allocation to groups is more directly related to interventions. In the former type of study, it is the exposure of interest, rather than the outcome, that is most susceptible to bias; review authors need to ask whether researchers assessing the exposure were masked to whether participants had experienced the outcome or not (i.e. were cases or controls). Case-control studies are well suited to investigating associations between rare outcomes and multiple exposures, so may have an important role in generating evidence about the potential adverse effects and unintended beneficial effects of interventions. They have also been used to evaluate large-scale public health interventions such as accident prevention and screening (MacLehose 2000), which are difficult or expensive to evaluate by randomized trials. However, review authors should familiarize themselves with epidemiological considerations that particularly apply to such studies (Rothman 1986). Note that some analyses of patient registries also have similarities with case-control studies: for example, if the entire database is divided into groups of patients who have or have not experienced a particular outcome and exposures associated with the outcome are investigated. Review authors

require a deeper knowledge of epidemiology when assessing the risk of bias in NRS, compared with randomized trials.

13.5.2.2 Confounding and adjustment

Researchers do not always make the same decisions concerning confounding factors, so the method used to control for confounding is an important source of heterogeneity between studies. There may be differences in the confounding factors considered, the method used to control for confounding and the precise way in which confounding factors were measured and included in analyses. Many (but not all) NRS describe the confounding factors that were considered and whether confounding was taken into account by the study design or analysis; most also report the baseline characteristics of the groups being compared. However, assessing what researchers actually did to control for confounding may be difficult; far fewer studies describe precisely how confounding factors were measured or fitted as covariates in regression models (e.g. as a continuous, ordinal, or grouped categorical variable).

Some specific suggestions for assessing risk of selection bias are as follows.

- At the stage of writing the protocol, list potential confounding factors.
- Identify the confounding factors that the researchers have considered and those that have been omitted. Note the ways in which they have been measured (the ability to control for a confounding factor depends on the precision with which the factor is measured).
- Assess the balance between comparator groups at baseline with respect to the main prognostic or confounding factors.
- Identify what researchers did to control for selection bias, i.e. any design features used for this purpose (e.g. matching or restriction to particular subgroups) and the methods of analysis (e.g. stratification or regression modelling with propensity scores or covariates).

There is no established method for identifying a pre-specified set of important confounders. Listing potential confounding factors should certainly be done ‘independently’ and, one might argue, ‘systematically’. The list should not be generated solely on the basis of factors considered in primary studies included in the review (at least, not without some form of independent validation), since the number of potential confounders is likely to increase over time (hence, older studies may be out of date) and researchers themselves may simply choose to measure confounders considered in previous studies (hence, such a list could be selective). (Researchers investigating aetiological associations often do not explain their choice of confounding factors (Pocock 2004).) Rather, the list should be based on evidence (although undertaking a systematic review to identify all potential prognostic factors is extreme) and expert opinion from members of the review team and advisors.

Reporting results of assessments of confounders in a Cochrane review may best be achieved by creating additional tables listing the pre-stated confounders as columns, the studies as rows, and indicating whether each study: (i) restricted participant selection so that all groups had the same value for the confounder (e.g. restricting the study to male participants only); (ii) demonstrated balance between groups for the confounder; (iii) matched on the confounder; or (iv) adjusted for the confounder in statistical analyses to quantify the effect size.

13.5.2.3 Tools for assessing methodological quality or risk of bias in non-randomized studies

Chapter 8 (Section 8.5) describes the ‘Risk of bias’ tool that review authors are expected to use for assessing risk of bias in randomized trials. This involves consideration of six features: sequence generation, allocation sequence concealment, blinding, incomplete outcome data, selective outcome reporting and ‘other’ potential sources of bias. Items are assessed by: (i) providing a description of

what happened in the study; (ii) providing a judgement on the adequacy of the study with regard to the item. The judgement is formulated by answering a pre-specified question, such that an answer of ‘Yes’ indicates low risk of bias, an answer of ‘No’ indicates high risk of bias, and an answer of ‘Unclear’ indicates unclear or unknown risk of bias. The tool was not developed with NRS in mind, and the six domains are not necessarily appropriate for NRS. However, the general structure of the tool and the assessments seems useful to follow when creating risk of bias assessments for NRS.

For experimental and controlled studies, and for prospective cohort studies (see [Box 13.1.a](#) and [Section 13.2.2](#)), the six domains in the standard ‘Risk of bias’ tool could usefully be assessed, whether allocation is randomized or not. This is the minimum assessment review authors should carry out and more details will usually be required. An additional component is to assess the risk of bias due to confounding. The depth of this assessment is likely to depend on the heterogeneity between studies and whether the review authors propose a quantitative synthesis (see [Section 13.6](#)). If studies are heterogeneous and no quantitative synthesis is proposed, then a less detailed assessment can nevertheless serve the purposes of illustrating the heterogeneity and informing interpretation of the findings of the review.

Many instruments for assessing methodological quality of non-randomized studies of interventions have been created, and were reviewed systematically by Deeks et al. (Deeks 2003). In their review they located 182 tools, which they reduced to a shortlist of 14, and identified six as potentially useful for systematic reviews as they “force the reviewer to be systematic in their study assessments and attempt to ensure that quality judgements are made in the most objective manner possible”. However, all six required a degree of adjustment as they neglected to elicit detailed information about how study participants were allocated to groups, which in terms of the risk of selection bias is likely to be critical. Not all of the six tools were suitable for different study designs. In common with some tools for assessing the quality of randomized trials, some did not distinguish items relating to the quality of the study and the quality of reporting of the study. The two most useful tools identified in this review are the Downs and Black instrument and the Newcastle-Ottawa Scale (Downs 1998, Wells 2008).

The Downs and Black instrument has been modified for use in a methodological systematic review (MacLehose 2000). The reviewers found that some of the 29 items were difficult to apply to case-control studies, that the instrument required considerable epidemiological expertise and that it was time consuming to use. The Newcastle-Ottawa Scale, which has been used in NRSMG workshops to illustrate issues in data extraction from primary NRS, contains only eight items and is simpler to apply (Wells 2008). However, the items may still need to be customized to the review question of interest. Review authors also need to be aware of differences in epidemiological terminology in different countries; for example, the Newcastle-Ottawa Scale uses the term ‘selection bias’ to describe what others may call ‘applicability’ or ‘generalizability’.

Acknowledging the importance of distinguishing between ‘what researchers do’ and ‘what researchers report’, review authors may also find it helpful to consider items included in reporting statements for randomized trials (Moher 2001) and observational epidemiological studies (Vandenbroucke 2007) in order to highlight gaps in reporting (and execution) in NRS (Reeves 2004, Reeves 2007).

13.5.2.4 Practical limitations in assessing risk of bias in non-randomized studies

Two studies of systematic reviews that included NRS have commented that only a minority of reviews assessed the methodological quality of included studies (Audige 2004, Golder 2006a). Members of the NRSMG have gained experience of trying to assess risk of bias in non-randomized studies. Anecdotally, review authors have reported that NRS are generally of poor methodological quality, or are poorly reported so that assessing methodological quality and risk of bias consistently across primary studies is difficult or impossible (Kwan 2004). Even the Newcastle-Ottawa scale has been

reported to be difficult to apply, so agreement between review authors is likely to be modest. Methodological information can be difficult to find in papers, making the task frustrating, especially when using some of the more detailed instruments; review authors may spend a long time searching for details of what researchers did, only to conclude that the information was not reported. Nevertheless, collecting some factual information (for example, the confounders considered and what researchers did about confounding) can still be useful since such information illustrates the extent of heterogeneity between studies.

13.5.3 Summary

- At the stage of writing the protocol for the review, compile a list of potential confounding factors and justify the choice.
- At the stage of writing the protocol for the review, decide how the risk of bias in primary studies will be assessed, including the extent of control for confounding.
- For NRS conducted entirely prospectively, apply the methods that the Collaboration recommends for randomized trials.
- There is no single recommended instrument, so review authors are likely to need to include supplementary risk of bias instruments or items.
- Issues such as confounding cannot easily be addressed with in the format of the new risk of bias tool and require creation of additional tables for reporting assessments.
- Collecting some factual information (for example, the confounders considered and what researchers did about confounding) is useful since such information illustrates the extent of heterogeneity between studies.
- Review authors who choose to include case-control studies in a Cochrane review should ensure that they are familiar with common pitfalls that can affect such studies and that they assess their susceptibility to bias using an instrument designed for this purpose.
- Review authors may decide that collecting great detail about the risk of confounding and other biases is not warranted. However, if this approach is taken, review authors must acknowledge the potential extent of the heterogeneity between studies with respect to potential residual confounding and other biases and demonstrate that they have considered this source of heterogeneity in their interpretation of the findings of the primary NRS reviewed.

13.6 Synthesis of data from non-randomized studies

13.6.1 What is different when including non-randomized studies?

Review authors should expect greater heterogeneity in a systematic review of NRS than a systematic review of randomized trials. This is due to the increased potential for methodological diversity through variation between primary studies in their risk of selection bias, variation in the way in which confounding is considered in the analysis and greater risk of other biases through poor design and execution. There is no way of controlling for these biases in the analysis of primary studies and no established method for assessing how, or the extent to which, these biases affect primary studies (but see Chapter 8).

There is a body of opinion that it is appropriate to pool results of non-randomized studies when they have large effects, but the logic of this view can be questioned. NRS with large effects are as likely (perhaps more likely) to be biased and to be heterogeneous as NRS with small effects. Judgements about the risk of bias and heterogeneity should be based on critical appraisal of the characteristics and methods of included studies, not on their results.

When assessing similarity of studies prior to a meta-analysis, review authors should also keep in mind that some features of studies, for example assessment of outcome not masked to intervention allocation, may be relatively homogeneous across NRS but still leave all studies at risk of bias.

If authors judge that included NRS are both reasonably resistant to biases and relatively homogeneous in this respect, they may wish to combine data across studies using meta-analysis (Taggart 2001). Unlike for randomized trials, it will usually be appropriate to analyse adjusted, rather than unadjusted, effect estimates, i.e. analyses that attempt to ‘control for confounding’. This may require authors to choose between alternative adjusted estimates reported for one study. Meta-analysis of adjusted estimates can be performed as an inverse-variance weighted average, for example using the ‘Generic inverse-variance’ outcome type in RevMan (see Chapter 9, Section 9.4.3). In principle, any effect measure used in meta-analysis of randomized trials can also be used in meta-analysis of non-randomized studies (see Chapter 9, Section 9.2), although the odds ratio will commonly be used as it is the only effect measure for dichotomous outcomes that can be estimated from case-control studies, and is estimated when logistic regression is used to adjust for confounders.

One danger is that a very large NRS of poor methodological quality (for example based on routinely collected data) may dominate the findings of other smaller studies at less risk of bias (perhaps carried out using customized data collection). Authors need to remember that the confidence intervals for effect estimates from larger NRS are less likely to represent the true uncertainty of the observed effect than are the confidence intervals for smaller NRS (see Section 13.5.1.2), although there is no way of estimating or correcting for this.

13.6.2 Guidance and resources available to support review authors

13.6.2.1 Controlling for confounding

Imbalances in prognostic factors in NRS (e.g. ‘confounding by indication’ (Grobbee 1997)) must be accounted for in the statistical analysis. There are several methods to control for confounding. Matching, i.e. the generation of similar intervention groups with respect to important prognostic factors, can be used to lessen confounding at the study design stage. Stratification and regression modelling are statistical approaches to control for confounding, which result in an estimated intervention effect adjusted for imbalances in observed prognostic factors. Some analyses use propensity score methods as part of a two-stage analysis. The probability of an individual receiving the experimental intervention (the propensity score) is first estimated according to their characteristics using a logistic regression model. This single summary measure of case-mix is then used for matching, stratification or in a regression model.

Matching

The selection of patients with similar values for important prognostic factors results in more comparable groups. Therefore, matching can be seen as a type of confounder adjustment. Matching can be either at the level of individual patients (i.e. one or more control participants are selected who have similar characteristics to an intervention participants) or at the level of participants strata (i.e. selecting participants so that there are roughly the same number of control participants in one stratum, for example 60 years or older, as in the intervention group). Where direct matching has been used, the paired nature of the data has to be considered in the statistical analysis of a single study in order to obtain appropriate confidence intervals for the estimated effect of the intervention. Matching on a single measure such as the propensity score is easier to achieve than matching individuals with a particular set of characteristics.

Stratification

Stratification involves the division of participants into subgroups with respect to categorical (or categorized quantitative) prognostic factors, for example classifying age into decades, or weight into quartiles. The intervention effect is then estimated in each stratum and a pooled estimate is calculated across strata. This procedure can be interpreted as a meta-analysis at the level of an individual study. For dichotomous outcomes, the Mantel-Haenszel method is often used to estimate the overall intervention effect, with versions available for the odds ratio, the risk ratio and the risk difference as measures of intervention effect. Again, the propensity score may be used as the stratification variable.

Modelling

In a modelling approach, information on intervention and prognostic factors is incorporated into a regression equation. Advantages of regression models include the possibility of incorporating quantitative factors without categorization and the possibility of modelling trends in confounders measured on an ordinal scale. For dichotomous outcomes, a logistic regression model is almost always used to estimate the adjusted intervention effect. Thus, the odds ratio is (implicitly) used as the measure of intervention effect. Regression models are also available for risk ratio and absolute risk reduction measures of effect but these models are rarely used in practice. A linear regression model is typically used for continuous outcomes (perhaps after transformation of one or more variables), and a proportional hazards regression (Cox regression) model is typically used for time-to-event data. Regression models may also use the propensity score alone or in combination with other participant characteristics as explanatory variables.

Review authors should acknowledge that in any non-randomized study, even when experimental and control groups appear comparable at baseline, the effect size estimate is still at risk of bias due to residual confounding. This is because all methods to control for confounding are imperfect, for example for the following reasons.

- Unknown, and consequently unmeasured, confounding factors, which cannot be controlled for.
- Poor resolution in the measurement of confounders, e.g. co-morbidity assessed on a simple ordinal scale (Concato 1992), which represents non-differential error misclassification with respect to confounders.
- Practical constraints on the resolution of matching, and the number of confounders on which participants can be matched, in matched analyses.
- Poor resolution in the way confounders are measured in stratified analyses, or handled in analyses, illustrated by the width of strata (e.g. decades of age); this limitation also applies to regression models when confounders are categorized and modelled discretely.
- Assumptions in the way confounders are modelled in regression analyses, because of imperfect knowledge of the shape of the association between confounder and outcome.

There is no established method for judging the likely extent of residual confounding. The direction of bias from confounding is unpredictable and may differ between studies.

13.6.2.2 Combining studies

Estimated intervention effects for different study designs can be expected to be influenced to varying degrees by different sources of bias (see Section 13.5). Results from different study designs should be expected to differ systematically, resulting in increased heterogeneity. Therefore, we recommend that NRS which used different study designs (or which have different design features), or randomized trials and NRS, should not be combined in a meta-analysis.

Because of the need to control for confounding as best as possible, the estimated intervention effect and its standard error (or confidence interval) are key pieces of information which should be used for pooling NRS in a meta-analysis. (Simple numerators and denominators, or means and standard errors, for intervention and control groups cannot control for confounding unless the groups have been matched at the design stage.) Consequently, meta-analysis methods based on estimates and standard errors, and in particular the generic inverse-variance method, will be suitable for NRS (see Chapter 9, Section 9.4.3).

It is straightforward to extract an adjusted effect estimate and its standard error for a meta-analysis if a single adjusted estimate is reported for a particular outcome in a primary NRS. However, many NRS report both unadjusted and adjusted effect estimates, and some NRS report multiple adjusted estimates from analyses including different sets of covariates. Review authors should record both unadjusted and adjusted effect estimates but it can be difficult to choose between alternative adjusted estimates. No general recommendation can be made for the selection of which adjusted estimate is preferable. Possible selection rules are:

- use the estimate from the model that adjusted for the maximum number of covariates;
- use the estimate that is identified as the primary adjusted model by the authors; and
- use the estimate from the model that includes the largest number of confounders considered important at the outset by the review authors.

Sensitivity analyses could be performed by pooling separately the most optimistic and pessimistic results from each included study.

There is a subtle statistical point regarding the different interpretation of adjusted and unadjusted effects when expressed as odds or hazard ratios. The unadjusted effect estimate is known as the population average effect, and if the estimate were unbiased would be the effect of intervention observed in a population with an average mixture of prognostic characteristics. When estimates are adjusted for prognostic characteristics, the estimated effects are known as conditional estimates and are the intervention effects that would be observed in groups with particular combinations of the adjusted covariates. Mathematical research has shown that conditional estimates are usually larger (further from an OR or HR of 1) than population average estimates. This phenomenon may not be observed in systematic reviews due to heterogeneity in the estimates of the studies.

13.6.2.3 Analysis of heterogeneity

The exploration of possible sources of heterogeneity between studies should be part of any Cochrane review, and is discussed in detail in Chapter 9 (Section 9.6). Non-randomized studies may be expected to be more heterogeneous than randomized trials, given the extra sources of methodological diversity and bias. The simplest way to show the variation in results of studies is by drawing a forest plot (see Chapter 11, Section 11.3.2).

It may be of value to undertake meta-regression analyses to identify important determinants of heterogeneity, even in reviews where studies are considered too heterogeneous to pool. Such analyses may help to identify methodological features which systematically relate to observed intervention effects, and help to identify the subgroups of studies most likely to yield valid estimates of intervention effects.

13.6.2.4 When pooling is judged not to be appropriate

Before undertaking a meta-analysis, review authors must ask themselves the standard question about whether primary studies are 'similar enough' to justify pooling (see Chapter 9). Forest plots in RevMan allow the presentation of estimates and standard errors for each study, using the 'Generic

If included studies are not sufficiently homogeneous to combine in a meta-analysis (which is expected to be the norm for reviews that include NRS), the NRSMG recommends displaying the results of included studies in a forest plot but suppressing the pooled estimate. Studies may be sorted in the forest plot (or shown in separate forest plots) by study design feature, or some other feature believed to reflect susceptibility to bias (e.g. number of Newcastle-Ottawa Scale ‘stars’ (Wells 2008)). Heterogeneity diagnostics and investigations (e.g. a test for heterogeneity, the I^2 statistic and meta-regression analyses) are worthwhile even when a judgement has been made that calculating a pooled estimate of effect is not (Higgins 2003, Siegfried 2003).

Narrative syntheses are, however, problematic, because it is difficult to set out or describe results without being selective or emphasizing some findings over others. Ideally, authors should set out in the review protocol how they plan to use narrative synthesis to report the findings of primary studies.

13.6.3 Summary

- Heterogeneity will be greater in a systematic review of NRS than in a systematic review of randomized trials. Therefore, authors should consider very carefully the likely extent of heterogeneity between included studies when deciding whether to pool findings quantitatively (i.e. by meta-analysis). We expect pooling of effect estimates from NRS to be the exception, rather than the rule.
- Effect estimates from NRS should not be combined with effect estimates from randomized trials, or across NRS that have dissimilar study design features.
- Forest plots should be used to summarize the findings from included studies.
- Heterogeneity diagnostics and investigations may be used irrespective of whether or not a decision has been taken to pool effect estimates from different studies.

13.7 Interpretation and discussion

13.7.1 Challenges in interpreting Cochrane reviews of effectiveness that include non-randomized studies

Review authors face great challenges in demonstrating convincingly that the result of a Cochrane review of NRS can give anything close to a definitive answer about the likely effect of an intervention (Deeks 2003). In many situations, reviews of NRS are likely to conclude that calculating an ‘average’ effect is not helpful (Siegfried 2003), that evidence from NRS is inadequate to prove effectiveness or harm (Kwan 2004) and that randomized trials should be undertaken (Taggart 2001).

Challenges arise at all stages of conducting a review of NRS: deciding which study designs to include, searching for studies, assessing studies for potential bias, and deciding whether to pool results. A review author needs to satisfy the reader of the review that these challenges have been adequately addressed, or should discuss how and why they cannot be met. In this section, the challenges are

illustrated with reference to issues raised in the different sections of this chapter. The Discussion section of the review should address the extent to which the challenges have been met.

13.7.1.1 Have all important and relevant studies been included?

Even if the choice of eligible study designs can be justified, it may be difficult to show that all relevant studies have been identified because of poor indexing and inconsistent use of study design labels by researchers. Comprehensive search strategies that focus only on the health condition and intervention of interest are likely to result in a very long list of citations including relatively few eligible studies; conversely, restrictive strategies will inevitably miss some eligible studies. In practice, available resources may make it impossible to process the results from a comprehensive search, especially since authors will often have to read full papers rather than abstracts to determine eligibility. The implications of using a more or less comprehensive search strategy are not known.

13.7.1.2 Has the risk of bias to included studies been adequately assessed?

Interpretation of the results of a review of NRS must include consideration of the likely direction and magnitude of bias. Biases that affect randomized trials also affect NRS but typically to a greater extent. For example, attrition in NRS is often worse (and poorly reported), intervention and outcome assessment are rarely conducted according to standardized protocols, and outcomes are rarely blind. Too often these limitations of NRS are seen as part of doing a NRS, and their implications for risk of bias are not properly considered. For example, some users of evidence may consider NRS that investigate long-term outcomes to have ‘better quality’ than randomized trials of short-term outcomes, simply on the basis of their relevance without appraising their risk of bias (see Section 13.2.1.4).

Assessing the magnitude of confounding in NRS is especially problematic. Review authors must not only have adequate methods for assessment but also collect and report adequate detail about the confounding factors considered by researchers and the methods used to control for confounding. The information may not be available from the reports of the primary studies, preventing the review authors from investigating differences in the methods of eligible studies and other sources of heterogeneity that were considered likely to be important when the protocol was written.

Authors must remember the following points about confounding:

- The direction of the bias introduced by confounding is unpredictable.
- Methods used by researchers to control for confounding are like to vary between studies.
- The extent of residual confounding in any particular study is unknown, and is likely to vary between studies.
- Residual confounding (and other biases) means that confidence intervals underestimate the true uncertainty around an effect estimate.
- It is important to identify the likely confounding factors that have not been adjusted for, as well as those that have been adjusted for.

The challenges described above affect all systematic reviews of NRS. However, challenges may be less extreme in some healthcare areas (e.g. confounding may be less of a problem in observational studies of long-term or adverse effects, or some public health primary prevention interventions).

One clue to the presence of bias is notable between-study heterogeneity. Although heterogeneity can arise through differences in participants, interventions and outcome assessments, the possibility that bias is the cause of heterogeneity in reviews of NRS must be considered seriously. However, lack of heterogeneity does not indicate lack of bias, since it is possible that a consistent bias applies in all studies.

Can the magnitude and direction of bias be predicted? This is a subject of ongoing research which is attempting to gather empirical evidence on factors (such as study design and intervention type) that determine the size and direction of these biases. The ability to predict both the likely magnitude of bias and the likely direction of bias would greatly improve the usefulness of evidence from systematic reviews of NRS. There is currently some evidence that in some limited circumstances the direction, at least, can be predicted (Henry 2001)

13.7.2 Evaluating the strength of evidence provided by reviews that include non-randomized studies

'Exposing' the evidence from NRS on a particular health question enables informed debate about its meaning and importance, and the certainty which can be attributed to it. Critically, there needs to be a debate about the chance that the observed findings could be misleading. Formal hierarchies of evidence all place NRS low down on the list, but above those of clinical opinion (Eccles 1996, National Health and Medical Research Council 1999, Oxford Centre for Evidence-based Medicine 2001). This emphasizes the general concern about biases in NRS, and the difficulties of attributing causality to the observed effects. The strength of evidence provided by a systematic review of NRS is likely to depend on meeting the challenges set out in Section 13.7.1. The ability to meet these challenges will vary with healthcare context and outcome. In some contexts little confounding is likely to occur. For example, little prognostic information may be known when infants are vaccinated, limiting possible confounding (Jefferson 2005).

Whether the debate concludes that there is a need for randomized trials or that the evidence from NRS is adequate for informed decision-making will depend on the cost placed on the uncertainty arising through use of potentially biased study designs, and the collective value of the observed effects. This value may depend on the wider healthcare context. It may not be possible to include assessments of the value within the review itself, and it may become evident only as part of the wider debate following publication.

For example, is evidence from NRS of a rare serious adverse effect adequate to decide that an intervention should not be used? The evidence is uncertain (due to a lack of randomized trials) but the value of knowing that there is the possibility of a potentially serious harm is considerable, and may be judged sufficient to withdraw the intervention. (It is worth noting that the judgement about withdrawing an intervention may depend on whether equivalent benefits can be obtained from elsewhere without such a risk; if not, the intervention may still be offered but with full disclosure of the potential harm.) Where evidence of benefit is not based on randomized trials and is therefore equivocal, the value attached to a systematic review of NRS of harm may be even greater.

In contrast, evidence of a small benefit of a novel intervention from a systematic review of NRS may not be sufficient for decision makers to recommend widespread implementation in the face of the uncertainty of the evidence and the substantial costs arising from provision of the intervention. In these circumstances, decision makers are likely to conclude that randomized trials should be undertaken if practicable and if the investment in the trial is likely to be repaid in the future.

The GRADE scheme for assessing the quality of a body of evidence is recommended for use in 'Summary of findings' tables in Cochrane reviews, and is summarized in Chapter 12 (Section 12.2). There are four quality levels: 'high', 'moderate', 'low' and 'very low'. A collection of studies that can be crudely categorized as randomized trials starts at the highest level, and may be downgraded due to study limitations (risk of bias), indirectness of evidence, heterogeneity, imprecision or publication bias. Collections of observational studies start at a level of 'low', and may be upgraded due to a large

magnitude of effect, lack of concern about confounders or a dose-response gradient. Review authors will need to make judgements about whether evidence from NRS should be upgraded from a low level or possibly (e.g. in the case of quasi-randomized trials) downgraded from a high level.

13.7.3 Guidance for potential review authors

Carrying out a systematic review of NRS is much more difficult than carrying out a systematic review of randomized trials. It is likely that complex decisions, requiring expert methodological or epidemiological advice, will need to be made at each stage of the review. Potential review authors should therefore seek to collaborate with epidemiologists or methodologists, irrespective of whether a review aims to investigate harms or benefits, short-term or long-term outcomes, frequent or rare events.

Healthcare professionals are keen to be involved in doing reviews of NRS in areas where there are few or no randomized trials because they have the ambition to improve the evidence-base in their specialty areas (the motivation for most Cochrane reviews). Methodologists are keen for more systematic reviews of NRS to inform the many areas of uncertainty in methodology highlighted by these chapters. However, healthcare professionals should also recognize that (a) the resources required to do a systematic review of NRS are likely to be much greater than for a systematic review of randomized trials and (b) the conclusions are likely to be much weaker and may make a relatively small contribution to the topic. Therefore, authors and CRG editors need to decide at an early stage whether the investment of resources is likely to be justified by the priority of the research question.

Bringing together the required team of healthcare professionals and methodologists may be easier for systematic reviews of NRS to estimate the effects of an intervention on long-term and rare adverse outcomes, for example when considering the side effects of drugs. However, these reviews may require the input of additional specialist authors, for example with relevant pharmacological expertise. There is a pressing need in many health conditions to supplement traditional systematic reviews of randomized trials of effectiveness with systematic reviews of adverse (unintended) effects. It is likely that these systematic reviews will usually need to include NRS.

13.8 Chapter information

Authors: Barnaby C Reeves, Jonathan J Deeks, Julian PT Higgins and George A Wells on behalf of the Cochrane Non-Randomised Studies Methods Group.

This chapter should be cited as: Reeves BC, Deeks JJ, Higgins JPT, Wells GA. Chapter 13: Including non-randomized studies. 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: We gratefully acknowledge Ole Olsen, Peter Gøtzsche, Angela Harden, Mustafa Soomro, Guido Schwarzer and Bev Shea for their early drafts of different sections. We also thank Laurent Audigé, Duncan Saunders, Alex Sutton, Helen Thomas and Gro Jamtved for comments on previous drafts.

Box 13.8.a: The Cochrane Non-Randomised Studies Methods Group

The Non-Randomised Studies Methods Group (NRSMG) of the Cochrane Collaboration advises the Steering Group to set policy and formulate guidance about the inclusion of non-randomized studies (NRS) of the effectiveness of healthcare interventions in Cochrane reviews. Membership of the group is open to anyone who wishes to contribute actively to the work of group. The work of the group is primarily methodological, rather than focused on particular healthcare interventions.

Activities of NRSMG members include:

- Developing guidelines to help decide when to include non-randomized data in Cochrane reviews.
- Conducting methodological research in the use of non-randomized studies, including search methods, quality assessment, meta-analysis, pitfalls and misuse.
- Conducting empirical research to compare bias in systematic reviews using both randomized and non-randomized studies, and to identify conditions under which randomized and non-randomized studies have led to similar conclusions, and situations in which the conclusions have been clearly contradictory.
- Collating examples of healthcare questions that (a) have been studied using both non-randomized studies and randomized trials, and (b) have not been (or which for a long period have not been) studied adequately by means of randomized trials.
- Providing training at annual Cochrane Colloquia.

13.9 References

Audige 2004

Audige L, Bhandari M, Griffin D, Middleton P, Reeves BC. Systematic reviews of nonrandomized clinical studies in the orthopaedic literature. *Clinical Orthopaedics and Related Research* 2004; 249-257.

Chan 2004

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004; 291: 2457-2465.

Concato 1992

Concato J, Horwitz RI, Feinstein AR, Elmore JG, Schiff SF. Problems of comorbidity in mortality after prostatectomy. *JAMA* 1992; 267: 1077-1082.

Deeks 2003

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomised intervention studies. *Health Technology Assessment* 2003; 7: 27.

Doll 1993

Doll R. Doing more good than harm: The evaluation of health care interventions: Summation of the conference. *Annals of the New York Academy of Sciences* 1993; 703: 310-313.

Downs 1998

Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health* 1998; 52: 377-384.

Eccles 1996

Eccles M, Clapp Z, Grimshaw J, Adams PC, Higgins B, Purves I, Russel I. North of England evidence based guidelines development project: methods of guideline development. *BMJ* 1996; 312: 760-762.

Fraser 2006

Fraser C, Murray A, Burr J. Identifying observational studies of surgical interventions in MEDLINE and EMBASE. *BMC Medical Research Methodology* 2006; 6: 41.

Furlan 2006

Furlan AD, Irvin E, Bombardier C. Limited search strategies were effective in finding relevant nonrandomized studies. *Journal of Clinical Epidemiology* 2006; 59: 1303-1311.

Glasziou 2007

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007; 334: 349-351.

Golder 2006a

Golder S, Loke Y, McIntosh HM. Room for improvement? A survey of the methods used in systematic reviews of adverse effects. *BMC Medical Research Methodology* 2006; 6: 3.

Golder 2006b

Golder S, McIntosh HM, Duffy S, Glanville J, Centre for Reviews and Dissemination and UK Cochrane Centre Search Filters Design Group. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. *Health Information and Libraries Journal* 2006; 23: 3-12.

Golder 2006c

Golder S, McIntosh HM, Loke Y. Identifying systematic reviews of the adverse effects of health care interventions. *BMC Medical Research Methodology* 2006; 6: 22.

Greenland 2004

Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *International Journal of Epidemiology* 2004; 33: 1389-1397.

Grobbée 1997

Grobbée DE, Hoes AW. Confounding and indication for treatment in evaluation of drug treatment for hypertension. *BMJ* 1997; 315: 1151-1154.

Henry 2001

Henry D, Moxey A, O'Connell D. Agreement between randomized and non-randomized studies: the effects of bias and confounding. *9th Cochrane Colloquium*, Lyon (France), 2001.

Higgins 2003

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; 327: 557-560.

Jefferson 2005

Jefferson T, Smith S, Demicheli V, Harnden A, Rivetti A, Di Pietrantonj C. Assessment of the efficacy and effectiveness of influenza vaccines in healthy children: systematic review. *The Lancet* 2005; 365: 773-780.

King 2005

King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, Sibbald B, Lai R. Impact of participant and physician intervention preferences on randomized trials: a systematic review. *JAMA* 2005; 293: 1089-1099.

Kwan 2004

Kwan J, Sandercock P. In-hospital care pathways for stroke. *Cochrane Database of Systematic Reviews* 2004, Issue 2. Art No: CD002924.

MacLehose 2000

MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment* 2000; 4: 1-154.

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).

National Health and Medical Research Council 1999

National Health and Medical Research Council. *A guide to the development, implementation and evaluation of clinical practice guidelines [Endorsed 16 November 1998]*. Canberra (Australia): Commonwealth of Australia, 1999.

Oxford Centre for Evidence-based Medicine 2001

Oxford Centre for Evidence-based Medicine. Levels of Evidence [May 2001]. Available from: <http://www.cebm.net/index.aspx?o=1047> (accessed 1 January 2008).

Peto 1995

Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *Journal of Clinical Epidemiology* 1995; 48: 23-40.

Petticrew 2001

Petticrew M. Systematic reviews from astronomy to zoology: myths and misconceptions. *BMJ* 2001; 322: 98-101.

Pocock 2004

Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, Kalish LA, Kasten LE, McCormack VA. Issues in the reporting of epidemiological studies: a survey of recent practice. *BMJ* 2004; 329: 883.

Reeves 2004

Reeves BC, Gaus W. Guidelines for reporting non-randomised studies. *Forschende Komplementärmedizin und klassische Naturheilkunde* 2004; 11 Suppl 1: 46-52.

Reeves 2006

Reeves BC. Parachute approach to evidence based medicine: as obvious as ABC. *BMJ* 2006; 333: 807-808.

Reeves 2007

Reeves BC, Langham J, Lindsay KW, Molyneux AJ, Browne JP, Copley L, Shaw D, Gholkar A, Kirkpatrick PJ. Findings of the International Subarachnoid Aneurysm Trial and the National Study of Subarachnoid Haemorrhage in context. *British Journal of Neurosurgery* 2007; 21: 318-23.

Rothman 1986

Rothman KJ. *Modern Epidemiology*. Boston (MA): Little, Brown & Company, 1986.

Shadish 2002

Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston (MA): Houghton Mifflin, 2002.

Siegfried 2003

Siegfried N, Muller M, Volmink J, Deeks J, Egger M, Low N, Weiss H, Walker S, Williamson P. Male circumcision for prevention of heterosexual acquisition of HIV in men. *Cochrane Database of Systematic Reviews* 2003, Issue 3. Art No: CD003362.

Taggart 2001

Taggart DP, D'Amico R, Altman DG. Effect of arterial revascularisation on survival: a systematic review of studies comparing bilateral and single internal mammary arteries. *The Lancet* 2001; 358: 870-875.

Vandenbroucke 2007

Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Medicine* 2007; 4: e297.

von Elm 2007

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLoS Medicine* 2007; 4: e296.

Wells 2008

Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm (accessed 1 January 2008).

Wieland 2005

Wieland S, Dickersin K. Selective exposure reporting and Medline indexing limited the search sensitivity for observational studies of the adverse effects of oral contraceptives. *Journal of Clinical Epidemiology* 2005; 58: 560-567.