Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions

Julian PT Higgins,a,b*† Craig Ramsay,c Barnaby C Reeves,d Jonathan J Deeks,e Beverley Shea,f Jeffrey C Valentine,g Peter Tugwellh and George Wellsi

Non-randomized studies may provide valuable evidence on the effects of interventions. They are the main source of evidence on the intended effects of some types of interventions and often provide the only evidence about the effects of interventions on long-term outcomes, rare events or adverse effects. Therefore, systematic reviews on the effects of interventions may include various types of non-randomized studies. In this second paper in a series, we address how review authors might articulate the particular non-randomized study designs they will include and how they might evaluate, in general terms, the extent to which a particular non-randomized study is at risk of important biases. We offer guidance for describing and classifying different non-randomized designs based on specific features of the studies in place of using non-informative study design labels. We also suggest criteria to consider when deciding whether to include non-randomized studies. We conclude that a taxonomy of study designs based on study design features is needed. Review authors need new tools specifically to assess the risk of bias for some non-randomized designs that involve a different inferential logic compared with parallel group trials.

Keywords: non-randomized studies; study design; bias; systematic reviews

1. Introduction

Non-randomized studies (NRS) may provide valuable evidence on the effects of interventions. They are the main source of evidence on the intended effects of many organizational or public health interventions and on interventions that cannot ethically be randomized. Furthermore, they often provide the only evidence about the effects of interventions on long-term outcomes, rare events or adverse effects. Systematic reviews on the effects of interventions, such as those prepared by The Cochrane Collaboration, the Campbell Collaboration or the Agency for Healthcare Research and Quality, may include various types of non-randomized study. In the first paper in this series, we described in general terms the perceived advantages and limitations of including NRS in
systematic reviews (Reeves et al., in preparation). In this second paper, we consider in more detail two related areas in the methodology of systematic reviews that include NRS. The first is the approach used to distinguish between different types of NRS. The second is the approach used to assess the risk of bias in a particular study, recognizing that there is a large variety of types of NRS.

When including NRS in systematic reviews of intervention effectiveness (either for harm or for benefit), one consideration is the potential for bias in the NRS compared with randomized trials. Risk of selection bias (understood here as differences in the baseline characteristics of individuals in different intervention groups (Higgins and Altman, 2008)) 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, that is, distributions of participant characteristics differ between groups and the characteristics are associated with outcome. One type of confounding is confounding by indication, whereby a clinical characteristic or medical condition both triggers the use of a particular intervention and is predictive of the outcome (e.g. where the choice of intervention is determined by the severity of disease) (Psaty et al., 1999). Confounding can occur by other mechanisms: for example in a historically controlled comparison, it may be caused by changes in referral pathways over time or changes in the availability of other interventions. Confounding can have two effects in a meta-analysis. If confounding produces biases in one direction, then the overall estimate of the intervention effect will be shifted (systematic bias). If biases vary across studies, then this will lead to increased variability of the observed effects, introducing excessive heterogeneity among the studies (Deeks et al., 2003), and the potential for true effects to be missed. It is important to consider both of these possible effects when undertaking and interpreting a meta-analysis of NRS.

The threat to validity posed by confounding may be less extreme for some research questions. For example, confounding may be less of a problem in studies of long-term or adverse effects (Golder et al., 2011), or studies of some public health, primary prevention, social, educational or crime and justice interventions, where prognostic factors are largely unknown and hence less likely to be indications for administering an intervention.

In the sections that follow, we summarize one strand of discussion at a workshop convened by the Non-Randomised Studies Methods Group (NRSMG) of the Cochrane Collaboration in June 2010. We first describe the methodological issues that were raised and then discuss the practical issues that face authors of systematic reviews. We proceed to propose guidance for review authors and follow this by presenting some of the unresolved issues worthy of further research.

2. Key methodological issues

2.1. Which types of non-randomized studies should be included in a systematic review?

There are many types of NRS design (Shadish et al., 2002). In this series of papers, we define an NRS as any quantitative study estimating the effects of an intervention (harm or benefit) that does not use randomization to allocate units to comparison groups. Examples include controlled trials with non-random allocation mechanisms such as alternation (sometimes called quasi-randomized studies) and designs based on comparisons before and after implementation of an intervention, as well as studies based on situations in which allocation occurs in the course of usual treatment decisions or peoples’ choices. The latter class of studies, frequently referred to as observational, includes many of the classical epidemiological designs such as cohort studies, case–control studies and cross-sectional studies, as well as analyses of clinical databases, possibly using statistical techniques such as propensity scores to attempt to control for confounding. An important distinction, often overlooked, is between studies that form comparison groups by classifying individuals and those that form them by classifying clusters of individuals (see Box 1).

An important stage in a systematic review is the decision as to which types of studies are eligible for inclusion. One possibility is to include any study that addresses the research question, irrespective of its research design, and to evaluate the limitations of each study identified. This is likely to be resource intensive, and efforts may be wasted on evaluating studies for which the risk of bias is too high for them to be informative. Risk of bias depends on how studies are carried out, including their design and their conduct. Therefore, for most research questions, eligibility criteria need to be used to limit the kinds of evidence included in a systematic review.

Many systematic reviews on the effects of interventions restrict their attention to randomized trials. Although there are several types of randomized design (including randomized cross-over trials, cluster-randomized trials and randomized factorial trials), the use or not of randomization is a dichotomy that is usually reasonably clear (although not always (Wu et al., 2009)), as are distinctions between specific types of
Clustering of observations by practitioner, or other unit of healthcare provision, is important because the statistical assumption that all observations are independent may not necessarily hold (Ukoumunne et al., 1999). Ignoring issues of clustering leads to confidence intervals for treatment effect estimates that are too narrow and p-values that are too small.

This distinction between level of allocation is most clearly seen in randomized trials, where it should be clear from the description of randomization that the unit being randomized is a cluster (e.g. practitioner, clinic, family or geographical area), rather than an individual. The issue of cluster-level allocation or assignment also applies to NRS, but it may be less obvious and not acknowledged or taken into account. The need for several clusters in the control and intervention groups is equally true for NRS as it is for cluster-randomized trials.

In NRS, clustering of patients within practitioners is potentially an issue even if practitioners provide both control and intervention treatments, that is, with individual-level allocation. Without the protection of randomization (randomized trials often stratify randomization by practitioners, and the effects of clustering are balanced across groups), there can be a differential effect of clustering by intervention/control group, and practitioners can become a confounding factor.

Given the possible non-random ways in which groups can be formed, it can be seen that almost every NRS will involve one or other type of clustering. The only obvious exception is when participants are classified into intervention or comparator groups on the basis of their personal choices, for example, for nutritional supplements or vitamins that are available over-the-counter, as is often the case for aetiological exposures.

randomized design. For NRS, the distinctions between different types of studies are much less precise, making the specification of eligible study types a challenge (Hartling et al., 2011). For example, even the commonly used term case-control study covers a wide variety of types of study with different strengths and weaknesses. For instance, the label could be applied to a traditional retrospective study in which cases and controls are separately identified and individuals interviewed about past exposures; to case-control studies nested within cohorts, in which past exposures have already been collected; and even to an analysis of a cross-sectional study in which participants are divided into cases and controls and cross-tabulated against some concurrently collected information about previous exposure. To make decisions on eligibility of NRS for a systematic review, different study types need to be distinguished from each other, and some sort of taxonomy will be needed to achieve this.

The workshop discussed the notion of absolute thresholds versus conditional criteria for including NRS in the review (Reeves et al., in preparation). The former would include any study meeting an absolute threshold, irrespective of whether any studies exist. This is the standard approach for Cochrane reviews and leads to some reviews that are ‘empty’ of included studies. The latter would include the best available evidence, so that, for example, if randomized trials are not found, then some types of NRS would be sought and if these are not found, then some other types of NRS would be sought. It quickly became apparent that setting this distinction at the level of a review was too simplistic and that review authors might need to set different eligibility criteria for different research questions within a review. One underlying reason for this is the differential risk of selection bias (confounding) for different outcomes that might be considered: for example, selection bias may be lower for a serious but rare harm (Vandenbroucke and Psaty, 2008). Studies addressing a primary prevention intervention, (Jefferson et al., 2005a; Jefferson et al., 2005b) or using administrative decisions to allocate organizational interventions, may be at lower risk of selection bias if diagnostic information is not available or is disregarded at the point of allocation.

2.1.2. Design labels or design features. Box 2 lists some commonly used design terms used to describe NRS. However, as we note earlier, many of these terms refer to a large variety of specific study designs, and the terms could prove problematic when used to specify unambiguous eligibility criteria. A particularly ambiguous pair of design labels is the term prospective study and the term retrospective study (Feinstein, 1985). At the extreme definition, prospective study should imply that all design aspects were planned. In practice, the degree of prospectiveness is more subtle, with the possibility of different aspects of the design being retrospective (e.g. recruitment of participants, hypothesis generation, outcome data and baseline data collection) and the concomitant potential for bias in those designs varies. There are, therefore, major inconsistencies in the reporting of studies as prospective and retrospective in the literature.
Box 2: Some traditional design labels for non-randomized studies.

<table>
<thead>
<tr>
<th>Design Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-control study</td>
<td>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.</td>
</tr>
<tr>
<td>Case series</td>
<td>Observations are made on a series of individuals, usually all receiving the same intervention, before and after an intervention but with no control group.</td>
</tr>
<tr>
<td>Cohort study</td>
<td>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.</td>
</tr>
<tr>
<td>Concurrently controlled study</td>
<td>A study that compares a group of participants with a contemporary control group, possibly using experimental (but non-random) allocation of interventions to individuals.</td>
</tr>
<tr>
<td>Cross-sectional study</td>
<td>A study that collects information on interventions (past or present) and current health outcomes, that is, 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.</td>
</tr>
<tr>
<td>Historically controlled study</td>
<td>A study that compares a group of participants receiving an intervention with a similar group from the past who did not.</td>
</tr>
<tr>
<td>Time series study</td>
<td>A study that compares measurements made before and after implementation of an intervention, possibly with a control group that does not receive the intervention.</td>
</tr>
</tbody>
</table>

In addition to the broad nature of many of these terms, simple design labels can be interpreted in different ways by different people or in different areas of application. For example, a design in which each individual receives two or more interventions may be known variously as a repeated-treatment design (psychology), a cross-over study (clinical trials) or a switchback study (agriculture), among other terms. Studies in which observations are made in two periods, before and after introducing the intervention of interest in some but not all participants, may be known as cohort controlled before and after studies (education and sociology), controlled before and after studies (health organization) or difference in differences studies (economics). Reliance on simple design labels can lead to potential confusion either among review author teams or among readers of reviews (Hartling et al., 2011).

Design labels have been used to classify studies within hierarchies of evidence, such that they can be ranked according to perceived risk of bias. Existing evidence hierarchies for studies of effectiveness (Eccles et al., 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, primarily on the basis of the fact that the temporal relationship between exposure and outcome is established. In practice, it is often unclear which biases have the greatest impact on effect size and how the biases vary according to the way in which studies are carried out. The precedence given to features indicating causality may not always be appropriate when the priority is unbiased quantification of the effect size for an intervention. A decision on which types of NRS to include in a systematic review might be better made using a fitness for purpose rather than a hierarchical paradigm (Tugwell et al., 2010), although such a decision may mean that additional resources are required for the tasks of bias assessment, data extraction and synthesis.

Using simple design labels, such as those in Box 2, to distinguish between NRS may lack the specificity necessary to determine whether a study is eligible and may fail to distinguish between studies at importantly different risks of bias. An alternative approach is to focus on particular features of the study. Indeed, it is the specific design features of a study, rather than its broad design, that determines its strength. The Cochrane NRSMG has proposed the use of four questions to help clarify the important features of a NRS:

1. Was there a comparison of outcomes following intervention and comparator?
2. How were study participants formed into groups?
3. Which parts of the study were carried out after the study was conceived?
4. On which variables was comparability between groups of different people receiving the intervention or the comparator assessed?

The NRSMG have organized potential responses to these questions into two lists, one for studies in which individuals are classified into intervention and comparator groups (Box 3) and one for studies in which clusters of individuals are classified into intervention and comparator groups (Box 4). Instructions for assigning features to studies are described.
in Box 5. The items are designed to characterize key features of studies that, on the basis of the experiences of NRS MG members and first principles (rather than evidence), are suspected to define the major study design categories or to be associated with varying risk of bias. The Cochrane Effective Practice and Organisation of Care Group uses design labels to specify which studies are eligible for inclusion in effectiveness reviews (Mowatt et al., 2001). However, explicit definitions for these are based on study design features; for example, a controlled before-and-after study is defined as one that has contemporaneous data collection in the control and intervention sites before and after implementation of an intervention, and a minimum of two control and two intervention sites (Cochrane Effective Practice and Organisation of Care Group, 2002).

**Box 3: List of study design features (studies formed by classifying individuals by intervention and comparator)**

<table>
<thead>
<tr>
<th><strong>Was there a relevant comparison:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Between two or more groups of participants receiving different interventions?</td>
</tr>
<tr>
<td>Within the same group of participants over time?</td>
</tr>
<tr>
<td><strong>Were groups of individuals formed by:</strong></td>
</tr>
<tr>
<td>Randomization?</td>
</tr>
<tr>
<td>Quasi-randomization?</td>
</tr>
<tr>
<td>Other action of researchers?</td>
</tr>
<tr>
<td>Time differences?</td>
</tr>
<tr>
<td>Location differences?</td>
</tr>
<tr>
<td>Healthcare decision makers?</td>
</tr>
<tr>
<td>Participants’ preferences?</td>
</tr>
<tr>
<td>On the basis of outcome?</td>
</tr>
<tr>
<td>Some other process? (specify)</td>
</tr>
<tr>
<td><strong>Were the features of the study described below carried out after the study was designed:</strong></td>
</tr>
<tr>
<td>Identification of participants?</td>
</tr>
<tr>
<td>Assessment before intervention?</td>
</tr>
<tr>
<td>Actions/choices leading to an individual becoming a member of a group?</td>
</tr>
<tr>
<td>Assessment of outcomes?</td>
</tr>
<tr>
<td><strong>On which variables was comparability between groups assessed:</strong></td>
</tr>
<tr>
<td>Potential confounders?</td>
</tr>
<tr>
<td>Assessment of outcome variables before intervention?</td>
</tr>
</tbody>
</table>

**Box 4: List of study design features (studies formed by classifying clusters by intervention and comparator)**

Note that cluster refers to an entity (e.g. an organization), not necessarily to a group of participants; group in a cluster-allocated study refers to one or more clusters (see Box 5).

<table>
<thead>
<tr>
<th><strong>Was there a relevant comparison:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Between two or more groups of clusters receiving different interventions?</td>
</tr>
<tr>
<td>Within the same group of clusters over time?</td>
</tr>
<tr>
<td><strong>Were groups of clusters formed by:</strong></td>
</tr>
<tr>
<td>Randomization?</td>
</tr>
<tr>
<td>Quasi-randomization?</td>
</tr>
<tr>
<td>Other action of researchers?</td>
</tr>
<tr>
<td>Time differences?</td>
</tr>
<tr>
<td>Location differences?</td>
</tr>
<tr>
<td>Policy/public health decisions?</td>
</tr>
<tr>
<td>Cluster preferences?</td>
</tr>
<tr>
<td>Some other process? (specify)</td>
</tr>
<tr>
<td><strong>Were the features of the study described below carried out after the study was designed:</strong></td>
</tr>
<tr>
<td>Identification of participating clusters?</td>
</tr>
<tr>
<td>Assessment of before intervention?</td>
</tr>
<tr>
<td>Actions/choices leading to a cluster becoming a member of a group?</td>
</tr>
<tr>
<td>Assessment of outcomes?</td>
</tr>
<tr>
<td><strong>On which variables was comparability between groups assessed:</strong></td>
</tr>
<tr>
<td>Potential confounders?</td>
</tr>
<tr>
<td>Assessment of outcome variables before intervention?</td>
</tr>
</tbody>
</table>
Box 5: Instructions for using the NRSMG design feature checklist

Clarity is needed about the way in which the terms group and cluster are used in Boxes 3 and 4. Box 3 refers only to groups, used in its conventional sense to mean a number of individual participants. Except for groups formed on the basis of outcome, group can be interpreted synonymously with intervention or comparator group. Box 4 refers to both clusters and groups, where clusters are typically an organizational entity such as a family health practice or administrative area, not an individual. As in Box 3, group is synonymous with intervention or comparator group, but in Box 4, the groups are formed by classifying clusters rather than individuals as having the intervention or comparator. 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; these individuals may be identical at both time points, include some individuals at both time points and other individuals at only one time or be formed from completely different individuals.

Was there a relevant comparison?

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

Were groups of individuals/clusters formed 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 a truly random sequence. Check carefully whether allocation was adequately concealed until subjects were definitively recruited.

Quasi-randomization: Allocation was carried out on the basis of a pseudo-random sequence, for example, odd/even hospital number or date of birth, alternation. Note that 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.

Other action of researchers: This is a catch-all category for situations in which allocation happened as the result of some decision or system applied by the researchers. Further details should be noted if the researchers report them. For example, subjects managed in particular units of provision (e.g. wards and general practices) may have been 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 time differences 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 time differences and healthcare decision makers 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 location differences and other action of researchers could be ticked for a single study.

Healthcare decision makers: 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 healthcare decision makers and location or time differences should be ticked.

Participant preferences: Intervention and control groups were formed by naturally occurring variation in participants’ 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 is compared with a group of people who did not, that is, a case–control study. Note that this option should be ticked for papers that report analyses of multiple risk factors for a particular outcome in a large series of subjects, that is, 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

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, for example, preferences made collectively or individually at the level of the cluster entity.

Were the features of the study described below carried out after the study was designed?

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. 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 between groups assessed at baseline?

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

2.1.3. Possible areas of confusion. The confusion around study design labels can be illustrated by secondary analyses of clinical databases or sources of routinely collected data, such as health insurance claims files or administrative databases. The data being analysed are typically collected alongside the delivery of interventions. The participants represented in the databases clearly constitute a cohort in the epidemiological sense. However, many secondary analyses are structured to investigate multiple exposures (on the basis of the available data) for a particular outcome, a situation which more closely resembles a traditional (nested) case–control design than a cohort study. For example, using data from the General Practice Research Database in the United Kingdom, Koro and colleagues were able to assemble a cohort of 19 637 patients who were being treated for schizophrenia (Angrist and Pischke, 2009). Within this group, there were 451 incident cases of diabetes with whom they matched 2696 non-diabetic controls (also from the schizophrenia cohort). They looked at multiple exposures with a particular focus on medications and observed that the risk of developing diabetes was greater with olanzapine than with other anti-psychotic drugs.

Using the NRSMG study design list, the groups in such situations would be described as having been formed on the basis of outcome. It can also be tempting to consider secondary analyses of databases as cross-sectional studies because, from the analyst’s perspective, information about the intervention/comparator and outcome are made available at the same time. However, this view clearly ignores the time dimension that applies to the data collection process, which usually ensures that the intervention or comparator was provided before the outcome was assessed.

These issues around the nature of a secondary analysis of a database also highlight a common difficulty in distinguishing between a study design and an analysis technique. Whereas epidemiologists tend to separate the two, in other disciplines, the distinction can be less clear. For instance, a regression discontinuity design is a distinctive study design, despite also being an analytic approach that can be used in numerous designs including a controlled before-and-after study and randomized cross-over trial. Some NRS are motivated explicitly by the availability of analysis techniques that allow valid inferences to be made, such as instrumental variables (Angrist et al., 1996) and propensity score matching (D’Agostino, 1998). Furthermore, a regression discontinuity design is a distinctive study design, despite also being named after the statistical technique used to analyse it (Thistlethwaite and Campbell, 1960).

2.2. Approaches to assessing the risk of bias in non-randomized studies

Whatever the eligibility criteria for including studies in a review, most systematic reviews—and all Cochrane reviews—include an assessment of the risk of bias in each individual primary study. This should consider both the likely direction and the magnitude of important biases, recognizing that there will usually be considerable uncertainty in these assessments. Considerations for assessing risk of bias in NRS are similar to those for randomized trials. However, potential biases are likely to be greater for NRS compared with randomized trials. Assessing the magnitude of confounding in NRS is especially problematic. Differences in how researchers attempt to control for confounding in NRS can introduce a source of heterogeneity not usually present in reviews that only include randomized trials (Valentine and Thompson, in press). A key difficulty is that studies may not have collected data on important confounders, so even a re-analysis of the raw data in an individual participant data meta-analysis cannot overcome this heterogeneity.
An assessment of each study should consider the following: (i) the inherent strengths and weaknesses of the designs that have been used, such as noting their potential to ascertain valid effect estimates for beneficial and for adverse outcomes; (ii) the execution of the studies, through a careful assessment of their specific conduct and, in particular, the potential for selection bias and confounding to which all NRS are suspect; and (iii) the potential for reporting biases, including those arising from selective reporting of outcomes and analyses (Norris et al., in press).

Not all NRS are equally at risk of bias. For example, investigations of adverse effects of pharmaceuticals may be reasonably free of bias when the biological pathway through which an adverse effect arises is different from that through which the intended effects of an intervention operate (Vandenbroucke and Psaty, 2008; Golder et al., 2011). Acceptance of causality between the intervention and the adverse event is based upon rejection of the idea that the observed difference in frequency of adverse events is due to bias or confounding.

Decisions based on evidence about the harms of an intervention may not always require precise and unbiased estimation of an effect size. It may be sufficient to establish beyond reasonable doubt that the intervention causes the harm being investigated, for example if the most favourable extreme of a confidence interval clearly indicates a large harmful effect that cannot be attributed to bias, or if a clear dose response effect is demonstrated and the intervention can be related to the harm through a known biological pathway (Hill, 1965). The confidence required to make a decision in the absence of a precise and unbiased estimate of the effect size will depend on the context for the research question. For example, if multiple drugs achieve similar benefits, only moderate suspicion about harm associated with one drug may lead to its withdrawal, whereas much higher quality evidence about harm would be needed to withdraw an orphan drug intended for only a small number of patients suffering from a very rare condition.

Can the magnitude and direction of bias be predicted? This is a subject of ongoing research that is attempting to gather empirical evidence on factors such as study design and intervention type that influence 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 et al., 2001; Deeks et al., 2003). However, studies that have compared randomized with NRS of the same question have not provided a consistent message. One comprehensive review examined eight studies that had made such comparisons across multiple interventions (Deeks et al., 2003). The findings of these studies were contradictory. The authors of the review identified weaknesses in each study, which they believed to be responsible for most of the divergence.

Many instruments for assessing methodological quality of NRS of interventions have been created, and those available in 2003 were reviewed systematically by Deeks et al. (2003). They located 182 tools, which they reduced to a shortlist of 14, and identified six as potentially useful for systematic reviews that ‘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 would need modification because they did not require assessors to extract detailed information about how study groups were formed, which is likely to be critical for risk of selection bias. 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 the review were the Downs and Black instrument and the Newcastle–Ottawa Scale (Downs and Black, 1998; Wells et al., 2008).

Methodological reviews have implicated design features of randomized trials such as random sequence generation and allocation concealment as sources of important bias. A new methodological review has recently been carried out (Bias in Randomised AND Observational studies, or BRANDO) combining data from all previous methodological reviews of the bias arising from sub-optimal features of randomized trial design, that is, sequence generation, allocation concealment and blinding (Savovic et al., in press). This review observed that the influence of these features depended on whether or not outcomes were objective (mainly all-cause mortality but including other outcomes such as resource use and withdrawals). This finding was unexpected and suggests that the features may be proxy markers of more general methodological quality. Importantly, the estimates of bias associated with different study features that have been obtained from initiatives such as BRANDO may provide an empirical basis for down-weighting potentially biased studies when combining them with less biased studies (Welton et al., 2009).

The Cochrane Collaboration has developed a tool that review authors are expected to use for assessing risk of bias in randomized trials. This involves consideration of six bias domains: selection bias, performance bias, detection bias, attrition bias, reporting bias and other bias. Specific items are addressed within these domains (e.g. random sequence generation and allocation concealment are addressed under selection bias). Items are assessed by (i) providing a judgement on the risk of important bias that might arise from the item (low risk, high risk and unclear risk) and (ii) providing a narrative explanation to provide support for the judgement (Higgins and Altman, 2008). The tool was not developed with NRS in mind, and the existing guidance for using it is not appropriate for all NRS. The Cochrane NRSMG and the Cochrane Effective Practice and Organisation of Care Review Group have both modified the tool to address some NRS designs, such as non-randomized controlled studies, controlled before-and-after designs and interrupted time series studies. A project to extend the existing risk of bias tool to NRS has recently been funded by The Cochrane Collaboration.
Risk of bias assessments for some NRS will need to appraise whether studies satisfy the conditions underpinning the design. For example, consider the regression discontinuity design, in which intervention and comparator groups are formed on the basis of a specified cut-off on a baseline predictor of outcome. Moss and Yeaton (2006) used this design to evaluate a developmental English course; students were required to take an English placement exam before they could take college-level English, and those with exam scores below a specified threshold were required to take the developmental course. In this example, the placement exam is the predictor of the outcome of interest, and the outcome of interest is the performance on the college-level English course. The developmental course was found to have a more positive effect on English achievement among those who had poorer scores on the placement exam; as placement exam scores approached the cut-score, the effect was smaller. This design tests for a change in level or slope of the association between pre-test and outcome at the specified pre-test cut-off as evidence of the effect of the intervention. Because the selection process is completely known, this is a strong research design. However, for the inferences regarding the effects of the intervention to be valid, several conditions need to be met. One of these conditions is that the relationship between the pre-test variable and the outcome in the absence of the intervention is known and can be specified. Violations can be hard to diagnose in the absence of the original data. Therefore, review authors who include regression discontinuity designs will need to develop specific items to assess this condition. The implication is that, because the category of NRS is so large, one assessment tool is unlikely to adequately address concerns across all of these designs, even if the general domains of potential bias are the same.

3. Practical considerations for review authors

Setting eligibility criteria for including primary studies is a crucial step in a systematic review. Criteria should be justified not only by the research question but also with consideration of the issues discussed earlier and described in the first paper in this series, which recommends consideration of the likely consequences of disseminating a systematic review on the basis of the chosen evidence (Reeves et al, in preparation). Here, we set out some further principles to guide the choice of eligibility criteria regarding study designs. We anticipate that review authors will still find this step difficult, not least because there are likely to be resource implications as well as considerations about risk of bias as we discuss in Section 4.

The practical consequences of searching for NRS can be far reaching and should be considered carefully by review authors. Whichever studies review authors decide should be included, there are currently no sensitive search filters for identifying studies on the basis of design labels or design features. If review authors want to search comprehensively for all eligible primary studies, the way to do this currently is usually by searching for any study evaluating the intervention of interest in relevant populations. For review questions about specific outcomes such as particular harms, review authors could limit their searches by further specifying the outcome(s) of interest (Reeves et al., 2008).

Study design labels may offer the possibility of designing search filters to identify studies by using particular design features, such as case–control studies. However, the design of such a filter will need to take into account the multiple ways in which studies using common design features can be described, the inappropriate use of study design labels and the fact that some researchers do not describe their studies with a commonly-used label. Applying design-related eligibility criteria at the initial screening of citations identified by searching may not be possible from abstracts alone because the abstracts may not report sufficient detail about the methods used, that is, even this step of the review may require review authors to consult full text papers. The likely lack of specificity of the searching, and the need for full papers to carry out the initial screen of citations, may have profound consequences on the resources needed to carry out a review.

The concept of an unpublished study may be difficult to apply to NRS. Depending on the definition of eligibility, the ‘universe’ of eligible studies may be difficult to define compared with the situation with randomized trials. This is largely because many NRS are conducted using data that have been collected for other purposes. Furthermore, NRS protocols are rarely written or published, and very few are prospectively registered in databases of ongoing research. In these circumstances, it is not clear that comprehensive searching represents the optimal strategy (Reeves et al., 2008).

Assessment of the risk of bias in NRS is more complicated than in randomized trials. However, a comprehensive methodological assessment of NRS is critical in a systematic review to fully document the weaknesses and uncertainties in the evidence. Methodologists are starting to develop tools to help, but the practical issue will remain that such tools will likely demand considerable epidemiological skills as well as an awareness of the topic area.

4. Guidance for review authors

Review authors must ensure that the review team includes authors with the required content and methodological expertise. At a minimum, the team must include the necessary methodological expertise in assessing risk of bias.
and correctness of analysis in the designs to be included. Thus, statistical and epidemiological skills are needed in addition to content expertise and information science expertise. For a review question about the harms of a drug, the team should also include pharmacoepidemiological expertise.

Eligibility criteria for study inclusion must be explicitly justified and should be stated unambiguously. Criteria may revolve around design features (as in Boxes 3 and 4) or design labels (as in Box 2). We prefer to concentrate on design features. Although simple study design labels can provide an easy way to communicate, they may not provide sufficient information to understand the design that was used. If study design labels are used, review authors should provide explicit descriptions of the key design features that they consider such studies to have. We recommend that review authors routinely map study design features on to their preferred labels, using the features described in Boxes 3 and 4. The use of ambiguous design labels such as *prospective study* and *retrospective study* should be avoided. Common design labels such as *cohort study*, *case–control study* and *interrupted time series* should be used with caution and always be accompanied by clear definitions; as they stand, these labels almost certainly do not give sufficient information about how a study was conducted. When assessing eligibility of specific studies against eligibility criteria, we recommend use of the guidance in Box 5.

The justification for which studies to include in a review should discuss the context of the review and the likely bias arising from including or excluding certain types of evidence. We propose four general considerations underpinning the choice of eligibility criteria for study design:

- The likelihood of randomized trials being carried out. This consideration may provide sound justification for including NRS if the review question is about unintended, rare (often harms), or long-term outcomes (benefit or harm), some organizational interventions and interventions that cannot be randomized for ethical reasons.
- The extent to which studies are resistant to performance bias (bias due to systematic differences between groups in care provided or exposure to other factors) and detection bias (bias due to systematic differences between groups in how outcomes are determined). This consideration may provide sound justification for including NRS if the review question is about objective outcomes, if data were collected before the hypothesis was conceived or if the evaluation incorporates a measure of the fidelity of the intervention. However, it is often difficult to assess performance and detection biases in NRS. In the absence of a protocol, delivery of interventions may be highly variable, misclassified and poorly reported; and outcome assessments may be poorly planned and not standardized in approach or timing.
- The extent to which studies are resistant to selection bias (bias due to systematic differences in baseline characteristics of the groups being compared). This consideration may provide sound justification for including NRS if the review question is about harms not on the same biological pathway or affecting a different organ system from the benefits of the intervention, is about primary prevention or is about an intervention where the factors predictive of the outcome(s) of interest are not known.
- The extent to which randomization might undermine the ability to perform a study of the research question.

There was no consensus at the workshop about whether eligibility criteria should be *absolute* (sticking rigidly to a pre-specified criterion) or *conditional* (modified sequentially until the best available evidence is included). The need to inform an imminent policy decision might provide a rationale for including the best available evidence, thus including studies that might otherwise be considered to be at too high a risk of bias. If review authors do take this approach, then the sequential strategy should be specified in advance, and appropriate warnings should be provided that the risk of bias may mean that the included evidence is misleading.

Review authors must assess the risk of bias in each included study, including the risk of confounding. The general domains for such assessments of NRS and randomized experiments are the same. The existing Cochrane risk of bias tool can in principle be used for everything other than case control studies, but the *Cochrane Handbook* provides guidance mainly for making judgements about randomized trials. The Cochrane tool classifies items as putting a study at low, unclear or high risk of bias. Three categories may be insufficient when dealing with NRS because there is unlikely to be sufficient discrimination between NRS at varying risk of bias.

To avoid the misuse, by readers of the review, of evidence that is at high risk of bias, it is important that assessments of risk of bias are presented clearly and are linked directly with the results of any analyses involving such evidence. One convenient way to achieve this is to present a summary of findings table, placing the numerical results alongside assessments of the quality of the body of evidence, for example, using the GRADE system (Schünemann *et al.*, 2008).

#### 5. Research priorities

##### 5.1. Study design taxonomy and design labels

An agreed framework is required for defining study design labels, perhaps building on the cross-tabulation of design features and labels described in the NRSMG Handbook chapter (Reeves *et al.*, 2008). The steps involved will include the following:
• Developing explicit definitions for designs based on features that are commonly used to estimate the benefits and harms of interventions.
• Presenting label synonyms (e.g. used by different disciplines) as appropriate.
• Developing a design feature checklist to ensure that options cover study designs used by different disciplines (e.g. regression discontinuity design) and modifying instructions to clarify that some commonly used terms are statistical methods to achieve control for confounding (e.g. use of an instrumental variable or propensity scores), rather than study designs per se.
• Considering whether there are contexts in which a hierarchical taxonomy of study designs, defined by features and varying in risk of bias, can be recommended.

5.2. Tools for review authors to assess the risk of bias in non-randomized studies

There was no consensus at the workshop on how judgements of risk of bias might be formulated and utilized. Some participants at the meeting felt that it was important that summary judgements about risks of bias should be made in a similar way to judgements about bias in randomized trials, to facilitate the comparison of the risk of bias across all studies, randomized or non-randomized. Others felt that each study should be evaluated only in the context of the particular design features used and the purpose of the study. Developing a tool for assessing the risk of bias in NRS could be tackled in two ways: (i) by extending an existing tool for randomized trials or (ii) by developing a tool specific to NRS or specific to studies with common design features. There is also a need to consider whether a generic framework for assessing the risk of bias in systematic reviews including NRS is required, especially if adopting the latter approach. A starting point for a framework could reflect the approach used in a systematic review of the effects of male circumcision on acquisition of HIV. The authors developed a tool that covers the domains of selection bias/confounding, performance bias, detection bias and attrition bias (similar domains to the Cochrane risk of bias tool) but developed different criteria for assessing them to address three types of study design (Siegfried et al., 2005).

Whilst there was no consensus on the exact approach that should be used for incorporating risk of bias assessments in reviews, a consistent generic framework is being applied by different review organizations. The framework is summarized in Box 6. The nature of the review question, the context of each study and the nature of each study are critical. The intensity of an assessment of the risk of bias is likely to depend on whether a quantitative synthesis is proposed because it is particularly undesirable to combine results of studies that are at widely varying risk of bias. The need to consider the risk of bias to the totality of the evidence base, as well as to each individual study, applies in the same way when including NRS (Schünemann et al., 2008).

Box 6: A framework for considering risk of bias in systematic reviews including NRS

1 Determine the nature of the question
   • For example, a question about harms or benefits, of comparative effectiveness or absolute effectiveness

2 Determine the nature of the eligible studies
   • Exploratory or confirmatory purpose
   • Emphasize design features rather than simple design labels

3 Pre-specify key sources of bias in primary studies
   • Determines key sources of potential bias and contents of any bias assessment tool
   • Must include pre-specifying of likely sources of confounding by indication

4 After including studies in the review, assess risk of bias in the findings of each included study
   • Document the principle sources of likely bias that are likely to affect the study question of interest
   • Any of the tools mentioned in this manuscript could be used
   • May assess adequacy of adjustments for confounders, implementation of specialized techniques and so forth

5 Integrate concerns into the review’s findings and conclusions
   • Document the principle sources of likely bias for the totality of evidence across studies for each question of interest, including pervasive reporting biases
   • May include presentation of a summary of findings table or use of a formal grading scheme (e.g. GRADE) for the overall quality of the body of evidence (Schünemann et al., 2008)
Part of the framework outlined in Box 6 is a detailed assessment of the risk of bias in a specific study, an assessment that needs to be informed by the nature and purpose of the study. The Cochrane Collaboration’s risk of bias tool, developed so far for randomized trials, provides a template that could be extended to any study design. The tool has the following core characteristics (Higgins and Altman, 2008) that could be carried forward:

- It is tailored to the type of study design.
- It identifies key domains of bias for the study type in question (for randomized trials, these are selection bias, performance bias, detection bias, attrition bias and reporting bias).
- It allows for other design, review question or health setting-specific biases through an other bias domain.
- It assesses the risk of important bias within each of the bias domains, by selecting from a list of judgements (e.g. low risk, high risk and unclear risk).
- It provides free-text explanations, preferably including quotations from reports of the primary study, in support of each of judgment.
- It assesses risk of bias at the level of the outcome rather than the study, recognizing that different biases may operate on different types of outcomes.

Specific research tasks may include the following:

- Develop a decision aid (as outlined in Box 6) to guide authors through assessment of (i) study design and (ii) risk of bias for specific design (and specific context).
- Identify additional items that may be required when assessing the risk of bias in studies with specific study design features.

5.3. Defining eligibility to avoid empty reviews

We previously described how there was no consensus about whether systematic reviews should routinely include studies that might otherwise be considered to be at too high a risk of bias if studies with design features less likely to give rise to bias are not available. We think that it is important to try to obtain evidence to inform whether such an approach leads, on balance, to net benefit for society. There may be a case for independent critical appraisal of a sample of systematic reviews carried out in these circumstances alongside the policy decisions that were taken on the basis of the review. If sufficient reviews were available, it might be possible to compare situations in which the policy decision was deemed appropriate with situations in which the decision was not deemed appropriate, to see whether the risks of bias in reviews varied systematically.

6. Conclusions

It can be important for systematic reviews of the effects of interventions to include NRS. Different disciplines may use different study design labels, but these are not always used consistently, even within disciplines. Therefore, review authors need to define eligibility criteria for including studies in a review on the basis of explicit study design features, even if these features are referred to by a study design label. The NRSMG has developed a checklist of study design features that are believed to be the most important ones.

The existing evidence based on assessing the risk of bias, and the consequences of high risk of bias, comes almost exclusively from methodological reviews examining randomized trials. Extending risk of bias assessment to NRS requires both a general framework to guide review authors through the assessments and specific additional items, in particular an item for confounding. We have described in this paper how this research agenda may be addressed.

Acknowledgements

We are grateful to all the workshop participants (see the Appendix to Paper 1); all of whom contributed to the discussions that provided the foundation for this paper. The workshop was supported financially by the Agency for Healthcare Quality and Research and by a grant from the Cochrane Collaboration Discretionary Fund. JPTh is supported by the Medical Research Council (Unit Programme number U105285807). The Health Services Research Unit is funded by the Scottish Government Executive Health Department. BCR is supported in part by the U.K. National Institute for Health Research Bristol Cardiovascular Biomedical Research Unit. The views expressed in this article are those of the authors and not necessarily those of the funding bodies, The Cochrane Collaboration or its registered entities, committees or working groups or the Campbell Collaboration. Some authors contribute to the Cochrane (JH, CR, BR, JD, BS, PT, GW) and Campbell Collaborations (JH, JV, PT, GW).

References


