Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance

Edited by Jonathan AC Sterne, Julian PT Higgins, Roy G Elbers and Barney C Reeves
on behalf of the development group for ROBINS-I

Updated 20 October 2016

To cite the ROBINS-I tool: Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JPT. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. *BMJ* 2016; **355**; i4919.

To cite this document: Sterne JAC, Higgins JPT, Elbers RG, Reeves BC and the development group for ROBINS-I. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance, updated 12 October 2016. Available from http://www.riskofbias.info [accessed {date}]

Contents

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1 Contributors

(Listed alphabetically within category)

Core group: Julian Higgins, Barney Reeves, Jelena Savović, Jonathan Sterne, Lucy Turner.

Additional core research staff: Roy Elbers, Alexandra McAleenan, Matthew Page.

Bias due to confounding: Nancy Berkman, Miguel Hernán, Pasqualina Santaguida, Jelena Savović, Beverley Shea, Jonathan Sterne, Meera Viswanathan.

Bias in selection of participants into the study: Nancy Berkman, Miguel Hernán, Pasqualina Santaguida, Jelena Savović, Beverley Shea, Jonathan Sterne, Meera Viswanathan.

Bias due to departures from intended interventions: David Henry, Julian Higgins, Peter Jüni, Lakhbir Sandhu, Pasqualina Santaguida, Jonathan Sterne, Peter Tugwell.

Bias due to missing data: James Carpenter, Julian Higgins, Terri Piggott, Hannah Rothstein, Ian Shrier, George Wells.

Bias in measurement of outcomes or interventions: Isabelle Boutron, Asbjørn Hróbjartsson, David Moher, Lucy Turner.

Bias in selection of the reported result: Doug Altman, Mohammed Ansari, Barney Reeves, An-Wen Chan, Jamie Kirkham, Jeffrey Valentine.

Cognitive testing leads: Nancy Berkman, Meera Viswanathan.

Piloting and cognitive testing participants: Katherine Chaplin, Hannah Christensen, Maryam Darvishian, Anat Fisher, Laura Gartshore, Sharea Ijaz, J Christiaan Keurentjes, José López-López, Natasha Martin, Ana Marušić, Anette Minarzyk, Barbara Mintzes, Maria Pufulete, Stefan Sauerland, Jelena Savović, Nandi Seigfried, Jos Verbeek, Marie Wetwood, Penny Whiting.

Other contributors: Belinda Burford, Rachel Churchill, Jon Deeks, Toby Lasserson, Yoon Loke, Craig Ramsay, Deborah Regidor, Jan Vandenbroucke, Penny Whiting.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

2 Background

The goal of a systematic review of the effects of an intervention is to determine its causal effects on one or more outcomes. When the included studies are randomized trials, causality can be inferred if the trials are methodologically sound, because successful randomization of a sufficiently large number of individuals should result in intervention and comparator groups that have similar distributions of both observed and unobserved prognostic factors. However, evidence from randomized trials may not be sufficient to answer questions of interest to patients and health care providers, and so systematic review authors may wish to include non-randomized studies of the effects of interventions (NRSIs) in their reviews.

Our ROBINS-I tool ("Risk Of Bias In Non-randomized Studies - of Interventions") is concerned with evaluating the risk of bias (RoB) in the results of NRSIs that compare the health effects of two or more interventions. The types of NRSIs that can be evaluated using this tool are quantitative studies estimating the effectiveness (harm or benefit) of an intervention, which did not use randomization to allocate units (individuals or clusters of individuals) to comparison groups. This includes studies where allocation occurs during the course of usual treatment decisions or peoples' choices: such studies are often called "observational". There are many types of such NRSIs, including cohort studies, case-control studies, controlled before-and-after studies, interrupted-time-series studies and controlled trials in which intervention groups are allocated using a method that falls short of full randomization (sometimes called "quasi-randomized" studies). This document provides guidance for using the ROBINS-I tool specifically for studies with a cohort-type of design, in which individuals who have received (or are receiving) different interventions are followed up over time.

The ROBINS-I tool is based on the Cochrane RoB tool for randomized trials, which was launched in 2008 and modified in 2011 (Higgins et al, 2011). As in the tool for randomized trials, risk of bias is assessed within specified bias domains, and review authors are asked to document the information on which judgements are based. ROBINS-I also builds on related tools such as the QUADAS 2 tool for assessment of diagnostic accuracy studies (Whiting et al, 2011) by providing signalling questions whose answers flag the potential for bias and should help review authors reach risk of bias judgements. Therefore, the ROBINS-I tool provides a systematic way to organize and present the available evidence relating to risk of bias in NRSI.

2.1 Context of the tool

Evaluating risk of bias in a systematic review of NRSI requires both methodological and content expertise. The process is more involved than the process of evaluating risk of bias in randomized trials, and typically involves three stages.

- First, at the planning stage, the review question must be clearly articulated, and important potential problems in NRSI should be identified. This includes a preliminary specification of key confounders (see the discussion below Table 1, and section 4.1) and co-interventions (see section 4.4).
- Second, each study should be carefully examined, considering all the ways in which it might be put at risk of bias. The assessment must draw on the preliminary considerations, to identify important issues that might not have been anticipated. For example, further key confounders, or problems with definitions of interventions, or important co-interventions, might be identified.
- Third, to draw conclusions about the extent to which observed intervention effects might be causal, the studies should be compared and contrasted so that their strengths and weaknesses can be considered jointly. Studies with different designs may present different types of bias, and "triangulation" of findings across these studies may provide assurance either that the biases are minimal or that they are real.

This document primarily addresses the second of these stages, by proposing a tool for assessing risk of bias in a NRSI. Some first-stage considerations are also covered, since these are needed to inform the assessment of each study.

2.2 Assessing risk of bias in relation to a target trial

Both the ROBINS-I tool and the Cochrane RoB tool for randomized trials focus on a study's **internal validity**, For both types of study, we define bias as a tendency for study results to differ systematically from the results expected from a randomized trial, conducted on the same participant group that had no flaws in its conduct. This would typically be a large trial that achieved concealment of randomized allocation; maintained blinding of

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

patients, health care professionals and outcome assessors to intervention received throughout follow up; ascertained outcomes in all randomized participants; and reported intervention effects for all measured outcomes. Defined in this way, bias is distinct from **issues of generalizability (applicability or transportability)** to types of individual who were not included in the study. For example, restricting the study sample to individuals free of comorbidities may limit the utility of its findings because they cannot be generalized to clinical practice, where comorbidities are common.

Evaluations of risk of bias in the results of NRSIs are therefore facilitated by considering each NRSI as an attempt to emulate (mimic) a hypothetical trial. This is the hypothetical pragmatic randomized trial that compares the health effects of the same interventions, conducted on the same participant group and without features putting it at risk of bias (Hernán 2011; Institute of Medicine 2012). **We refer to such a hypothetical randomized trial as a "target" randomized trial** (see section 3.1.1 for more details). Importantly, a target randomized trial need not be feasible or ethical.

ROBINS-I requires that review authors explicitly identify the interventions that would be compared in the target trial that the NRSI is trying to emulate. Often the description of these interventions will require subject-matter knowledge, because information provided by the investigators of the observational study is insufficient to define the target trial. For example, authors may refer to "use of therapy [A]," which does not directly correspond to the intervention "initiation of therapy [A]" that would be tested in an intention-to-treat analysis of the target trial. Meaningful assessment of risk of bias is problematic in the absence of well-defined interventions. For example, it would be harder to assess confounding for the effect of obesity on mortality than for the effect of a particular weight loss intervention (e.g., caloric restriction) in obese people on mortality.

To keep the analogy with the target trial, **this document uses the term "intervention" groups to refer to "treatment" or "exposure" groups in observational studies** even though in such studies no actual intervention was implemented by the investigators.

2.3 Domains of bias

The ROBINS-I tool covers seven domains through which bias might be introduced into a NRSI. These domains provide a framework for considering any type of NRSI, and are summarized in Table 1. The first two domains address issues before the start of the interventions that are to be compared ("baseline") and the third domain addresses classification of the interventions themselves. The other four domains address issues after the start of interventions. For the first three domains, risk of bias assessments for NRSIs are mainly distinct from assessments of randomized trials because randomization protects against biases that arise before the start of intervention. However, randomization does not protect against biases that arise after the start of intervention. Therefore, there is substantial overlap for the last four domains between bias assessments in NRSI and randomized trials.

Variation in terminology between contributors and between research areas proved a challenge to development of ROBINS-I and to writing guidance. The same terms are sometimes used to refer to different types of bias, and different types of bias are often described by a host of different terms. Table 1 explains the terms that we have chosen to describe each bias domain, and related terms that are sometimes used. The term *selection bias* is a particular source of confusion. It is often used as a synonym for *confounding* (including in the current Cochrane tool for assessing RoB in randomized trials), which occurs when one or more prognostic factors also predict whether an individual receives one or the other intervention of interest. We restrict our use of the term selection bias to refer to a separate type of bias that occurs when some eligible participants, or the initial follow up time of some participants, or some outcome events, are excluded in a way that leads to the association between intervention and outcome differing from the association that would have been observed in complete follow up of the target trial. *We discourage the use of the term selection bias to refer to confounding*, although we have done this in the past, for example in the context of the RoB tool for randomized trials. Work is in progress to resolve this difference in terminology between the ROBINS-I tool and the current Cochrane tool for assessing RoB in randomized trials.

By contrast with randomized trials, in NRSIs the characteristics of study participants will typically differ between intervention groups. The assessment of the risk of bias arising from uncontrolled confounding is therefore a major component of the ROBINS-I assessment. **Confounding** of intervention effects occurs when one or more prognostic factors (factors that predict the outcome of interest) also predict whether an individual receives one or the other intervention of interest. As an example, consider a cohort study of HIV-infected patients that

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

compares the risk of death from initiation of antiretroviral therapy A versus antiretroviral therapy B. If confounding is successfully controlled, the effect estimates from this observational study will be identical, except for sampling variation, to those from a trial that randomly assigns individuals in the same study population to either intervention A or B. However, failure to control for key confounders may violate the expectation of comparability between those receiving therapies A and B, and thus result in bias. A detailed discussion of assessment of confounding appears in section 4.1

Selection bias may arise when the analysis does not include all of the participants, or all of their follow-up after initiation of intervention, that would have been included in the target randomized trial. The ROBINS-I tool addresses two types of selection bias: (1) bias that arises when either all of the follow-up or a period of follow-up following initiation of intervention is missing for some individuals (for example, bias due to the inclusion of prevalent users rather than new users of an intervention), and (2) bias that arises when later follow-up is missing for individuals who were initially included and followed (for example, bias due to differential loss to follow-up that is affected by prognostic factors).We consider the first type of selection bias under "Bias in selection of participants into the study" (section 4.2), and aspects relating to loss to follow up are covered under "Bias due to missing data" (section 4.5). Examples of these types of bias are given within the relevant sections.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 1. Bias domains included in the ROBINS-I tool

Domain	Related terms	Explanation	
Pre-intervention			Pre-intervention or at-intervention domains for which risk of bias assessment is mainly distinct from assessments of randomized trials
Bias due to confounding	Selection bias *as it is sometimes used in relation to clinical trials* (and currently in widespread use within Cochrane); Allocation bias; Case-mix bias; Channelling bias.	Baseline confounding occurs when one or more prognostic variables (factors that predict the outcome of interest) also predicts the intervention received at baseline. ROBINS-I can also address time-varying confounding, which occurs when individuals switch between the interventions being compared and when post-baseline prognostic factors affect the intervention received after baseline.	
Bias in selection of participants into the study	Selection bias *as it is usually used in relation to observational studies and sometimes used in relation to clinical trials*; Inception bias; Lead-time bias; Immortal time bias. Note that this bias specifically excludes lack of external validity, which is viewed as a failure to generalize or transport an unbiased (internally valid) effect estimate to populations other than the one from which the study population arose.	When exclusion of some eligible participants, or the initial follow up time of some participants, or some outcome events, is related to both intervention and outcome, there will be an association between interventions and outcome even if the effects of the interventions are identical. This form of selection bias is distinct from confounding. A specific example is bias due to the inclusion of prevalent users, rather than new users, of an intervention.	
At intervention			
Bias in classification of interventions	Misclassification bias; Information bias; Recall bias; Measurement bias; Observer bias.	Bias introduced by either differential or non-differential misclassification of intervention status. Non-differential misclassification is unrelated to the outcome and will usually bias the estimated effect of intervention towards the null. Differential misclassification occurs when misclassification of intervention status is related to the outcome or the risk of the outcome, and is likely to lead to bias.	

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Post-intervention			Post-intervention domains for which there is substantial overlap with assessments of randomized trials
Bias due to deviations from intended interventions	Performance bias; Time-varying confounding	Bias that arises when there are systematic differences between experimental intervention and comparator groups in the care provided, which represent a deviation from the intended intervention(s). Assessment of bias in this domain will depend on the type of effect of interest (either the effect of assignment to intervention or the effect of starting and adhering to intervention).	
Bias due to missing data	Attrition bias; Selection bias *as it is sometimes used in relation to observational studies*	Bias that arises when later follow-up is missing for individuals initially included and followed (e.g. differential loss to follow-up that is affected by prognostic factors); bias due to exclusion of individuals with missing information about intervention status or other variables such as confounders.	
Bias in measurement of outcomes	Detection bias; Recall bias; Information bias; Misclassification bias; Observer bias; Measurement bias	Bias introduced by either differential or non-differential errors in measurement of outcome data. Such bias can arise when outcome assessors are aware of intervention status, if different methods are used to assess outcomes in different intervention groups, or if measurement errors are related to intervention status or effects.	
Bias in selection of the reported result	Outcome reporting bias; Analysis reporting bias	Selective reporting of results in a way that depends on the findings.	

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

2.4 Study designs

This document relates most closely to NRSIs with **cohort-like designs**, such as cohort studies, quasi-randomized trials and other concurrently controlled studies. Much of the material is also relevant to designs such as case-control studies, cross-sectional studies, interrupted time series and controlled before-after studies, although we are currently considering whether modifications to the signalling questions are required for these other types of studies.

2.5 Risk of bias assessments should relate to a specified intervention effect

This section relates to the effect of intervention that a study aims to quantify. The effect of interest in the target trial will be either

- the effect of **assignment** to the intervention at baseline (start of follow-up), regardless of the extent to which the intervention was received during follow-up (sometimes referred to as the "intention-to-treat" effect in the context of randomized trials); or
- the effect of **starting and adhering** to the intervention as specified in the trial protocol (sometimes referred to as the "per-protocol" effect in the context of randomized trials).

For example, to inform a health policy question about whether to recommend an intervention in a particular health system we would probably estimate the effect of *assignment to intervention*, whereas to inform a care decision by an individual patient we would wish to estimate the effect of *starting and adhering to* the treatment according to a specified protocol, compared with a specified comparator. Review authors need to define the intervention effect of interest to them in each NRSI, and apply the risk of bias tool appropriately to this effect. Issues relating to the choice of intervention effect are discussed in more detail in Section 3.2.2 below.

Note that in the context of ROBINS-I, specification of the intervention effect does not relate to choice of a relative or absolute measures, nor to specific PICO (patient, intervention, comparator, outcome) elements of the review question.

2.6 Structure of this document

Sections 3 and 4 of this document provide detailed guidance on use of ROBINS-I. This includes considerations during the process of writing the review protocol (section 3.1), issues in specifying the effect of interest (section 3.2.2), the use of signalling questions in assessments of risk of bias (section 3.3), the requirement for domain-level bias judgements (section 3.4), how these are used to reach an overall judgement on risk of bias (section 3.5) and the use of outcome-level assessments (section 3.6). Detailed guidance on bias assessments for each domain is provided in Section 4.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

3 Guidance for using the tool: general considerations

3.1 At protocol stage

3.1.1 Specifying the research question

The research question follows directly from the objective(s) of the review. It addresses the population, experimental intervention, comparator and outcomes of interest. The comparator could be no intervention, usual care, or an alternative intervention.

A review of NRSI should begin with consideration of what problems might arise, in the context of the research question, in making a causal assessment of the effect of the intervention(s) of interest on the basis of NRSI. It is helpful to think about what is to be studied, why it is to be studied, what types of study are likely to be found, and what problems are likely to be encountered in those studies. Identification of the problems that might arise will be based in part on subject matter experts' knowledge of the literature: the team should also address whether conflicts of interest might affect experts' judgements.

Features of the research question may highlight difficulties in defining the intervention being evaluated in a NRSI, or complexities that may arise with respect to the tools used to measure an outcome domain or the timing of measurements. Ideally, the protocol will specify how the review authors plan to accommodate such complexities in their conduct of the review as well as in preparing for the risk of bias assessment.

3.1.2 Listing the confounding domains relevant to all or most studies eligible for the review

Relevant confounding domains are the prognostic factors that predict whether an individual receives one or the other intervention of interest. They are likely to be identified both through the knowledge of subject matter experts who are members of the review group, and through initial (scoping) reviews of the literature. Discussions with health professionals who make intervention decisions for the target patient or population groups may also be helpful. These issues are discussed further in section 4.1.

3.1.3 Listing the possible co-interventions that could differ between intervention groups and have an impact on study outcomes

Relevant co-interventions are the interventions or exposures that individuals might receive after or with initiation of the intervention of interest, which are related to the intervention received and which are prognostic for the outcome of interest. These are also likely to be identified through the expert knowledge of members of the review group, via initial (scoping) reviews of the literature, and after discussions with health professionals. These issues are discussed further in section 4.4.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 1: The ROBINS-I tool (Stage 1): At protocol stage

Specify the review question

Participants	
Experimental intervention	
Comparator	
Outcomes	

List the confounding domains relevant to all or most studies

List co-interventions that could be different between intervention groups and that could impact on outcomes

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

3.2 Preliminary considerations for each study

3.2.1 Specifying a target trial specific to the study

Evaluations of risk of bias are facilitated by considering the NRSI as an attempt to emulate a pragmatic randomized trial, which we refer to as the **target trial**. The first part of a ROBINS-I assessment for a particular study is to specify a target trial (Box 2). The target trial is the hypothetical randomized trial whose results should be the same as those from the NRSI under consideration, in the absence of bias. Its key characteristics are the types of participant (including exclusion/inclusion criteria) and a description of the experimental and comparator interventions. These issues are considered in more detail by Hernán (2001). The differences between the target trial for the individual NRSI and the generic research question of the review relate to issues of heterogeneity and/or generalizability rather than risk of bias.

Because it is hypothetical, ethics and feasibility need not be considered when specifying the target trial. For example there would be no objection to a target trial that compared individuals who did and did not start smoking, even though such a trial would be neither ethical nor feasible in practice.

Selection of a patient group that is eligible for a target trial may require detailed consideration, and lead to exclusion of many patients. For example, Magid et al, (2010) studied the comparative effectiveness of ACE inhibitors compared to beta-blockers as second-line treatments for hypertension. From an initial cohort of 1.6m patients, they restricted the analysis population to (1) persons with incident hypertension, (2) who were initially treated with a thiazide agent, and (3) who had one of the two drugs of interest added as a second agent for uncontrolled hypertension, and (4) who did not have a contraindication to either drug. Their "comparative effectiveness" cohort included 15,540 individuals: less than 1% of the original cohort.

3.2.2 Specifying the effect of interest

In the target trial, the effect of interest for any specific research question will be either the effect of **assignment** to the interventions at baseline, regardless of the extent to which the interventions were received during the follow-up, or the effect of **starting and adhering to** the interventions as specified in the protocol (Box 2). The choice between these effects is a decision of the review authors, and is not determined by the choice of analyses made by authors of the NRSI. However, the analyses of an NRSI may correspond more closely to one of the effects of interest, and therefore be biased with respect to the other one.

In the context of randomized trials, the effect of assignment to intervention can be estimated via an **intention-to-treat (ITT) analysis**, in which participants are analysed in the intervention groups to which they were randomized. In the presence of non-adherence to randomized intervention, an ITT analysis of a placebo-controlled trial underestimates the intervention effect that would have been seen if all participants had adhered to the randomized allocation. Although ITT effects may be regarded as conservative with regard to desired effects of interventions estimated in placebo-controlled trials, they may not be conservative in trials comparing two or more active interventions, and are problematic for non-inferiority or equivalence studies, or for estimating harms.

Patients and other stakeholders are often interested in the effect of starting and adhering to the intervention as described in the trial protocol (sometimes referred to as the **per protocol effect**). This is also the effect that is likely to be of interest when considering adverse (or unintended) effects of interventions. It is possible to use data from randomized trials to estimate the effect of starting and adhering to intervention. However, approaches used to do so in papers reporting on randomized trials are often problematic. In particular, unadjusted analyses based on the treatment actually received, or naïve "per protocol" analyses restricted to individuals in each intervention group who (or the follow up during which they) adhered to the trial protocol can be biased, if prognostic factors influenced treatment received. Advanced statistical methods permit appropriate adjustment for such bias, although applications of such methods are relatively rare. Alternative methods that use randomization status as an instrumental variable bypass the need to adjust for such prognostic factors, but they are not always applicable.

Analogues of these effects can be defined for NRSI. For example, the intention-to-treat effect can be approximated by the effect of *starting* experimental intervention versus *starting* comparator intervention, which corresponds to the intention-to-treat effect in a trial in which participants assigned to an intervention always start that intervention). This differs slightly from the ITT effect in randomized trials, because some individuals randomly assigned to a particular intervention may never initiate it. An analogue of the effect of starting and adhering to

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

the intervention as described in the trial protocol is *starting and adhering to* experimental intervention versus *starting and adhering to* comparator intervention unless medical reasons (e.g. toxicity) indicate discontinuation.

For example, in a study of cancer screening the effect of interest might relate either to receipt (or not) of an invitation to screening (the effect estimated in an ITT analysis of a randomized trial of screening), or to uptake (or not) of an invitation to screening.

For both randomized trials and NRSI, unbiased estimation of the effect of starting and adhering to intervention requires appropriate adjustment for prognostic factors that predict deviations from the intended interventions ("time-varying confounders", see detailed discussion in sections 4.1.9 and 4.4). Review authors should seek specialist advice when assessing intervention effects estimated using methods that adjust for time-varying confounding.

In both randomized trials and NRSI, risk of bias assessments should be in relation to a specified effect of interest. **When the effect of interest is that of assignment to the intervention at baseline (randomized trials) or starting intervention at baseline (NRSI), risk of bias assessments for both types of study need not be concerned with post-baseline deviations from intended interventions** that reflect the natural course of events (for example, a departure from randomized intervention that was clinically necessary because of a sudden worsening of the patient's condition) rather than potentially biased actions of researchers. When the effect of interest is starting and adhering to the intended intervention, risk of bias assessments of both randomized trials and NRSI may have to consider adherence and differences in additional interventions ("co-interventions") between intervention groups. More detailed discussions of these issues are provided in sections 4.1.8, 4.1.9 and 4.4.

3.2.3 Preliminary considerations of confounders and co-interventions

We recommend that the study be examined in detail in two key areas before completing the tool proper (Box 3). These two areas are confounders and co-interventions. The process should determine whether the critical confounders and co-interventions as specified in the protocol were measured or administered in the study at hand, and whether additional confounders and co-interventions were identified in the study. Further guidance and a structure for the assessment is provided in sections 4.1 and 4.4.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 2: The ROBINS-I tool (Stage 2, part 1): For each study: setting up the assessment

Specify a target randomized trial specific to the study

Design	Individually randomized / Cluster randomized / Matched (e.g. cross-over)
Participants	
Experimental intervention	
Comparator	

Is your aim for this study…?

☐ to assess the effect of *assignment to* intervention

☐ to assess the effect of *starting and adhering to* intervention

Specify the outcome

Specify which outcome is being assessed for risk of bias (typically from among those earmarked for the Summary of Findings table). Specify whether this is a proposed benefit or harm of intervention.

Specify the numerical result being assessed

In case of multiple alternative analyses being presented, specify the numeric result (e.g. RR = 1.52 (95% CI 0.83 to 2.77) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 3: The ROBINS-I tool (Stage 2, part 2): For each study: evaluation of confounding domains and co-interventions

Preliminary consideration of confounders

Complete a row for each important confounding domain (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as potentially important.

"Important" confounding domains are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the intervention. "Validity" refers to whether the confounding variable or variables fully measure the domain, while "reliability" refers to the precision of the measurement (more measurement error means less reliability).

(i) Confounding domains listed in the review protocol				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?*	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favour the experimental intervention or the comparator?
			Yes / No / No information	Favour experimental / Favour comparator / No information

(ii) Additional confounding domains relevant to the setting of this particular study, or which the study authors identified as important				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?*	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favour the experimental intervention or the comparator?
			Yes / No / No information	Favour experimental / Favour comparator / No information

* In the context of a particular study, variables can be demonstrated not to be confounders and so not included in the analysis: (a) if they are not predictive of the outcome; (b) if they are not predictive of intervention; or (c) because adjustment makes no or minimal difference to the estimated effect of the primary parameter. Note that "no statistically significant association" is not the same as "not predictive"

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Preliminary consideration of co-interventions

Complete a row for each important co-intervention (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as important.
"Important" co-interventions are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the intervention.

(i) Co-interventions listed in the review protocol		
Co-intervention	Is there evidence that controlling for this co-intervention was unnecessary (e.g. because it was not administered)?	Is presence of this co-intervention likely to favour outcomes in the experimental intervention or the comparator
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information

(ii) Additional co-interventions relevant to the setting of this particular study, or which the study authors identified as important		
Co-intervention	Is there evidence that controlling for this co-intervention was unnecessary (e.g. because it was not administered)?	Is presence of this co-intervention likely to favour outcomes in the experimental intervention or the comparator
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

3.3 Signalling questions

A key feature of the tool is the inclusion of signalling questions within each domain of bias. These are reasonably factual in nature and aim to facilitate judgements about the risk of bias.

The **response options for the signalling questions** are:

(1) Yes;
(2) Probably yes;
(3) Probably no;
(4) No; and
(5) No information.

One exception to this system is the opening signalling question (1.1, in the assessment of bias due to confounding) does not have a "No information" option.

Some signalling questions are only answered in certain circumstances, for example if the response to a previous question is "Yes" or "Probably yes" (or "No" or "Probably no"). When questions are not to be answered, a response option of "Not applicable" may be selected. Responses underlined in green in the tool are potential markers for low risk of bias, and responses in red are potential markers for a risk of bias. Where questions relate only to sign posts to other questions, no formatting is used.

Responses of "Yes" and "Probably yes" (also of "No" and "Probably no") have similar implications, but allow for a distinction between something that is known and something that is likely to be the case. The former would imply that firm evidence is available in relation to the signalling question; the latter would imply that a judgement has been made. If measures of agreement are applied to answers to the signalling questions, we recommend grouping these pairs of responses.

3.3.1 Free-text boxes alongside signalling questions

There is space for free text alongside each signalling question. This should be used to provide support for each answer. Brief direct quotations from the text of the study report should be used when possible to support responses.

3.4 Domain-level judgements about risk of bias

ROBINS-I is conceived hierarchically: responses to signalling questions (relatively factual, "what happened" or "what researchers did") provide the basis for domain-level judgements about RoB, which then provide the basis for an overall RoB judgement for a particular outcome. Use of the word "judgement" to describe the second and third stages is very important, since the review author needs to consider both the severity of the bias in a particular domain and the relative consequences of bias in different domains. The key to applying the tool is to make domain-level judgements about risk of bias that mean the same across domains with respect to concern about the impact of bias on the trustworthiness of the result. If domain-level judgements are made consistently, then judging the overall RoB for a particular outcome is relatively straightforward (see 3.5).

Criteria for reaching risk of bias judgements for the seven domains are provided. If none of the answers to the signalling questions for a domain suggest a potential problem then risk of bias for the domain can be judged to be low. Otherwise, potential for bias exists. Review authors must then make a judgement on the extent to which the results of the study are at risk of bias. "Risk of bias" is to be interpreted as "**risk of material bias**". That is, concerns should be expressed only about issues that are likely to affect the ability to draw valid conclusions from the study: a serious risk of a very small degree of bias should not be considered "Serious risk" of bias

The "no information" category should be used only when insufficient data are reported to permit a judgment.

The **response options for each domain-level RoB judgement** are:

(1) Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain);
(2) Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial);

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

(3) Serious risk of bias (the study has some important problems in this domain);
(4) Critical risk of bias (the study is too problematic in this domain to provide any useful evidence on the effects of intervention); *and*
(5) No information on which to base a judgement about risk of bias for this domain.

The "low risk of bias" category exists to emphasize the distinction between randomized trials and non-randomized studies. These distinctions apply in particular to the "pre-intervention" and "at-intervention" domains (see Table 1). In particular, we anticipate that only rarely design features of a non-randomized study will lead to a classification of low risk of bias due to confounding. Randomization does not protect against post-intervention biases, and we therefore expect more overlap between assessments of randomized trials and assessments of NRSI for the post-intervention domains. However other features of randomized trials, such as blinding of participants, health professionals or outcome assessors, may protect against post-intervention biases.

3.4.1 Free-text boxes alongside risk of bias judgements

There is space for free text alongside each RoB judgement to explain the reasoning that underpins the judgement. It is essential that the reasons are provided for any judgements of "Serious" or "Critical" risk of bias.

3.4.2 Direction of bias

It would be highly desirable to know the magnitude and direction of any potential biases identified, but this is considerably more challenging than judging the risk of bias. The tool includes an optional component to judge the direction of the bias for each domain and overall. For some domains, the bias is most easily thought of as being towards or away from the null. For example, suspicion of selective non-reporting of statistically non-significant results would suggest bias against the null. However, for other domains (in particular confounding, selection bias and forms of measurement bias such as differential misclassification), the bias needs to be thought of not in relation to the null, but as an increase or decrease in the effect estimate (i.e. to favour either the experimental intervention or comparator). For example, confounding bias that decreases the effect estimate would be towards the null if the true risk ratio were greater than 1, and away from the null if the risk ratio were less than 1. **If review authors do not have a clear rationale for judging the likely direction of the bias, they should not attempt to guess it.**

3.5 Reaching an overall judgement about risk of bias

The **response options for an overall RoB judgement** are:

(1) Low risk of bias (the study is comparable to a well-performed randomized trial);
(2) Moderate risk of bias (the study provides sound evidence for a non-randomized study but cannot be considered comparable to a well-performed randomized trial);
(3) Serious risk of bias (the study has some important problems);
(4) Critical risk of bias (the study is too problematic to provide any useful evidence and should not be included in any synthesis); *and*
(5) No information on which to base a judgement about risk of bias

Table 2 shows the basic approach to be used to map RoB judgements within domains to a single RoB judgement across domains for the outcome.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 2. Reaching an overall RoB judgement for a specific outcome.

RESPONSE OPTION	CRITERIA
Low risk of bias (the study is comparable to a well-performed randomized trial);	The study is judged to be at **low risk of bias for all domains**.
Moderate risk of bias (the study appears to provide sound evidence for a non-randomized study but cannot be considered comparable to a well-performed randomized trial);	The study is judged to be at **low or moderate risk of bias for all domains**.
Serious risk of bias (the study has some important problems);	The study is judged to be at **serious risk of bias** in at least one domain, but not at critical risk of bias in any domain.
Critical risk of bias (the study is too problematic to provide any useful evidence and should not be included in any synthesis);	The study is judged to be at **critical risk of bias in at least one domain.**
No information on which to base a judgement about risk of bias.	There is no clear indication that the study is at serious or critical risk of bias *and* there is a lack of information in one or more key domains of bias (*a judgement is required for this*).

Declaring a study to be at a particular level of risk of bias for an individual domain will mean that the study as a whole has a risk of bias at least this severe (for the outcome being assessed). Therefore, a judgement of "Serious risk of bias" within any domain should have similar implications for the study as a whole, irrespective of which domain is being assessed.

Because it will be rare that an NRSI is judged as at low risk of bias due to confounding, we anticipate that most NRSI will be judged as at least at moderate overall risk of bias.

The mapping of domain-level judgements to overall judgements described in Table 2 is a programmable algorithm. However, in practice some "Serious" risks of bias (or "Moderate" risks of bias) might be considered to be additive, so that "Serious" risks of bias in multiple domains can lead to an overall judgement of "Critical" risk of bias (and, similarly, "Moderate" risks of bias in multiple domains can lead to an overall judgement of "Serious" risk of bias).

3.6 Assessing risk of bias for multiple outcomes in a review

ROBINS-I addresses the risk of bias in a specific result from a NRSI. The risk of bias in the effect of an intervention may be very different for different analyses of the same outcome (e.g. when different analyses adjust for different confounders), as well as for different outcomes. NRSI included in systematic reviews will frequently (if not usually) contribute results for multiple outcomes, so several risk of bias assessments may be needed for each study. Table 3 shows examples of possible assessments for a hypothetical NRSI that addresses three outcomes, O1 (e.g. mortality), O2 (e.g. viral load) and O3 (e.g. quality of life).

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 3. Reaching an overall RoB judgement for a specific outcome.

Domain	Assessments by outcome	Comment
Bias due to confounding	O1: Serious risk	e.g. only counts available (no adjustment for confounders)
	O2: Moderate risk	e.g. appropriately adjusted
	O3: Serious risk	e.g. insufficient adjustment
Bias in selection of participants into the study	Grouped (O1, O2, O3): Low risk	e.g. same issues thought to apply to all
Bias in classification of interventions	Grouped (O1, O2, O3): Low risk	e.g. same issues thought to apply to all
Bias due to deviations from intended interventions	Grouped (O1, O2, O3): Moderate risk	e.g. same issues thought to apply to all
Bias due to missing data	O1: Low risk	e.g. everyone followed up through records
	Grouped (O2, O3): No information	e.g. due to attrition; same participants
Bias in measurement of outcomes	Grouped (O1, O2): Low risk	e.g. both objective measures
	O3: Serious risk	e.g. prone to biases due to lack of blind outcome assessment
Bias in selection of the reported result	O1: Moderate risk	e.g. unlikely to be manipulated
	O2: Moderate risk	e.g. unlikely to be manipulated
	O3: Serious risk	e.g. cut-point used without justification

This would give us the RoB profiles (which might accompany meta-analyses and/or GRADE assessments) shown in Table 4.

Table 4. Illustration of different RoB judgements for different outcomes

Domain	O1	O2	O3
Bias due to confounding	Serious risk	Moderate risk	Serious risk
Bias in selection of participants into the study	Low risk	Low risk	Low risk
Bias in classification of interventions	Low risk	Low risk	Low risk
Bias due to deviations from intended interventions	Moderate risk	Moderate risk	Moderate risk
Bias due to missing data	Low risk	No info	No info
Bias in measurement of outcomes	Low risk	Low risk	Serious risk
Bias in selection of the reported result	Moderate risk	Moderate risk	Serious risk
Overall*	*Serious risk*	*Moderate risk*	*Serious risk*

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

4 Guidance for using the tool: detailed guidance for each bias domain

4.1 Detailed guidance: Bias due to confounding

4.1.1 Background

A confounding domain is a pre-intervention prognostic factor that predicts whether an individual receives one or the other intervention of interest. Some common examples are severity of pre-existing disease, presence of comorbidities, health care utilization, adiposity, and socioeconomic status. Confounding domains can be characterised by measuring one or more of a range of specific variables. The relevant confounding domains vary across study settings. For example, socioeconomic status might not introduce confounding in studies conducted in countries in which access to the interventions of interest is universal and therefore socioeconomic status does not influence intervention received.

The tool addresses two types of confounding: baseline confounding and time-varying confounding.

4.1.2 Baseline confounding

Baseline confounding occurs when one or more pre-intervention prognostic factors predict the intervention received at start of follow up. A **pre-intervention variable** is one that is measured before the start of interventions of interest. For example, a non-randomized study comparing two antiretroviral drug regimens should control for CD4 cell count measured before the start of antiretroviral therapy, because this is strongly prognostic for AIDS and death and is likely to influence choice of regimen. Baseline confounding is likely to be an issue in most or all NRSI.

4.1.3 Time-varying confounding

Time-varying confounding occurs when the intervention received can change over time (for example, if individuals switch between the interventions being compared), and when post-baseline prognostic factors affect the intervention received after baseline. A **post-baseline variable** is one that is measured after baseline: for example CD4 cell count measured 6 months after initiation of therapy. **Time-varying confounding needs to be considered in studies that partition follow-up time for individual participants into time spent in different intervention groups**.

For example, suppose a study of patients treated for HIV partitions follow-up time into periods during which patients were receiving different antiretroviral regimens and compares outcomes during these periods in the analysis. CD4 cell count (as a post-baseline prognostic variable) might influence switches between the regimens of interest. When post-baseline prognostic variables are affected by the interventions themselves (for example, antiretroviral regimen may influence post-baseline CD4 count), conventional adjustment for them in statistical analyses is not appropriate as a means of controlling for confounding. For example, CD4 count measured after start of antiretroviral therapy (a post-baseline prognostic variable) might influence switches between the regimens of interest (Hernán et al, 2002). When post-baseline prognostic variables are affected by the interventions themselves (for example, antiretroviral regimen may influence post-baseline CD4 count), conventional adjustment for them in statistical analyses is not appropriate as a means of controlling for confounding (Hernán et al, 2002; Hernán et al, 2004). Note that when individuals switch between the interventions being compared the effect of interest is that of starting and adhering to intervention, not the effect of assignment to intervention.

As a further example, a large open comparative NRSI compared cardiovascular events in patients while taking a new medication for diabetes with those in control patients while receiving older therapies. Research evidence published during the study's follow up period suggested that the new diabetes medication increased the risk of vascular events. Patients whose blood pressure or lipid levels deteriorated after study entry were switched away from the new drug by physicians concerned about the cardiovascular risk. Because blood pressure and lipid levels were prognostic for cardiovascular events **and** predicted the intervention received after baseline, the study was at risk of bias due to time-varying confounding. These issues are discussed in sections 4.1.8 and 4.1.9.

4.1.4 Identifying confounding domains

Important confounding domains should be pre-specified in the protocol of a review of NRSI. The identification of potential confounding domains requires subject-matter knowledge. For example, in an observational study

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

comparing minimally invasive and open surgical strategies, lack of adjustment for pre-intervention fitness for surgery (comorbidity), measured by American Society of Anesthesiologists (ASA) class or Charlson index, would result in confounding if this factor predicted choice of surgical strategy. Experts on surgery are best-placed to identify prognostic factors that are likely to be related to choice of surgical strategy. The procedures described below are therefore designed to be used by raters who have good knowledge of the subject matter under study. **We recommend that subject-matter experts be included in the team writing the review protocol, and encourage the listing of confounding domains in the review protocol, based on initial discussions among the review authors and existing knowledge of the literature.**

It is likely that new ideas relating to confounding and other potential sources of bias will be identified after the drafting of the review protocol, and even after piloting data collection from studies selected for inclusion in the systematic review. For example, such issues may be identified because they are mentioned in the introduction and/or discussion of one or more papers. This could be addressed by explicitly recording whether potential confounders or other sources of bias are mentioned in the paper, as a field for data collection.

For rare or unusual adverse effects the underlying risk factors may not be known, and it may prove difficult to identify sources of confounding beforehand. For instance, nephrogenic systemic fibrosis is a rare, recently discovered adverse event where the aetiological factors and natural history have yet to be elucidated. In this specific situation, review authors may not be able to specify relevant sources of confounding beforehand or to judge if studies assessing this adverse event have adequately addressed confounding. On the other hand, review authors could judge confounding to be implausible if they believed that those assigning interventions were not aware of the possibility of an adverse effect and so unlikely to make treatment decisions based on risk factors for that adverse effect. Note that if the adverse effect is a result of, or correlated with, a known adverse event (for example, poor kidney function in the nephrogenic systemic fibrosis example above) of treatment, then confounding may still be present.

4.1.5 Residual and unmeasured confounding

Because confounding domains may not be directly measured, investigators measure specific variables (often referred to as confounders) in an attempt to fully or partly adjust for these confounding domains. For example, baseline CD4 cell count and recent weight loss may be used to adjust for disease severity; hospitalizations and number of medical encounters in the 6 months preceding baseline may be used to adjust for healthcare utilization; geographic measures to adjust for physician prescribing practices; body mass index and waist-to-hip ratio to adjust for adiposity; and income and education to adjust for socioeconomic status.

We can identify two broad reasons that confounding is not fully controlled. **Residual confounding** occurs when a confounding domain is measured with error, or when the relation between the confounding domain and the outcome or exposure (depending on the analytic approach being used) is imperfectly modelled. For example, in a NRSI comparing two antihypertensive drugs, we would expect residual confounding if pre-intervention blood pressure was measured 3 months before the start of intervention, but the blood pressures used by clinicians to decide between the drugs at the point of intervention were not available in our dataset. **Unmeasured confounding** occurs when a confounding domain has not been measured, or when it is not controlled in the analysis. This would be the case if no pre-intervention blood pressure measurements were available, or if the analysis failed to control for pre-intervention blood pressure despite it being measured.

Note that when intervention decisions are made by health professionals, measurement error in the information available to them does not necessarily introduce residual confounding. For example, pre-intervention blood pressure will not perfectly reflect underlying blood pressure. However, if intervention decisions were made based on two pre-intervention measurements, and these measurements were available in our dataset, it would be possible to adjust fully for the confounding.

For some review questions the confounding may be intractable, because it is not possible to measure all the confounding domains that influence treatment decisions. For example, consider a study of the effect of treating type 2 diabetes with insulin when oral antidiabetic drugs fail. The patients are usually older, and doctors may, without recording their decisions, prescribe insulin treatment mostly to those without cognitive impairment and with sufficient manual dexterity. This creates potentially strong confounding that may not be measurable.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

4.1.6 Control of confounding

When all confounders are measured without error, confounding may be controlled either by design (for example by restricting eligibility to individuals who all have the same value of the baseline confounders) or through statistical analyses that adjust ("control") for the confounding factor(s). If, in the context of a particular study, a confounding factor is unrelated to intervention or unrelated to outcome, then there is no need to control for it in the analysis. It is however important to note that in this context "unrelated" means "not associated" (for example, risk ratio close to 1) and does not mean "no statistically significant association".

Appropriate control of confounding requires that the variables used are valid and reliable measures of the confounding domains. In this context, "validity" refers to whether the variable or variables fully measures the domain, while "reliability" refers to the precision of the measurement (more measurement error means less reliability) (Streiner and Norman, 2003). For some topics, a list of valid and reliable measures of confounding domains will be available in advance and should be specified in the review protocol. For other topics, such a list may not be available. Study authors may cite references to support the use of a particular measure: reviewers can then base their judgment of the validity and reliability of the measure based on these citations (Cook and Beckman, 2006). Some authors may control for confounding variables with no indication of their validity or reliability. In such instances, review authors should pay attention to the subjectivity of the measure. Subjective measures based on self-report may tend to have lower validity and reliability relative to objective measures such as clinical reports and lab findings (Cook et al, 1990).

It is important to consider whether inappropriate adjustments were made. In particular, **adjusting for post-intervention variables is usually not appropriate**. Adjusting for **mediating variables** (those on the causal pathway from intervention to outcome) restricts attention to the effect of intervention that does not go via the mediator (the "direct effect") and may introduce confounding, even for randomized trials. Adjusting for **common effects** of intervention and outcome causes bias. For example, in a study comparing different antiretroviral drug combinations it will usually be essential to adjust for pre-intervention CD4 cell count, but it would be inappropriate to adjust for CD4 cell count 6 months after initiation of therapy.

4.1.7 Negative controls

Use of a "**negative control**" – exploration of an alternative analysis in which no association should be observed – can sometimes address the likelihood of unmeasured confounding. Lipsitch *et al* (2010) discussed this issue, and distinguished two types of negative controls: exposure controls and outcome controls. One example discussed by these authors relates to observational studies in elderly persons that have suggested that vaccination against influenza is associated with large reductions in risk of pneumonia/influenza hospitalization and in all-cause mortality. To test this hypothesis, Jackson *et al* (2006) reproduced earlier estimates of the protective effect of influenza vaccination, then repeated the analysis for two sets of negative control outcomes. First, they compared the risk of pneumonia/influenza hospitalization and all-cause mortality in vaccinated and unvaccinated persons before, during, and after influenza season ("exposure control"). They reasoned that if the effect measured in previous studies was causal, it should be most prominent during influenza season. Despite efforts to control for confounding, they observed that the protective effect was actually greatest before, intermediate during, and least after influenza season. They concluded that this is evidence that confounding, rather than protection against influenza, accounts for a substantial part of the observed "protection." Second, they postulated that the protective effects of influenza vaccination, if real, should be limited to outcomes plausibly linked to influenza. They repeated their analysis, but substituted hospitalization for injury or trauma as the end point ("outcome control"). They found that influenza vaccination was also "protective" against injury or trauma hospitalization. This, too, was interpreted as evidence that some of the protection observed for pneumonia/influenza hospitalization or mortality was due to inadequately controlled confounding. A second example of "outcome control" is that studies of smoking and suicide also found an association between smoking and homicide (Davey Smith et al, 1992).

4.1.8 Switches between interventions

In some (perhaps many) NRSI, particularly those based on routinely collected data, the intervention received by participants may change, during follow up, from the intervention that they received at baseline to another of the interventions being compared in the review. This may result in "**switches between interventions of interest**", a phenomenon that we consider here under the confounding domain (see "time-varying confounding" below). If one of the intervention groups being compared is no intervention, then such switches include discontinuation of

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

active intervention, or starting active treatment for individuals assigned to control. On the other hand, change from the baseline intervention may result in switching to an intervention that is not of interest to the review question. We consider switches of this kind under "Deviations from intended intervention".

For studies in which participants switch between interventions, risk of bias assessments will depend on the effect of interest. There are two broad approaches:

1. The effect of interest is the effect of assignment to (or starting) experimental intervention versus assignment to (or starting) comparator intervention) and participants are analysed in groups defined by the initial intervention received. In this circumstance, switches between interventions during follow up do not cause bias. For example, consider a study in which men with screen-detected localized prostate cancer are assigned to either immediate surgery or active monitoring of their cancer. Some men subsequently receive surgery, but they would be analysed according to the initial intervention. As another example, a study examining the effect of women's choice of oral contraceptive on their subsequent risk of breast cancer would include all follow-up time, regardless of whether women stopped using contraception because they wished to conceive.
2. The effect of interest is the effect of starting and adhering to intervention, and follow-up time is split into time during which different interventions were received. For example, in a 12-month study comparing two selective serotonin-reuptake inhibitors (SSRIs) A and B with no intervention, a patient might spend 6 months on A, two months on no intervention and four months on B, and these follow up periods are assigned to the different interventions in the analysis. Such studies and analyses depend on an assumption that the risk of the outcome of interest changes soon after change of intervention: for example the study authors may believe that any change in the risk of venous thrombosis that is associated with the use of a particular oral contraceptive stops soon after the use of that intervention. By contrast, study authors may believe that changes in the risk of breast cancer are sustained for a considerable period after cessation of the oral contraceptive: estimation of "per protocol" effects would then be very difficult because it would be necessary to make strong assumptions about the contributions of previous and current interventions to breast cancer risk during a particular period of follow-up.

4.1.9 Time-varying confounding

When follow-up time is split according to the intervention received we need to assess the risk of bias due to time-varying confounding. If the values of factors that are prognostic for the outcome of interest and predict a switch of intervention also change with time, then adjusting only for baseline confounding is insufficient. For example, in a study comparing the effect of non-steroidal anti-inflammatory drugs (NSAIDS) on mortality, in which participants switched during follow-up between the NSAIDS being compared, time-varying confounding would occur if episodes of gastrointestinal bleeding during follow up were prognostic for mortality and also predicted switches between NSAIDS.

4.1.10 Technical note: adjusting for time-varying confounding

Time-varying confounding occurs when time-varying factors that predict the outcome also affect changes of intervention. If, in addition, past changes to intervention affect subsequent values of the same factors, standard statistical methods (such as Cox regression models including the time-varying factor) are not able to adjust appropriately for the confounding, even if the factor concerned is perfectly measured and its effect perfectly modelled. Studies of antiretroviral therapy (ART) for HIV infection provide an example: CD_4 cell count is a prognostic factor for AIDS that might predict adherence to ART, and adherence to ART affects subsequent CD_4 counts. In these circumstances, estimation of the effect of continuous intervention is in principle possible, but requires use of methods that can deal with time-varying confounding. A commonly used method is inverse probability weighting of marginal structural models, but their implementation is technically challenging. Specialist advice should be sought for risk of bias assessments of studies employing these methods.

4.1.11 Risk of bias assessment for bias due to confounding

The signalling questions and risk of bias assessments are given in Box 4 and Table 5. If there is potential for confounding, risk of bias judgements should be based on answers to questions 1.4 to 1.6 for studies in which participants remained in their initial intervention group during follow up or for which time-varying confounding is not expected, and on answers to questions 1.7 and 1.8 for studies in which participants switched between interventions of interest and time-varying confounding is expected.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 4: The ROBINS-I tool (Stage 2, part 3): Risk of bias due to confounding

Signalling questions	Elaboration	Response options
1.1 Is there potential for confounding of the effect of intervention in this study? **If N/PN to 1.1:** the study can be considered to be at low risk of bias due to confounding and no further signalling questions need be considered **If Y/PY to 1.1**: determine whether there is a need to assess time-varying confounding:	In rare situations, such as when studying harms that are very unlikely to be related to factors that influence treatment decisions, no confounding is expected and the study can be considered to be at low risk of bias due to confounding, equivalent to a fully randomized trial. There is no NI (No information) option for this signalling question.	Y / PY / PN / N
1.2. Was the analysis based on splitting participants' follow up time according to intervention received? **If N/PN**, answer questions relating to baseline confounding (1.4 to 1.6) **If Y/PY**, go to question 1.3.	If participants could switch between intervention groups then associations between intervention and outcome may be biased by time-varying confounding. This occurs when prognostic factors influence switches between intended interventions.	NA / Y / PY / PN / N / NI
1.3. Were intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome? **If N/PN**, answer questions relating to baseline confounding (1.4 to 1.6) **If Y/PY**, answer questions relating to both baseline and time-varying confounding (1.7 and 1.8)	If intervention switches are unrelated to the outcome, for example when the outcome is an unexpected harm, then time-varying confounding will not be present and only control for baseline confounding is required.	NA / Y / PY / PN / N / NI

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Signalling questions	Elaboration	Response options
Questions relating to baseline confounding only		
1.4. Did the authors use an appropriate analysis method that controlled for all the important confounding domains?	Appropriate methods to control for measured confounders include stratification, regression, matching, standardization, and inverse probability weighting. They may control for individual variables or for the estimated propensity score. Inverse probability weighting is based on a function of the propensity score. Each method depends on the assumption that there is no unmeasured or residual confounding.	NA / Y / PY / PN / N / NI
1.5. **If Y/PY to 1.4**: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?	Appropriate control of confounding requires that the variables adjusted for are valid and reliable measures of the confounding domains. For some topics, a list of valid and reliable measures of confounding domains will be specified in the review protocol but for others such a list may not be available. Study authors may cite references to support the use of a particular measure. If authors control for confounding variables with no indication of their validity or reliability pay attention to the subjectivity of the measure. Subjective measures (e.g. based on self-report) may have lower validity and reliability than objective measures such as lab findings.	NA / Y / PY / PN / N / NI
1.6. Did the authors control for any post-intervention variables that could have been affected by the intervention?	Controlling for post-intervention variables that are affected by intervention is not appropriate. Controlling for mediating variables estimates the direct effect of intervention and may introduce bias. Controlling for common effects of intervention and outcome introduces bias.	NA / Y / PY / PN / N / NI
Questions relating to baseline and time-varying confounding		
1.7. Did the authors use an appropriate analysis method that controlled for all the important confounding domains and for time-varying confounding?	Adjustment for time-varying confounding is necessary to estimate the effect of starting and adhering to intervention, in both randomized trials and NRSI. Appropriate methods include those based on inverse probability weighting. Standard regression models that include time-updated confounders may be problematic if time-varying confounding is present.	NA / Y / PY / PN / N / NI
1.8. **If Y/PY to 1.7**: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?	See 1.5 above.	NA / Y / PY / PN / N / NI

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Risk of bias judgement Optional: What is the predicted direction of bias due to confounding?	See Table 5. Can the true effect estimate be predicted to be greater or less than the estimated effect in the study because one or more of the important confounding domains was not controlled for? Answering this question will be based on expert knowledge and results in other studies and therefore can only be completed after all of the studies in the body of evidence have been reviewed. Consider the potential effect of each of the unmeasured domains and whether all important confounding domains not controlled for in the analysis would be likely to change the estimate in the same direction, or if one important confounding domain that was not controlled for in the analysis is likely to have a dominant impact.	Low / Moderate / Serious / Critical / NI Favours experimental / Favours comparator / Unpredictable

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 5: Reaching risk of bias judgements for bias due to confounding

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	No confounding expected.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) Confounding expected, all known important confounding domains appropriately measured and controlled for; *and* (ii) Reliability and validity of measurement of important domains were sufficient, such that we do not expect serious residual confounding.
Serious risk of bias (the study has some important problems)	(i) At least one known important domain was not appropriately measured, or not controlled for; *or* (ii) Reliability or validity of measurement of an important domain was low enough that we expect serious residual confounding.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(i) Confounding inherently not controllable *or* (ii) The use of negative controls strongly suggests unmeasured confounding.
No information on which to base a judgement about risk of bias for this domain	No information on whether confounding might be present.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

4.2 Detailed guidance: Bias in selection of participants into the study

4.2.1 Introduction

Selection bias occurs when some eligible participants, or the initial follow up time of some participants, or some outcome events, are excluded in a way that leads to the association between intervention and outcome differing from the association that would have been observed in the target trial. As explained in section 2.3, **this phenomenon is distinct from that of confounding**, although the term selection bias is sometimes used to mean confounding.

Our use of the term "selection bias" is intended to refer only to **biases that are internal to the study**, and **not to issues of indirectness (generalizability, applicability** or **transferability** to people who were excluded from the study) (Schunemann et al, 2013). For example, restricting the study sample to individuals free of comorbidities may limit the generalizability of its findings to clinical practice, where comorbidities are common. However it does not bias the estimated effect of intervention in individuals free of comorbidities.

4.2.2 When selection of participants into the study may introduce bias

Selective recruitment of participants into a study does not necessarily cause bias. For example, consider a study that selected (at random) only half of eligible men, but included all eligible women. The effect of intervention in men will be less precisely estimated than if all men had been included, but if the true effect of the intervention in women is the same as in men, the effect estimate will not be biased by the selection.

Selection bias occurs when selection of participants is **related to both intervention and outcome**. For example, studies of folate supplementation to prevent neural tube defects were biased because they were restricted to live births (Hernán et al, 2002). The bias arises because stillbirths and therapeutic abortions (which were excluded from the sample), are related to both the intervention and the outcome (Velie and Shaw, 1996, Hernán et al, 2002).

For the same reason, selection bias can occur when some follow up time is excluded from the analysis. For example, consider the potential for bias when prevalent, rather than new (incident), users of the intervention are included in analyses. This is analogous to starting the follow-up of the target trial some time after the start of intervention, so that some individuals who experienced the outcome after starting the intervention will have been excluded. This is a type of selection bias that has also been termed **inception bias** or **lead time bias**. If participants are not followed from the start of the intervention (inception), as they would be in a randomized trial, then a period of follow up has been excluded, and individuals who experienced the outcome soon after intervention will be missing from analyses. The key problem is that there is no reason to expect the effect of the intervention to be constant over time. Therefore, excluding follow up immediately after intervention may bias the estimated effect either upwards or downwards. Studies that report estimates of the effect of intervention stratified into follow up periods may provide information on the extent to which the effect of intervention varies with time since the start of intervention.

For example, analyses based on prevalent users of a drug may tend to select those who tolerate the drug well: "depletion of the susceptible" will already have taken place. As a result we may underestimate the rate of adverse effects in the intervention group: pharmacoepidemiological studies therefore often specify that there should have been no record of use of the drug in the previous 12 months. For example, there was an apparently increased risk of venous thromboembolism with the newer oral contraceptive progestogens when investigated in NRSI (Ray *et al*, 2003; Suissa *et al*, 2000).

Users of the newer agents had started treatment more recently than users of older agents and the risk of venous thromboembolism is greatest early in the course of treatment. Contemporary methodological standards emphasize the importance both of identifying cohorts of new users of health technologies and of commencing follow-up from the date of the treatment decision, not commencement of treatment, in order to avoid biases like this (Ray *et al*, 2003; Suissa, 2008).

A related bias – **immortal time bias** – occurs when the interventions are defined in such a way that there is a period of follow up during which the outcome cannot occur. For example, a study followed cohorts of subjects with chronic obstructive pulmonary disease or chronic heart failure and considered them to be in two groups according to whether they received telehomecare or standard care. However, to get telehomecare, patients had to survive for several weeks after the index hospitalization: therefore the time between hospitalization and start of telehomecare was "immortal time". Exclusion of this follow up period, and of the deaths that occur during the

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

period, will bias the study towards finding that telehomecare reduces mortality. Comparison with a target trial should facilitate identification of such bias, because in a trial participants would be followed from the time of randomization even if implementation of intervention occurred some time later.

4.2.3 Technical note: adjusting for selection bias

There are analytic approaches to adjust for these types of selection bias and statistical analyses that protect against selection bias. The key issue is whether measured variables that permit meaningful adjustment (for example via inverse-probability-weighting) are available. In many situations this will not be the case, so that these design issues will lead to a classification of serious or critical risk of bias.

4.2.4 Risk of bias assessment for bias in selection of participants into the study

The signalling questions and risk of bias assessments are given in Box 5 and Table 6.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 5: The ROBINS-I tool (Stage 2, part 4): Risk of bias in selection of participants into the study

Signalling questions	Elaboration	Response options
2.1. Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention? **If N/PN to 2.1:** go to 2.4	This domain is concerned only with selection into the study based on participant characteristics observed *after* the start of intervention. Selection based on characteristics observed *before* the start of intervention can be addressed by controlling for imbalances between experimental intervention and comparator groups in baseline characteristics that are prognostic for the outcome (baseline confounding).	Y / PY / PN / N / NI
2.2. **If Y/PY to 2.1**: Were the post-intervention variables that influenced selection likely to be associated with intervention?	Selection bias occurs when selection is related to an effect of either intervention or a cause of intervention **and** an effect of either the outcome or a cause of the outcome. Therefore, the result is at risk of selection bias if selection into the study is related to both the intervention and the outcome.	NA / Y / PY / PN / N / NI
2.3 **If Y/PY to 2.2**: Were the post-intervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome?		NA / Y / PY / PN / N / NI
2.4. Do start of follow-up and start of intervention coincide for most participants?	If participants are not followed from the start of the intervention then a period of follow up has been excluded, and individuals who experienced the outcome soon after intervention will be missing from analyses. This problem may occur when prevalent, rather than new (incident), users of the intervention are included in analyses.	Y / PY / PN / N / NI
2.5. **If Y/PY to 2.2 and 2.3, or N/PN to 2.4**: Were adjustment techniques used that are likely to correct for the presence of selection biases?	It is in principle possible to correct for selection biases, for example by using inverse probability weights to create a pseudo-population in which the selection bias has been removed, or by modelling the distributions of the missing participants or follow up times and outcome events and including them using missing data methodology. However such methods are rarely used and the answer to this question will usually be "No".	NA / Y / PY / PN / N / NI
Risk of bias judgement	See Table 6.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to selection of participants into the study?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 6: Reaching risk of bias judgements in selection of participants into the study

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	(i) All participants who would have been eligible for the target trial were included in the study; *and* (ii) For each participant, start of follow up and start of intervention coincided.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) Selection into the study may have been related to intervention and outcome; *and* The authors used appropriate methods to adjust for the selection bias; *or* (ii) Start of follow up and start of intervention do not coincide for all participants; *and* (a) the proportion of participants for which this was the case was too low to induce important bias; *or* (b) the authors used appropriate methods to adjust for the selection bias; *or* (c) the review authors are confident that the rate (hazard) ratio for the effect of intervention remains constant over time.
Serious risk of bias (the study has some important problems)	(i) Selection into the study was related (but not very strongly) to intervention and outcome; *and* This could not be adjusted for in analyses; *or* (ii) Start of follow up and start of intervention do not coincide; *and* A potentially important amount of follow-up time is missing from analyses; *and* The rate ratio is not constant over time.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(i) Selection into the study was very strongly related to intervention and outcome; *and* This could not be adjusted for in analyses; *or* (ii) A substantial amount of follow-up time is likely to be missing from analyses; *and* The rate ratio is not constant over time.
No information on which to base a judgement about risk of bias for this domain	No information is reported about selection of participants into the study or whether start of follow up and start of intervention coincide.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

4.3 Detailed guidance: Bias in classification of interventions

4.3.1 Introduction

Bias may be introduced if intervention status is misclassified. This is seldom a problem in randomized trials and other experimental studies, because interventions are actively assigned by the researcher and their accurate recording is a key feature of the study. However, in observational studies information about interventions allocated or received needs to be collected.

Possible methods for data collection include:

- systematic assessment of patients (clinical examinations, interviews, diagnostic tests);
- administrative or in-house databases (prospective recording of data with no pre-specified purpose);
- extraction from medical records; and
- organizational records or policy documents (e.g. for organizational or public health interventions).

4.3.2 Differential and non-differential misclassification

Misclassification of intervention status may be non-differential or differential. **Non-differential misclassification** is unrelated to the outcome: for example in a comparison of smoke alarm installation with no smoke alarm installation, receipt of intervention may be incompletely recorded so that some people who installed a smoke alarm are incorrectly allocated to the "no alarm" group. Provided that such misclassification is unrelated to subsequent outcomes (e.g. risk of fire-related injury is unrelated to the reasons for failing to identify smoke alarm installation), the misclassification is non-differential and will usually bias the estimated effect of intervention towards the null (no intervention effect or no difference between interventions).

Differential misclassification occurs when misclassifications of intervention status is related to subsequent outcome or to the risk of the outcome. It is important that, wherever possible, interventions are defined and categorized without knowledge of subsequent outcomes. A well-known example of differential misclassification, when this might not be the case, is **recall bias** in a case-control study, whereby knowledge of case-control status affects recall of previous intervention: typically the cases are more likely than controls to recall potentially important events.

Differential misclassification can occur in cohort studies, if information about intervention status is obtained retrospectively. This can happen if the information (or availability of information) on intervention status is influenced by outcomes: for example a cohort study in elderly people in which the outcome is dementia, and participants' recall of past intervention status at study inception was affected by pre-existing mild cognitive impairment. Alternatively, a research assistant may search more diligently for past intervention status when the participant has dementia. Other mechanisms may lead to differential misclassification of intervention status. For instance, information on the vaccination status of children in parts of Africa is collected by examining vaccination cards on periodic visits to family homes, and if no card is found, a child is assumed to be unvaccinated. In some cultures, vaccination cards are destroyed if a child dies. Vaccination status for such children may be differentially misclassified if they are analysed as unvaccinated in studies of the effect of vaccination on mortality. Such problems can be avoided if information about intervention status is collected at the time of the intervention and the information is complete and accessible to those undertaking the NRSI.

4.3.3 Risk of bias assessment for bias in classification of interventions

The signalling questions and risk of bias assessments are given in Box 6 and Table 7.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 6: The ROBINS-I tool (Stage 2, part 5): Risk of bias in classification of interventions

Signalling questions	Elaboration	Response options
3.1 Were intervention groups clearly defined?	A pre-requisite for an appropriate comparison of interventions is that the interventions are well defined. Ambiguity in the definition may lead to bias in the classification of participants. For individual-level interventions, criteria for considering individuals to have received each intervention should be clear and explicit, covering issues such as type, setting, dose, frequency, intensity and/or timing of intervention. For population-level interventions (e.g. measures to control air pollution), the question relates to whether the population is clearly defined, and the answer is likely to be 'Yes'.	Y / PY / PN / N / NI
3.2 Was the information used to define intervention groups recorded at the start of the intervention?	In general, if information about interventions received is available from sources that could not have been affected by subsequent outcomes, then differential misclassification of intervention status is unlikely. Collection of the information at the time of the intervention makes it easier to avoid such misclassification. For population-level interventions (e.g. measures to control air pollution), the answer to this question is likely to be 'Yes'.	Y / PY / PN / N / NI
3.3 Could classification of intervention status have been affected by knowledge of the outcome or risk of the outcome?	Collection of the information at the time of the intervention may not be sufficient to avoid bias. The way in which the data are collected for the purposes of the NRSI should also avoid misclassification.	Y / PY / PN / N / NI
Risk of bias judgement	See Table 7.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to measurement of outcomes or interventions?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 7: Reaching risk of bias judgements for bias in classification of interventions

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	(i) Intervention status is well defined; *and* (ii) Intervention definition is based solely on information collected at the time of intervention.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) Intervention status is well defined; *and* (ii) Some aspects of the assignments of intervention status were determined retrospectively.
Serious risk of bias (the study has some important problems)	(i) Intervention status is not well defined; *or* (ii) Major aspects of the assignments of intervention status were determined in a way that could have been affected by knowledge of the outcome.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(Unusual) An extremely high amount of misclassification of intervention status, e.g. because of unusually strong recall biases.
No information on which to base a judgement about risk of bias for this domain	No definition of intervention or no explanation of the source of information about intervention status is reported.

4.4 Detailed guidance: Bias due to deviations from intended interventions

4.4.1 Introduction

We consider in this domain biases that arise **when there are systematic differences between the care provided to experimental intervention and comparator groups, beyond the assigned interventions**. These differences reflect additional aspects of care, or intended aspects of care that were not delivered.

It is important to distinguish between:

(a) deviations from intended intervention that arise because of knowledge of the intervention applied and **because of expectation of finding a difference between experimental intervention and comparator consistent with the hypothesis being tested in the study**. Such deviations are not part of usual practice.

(b) deviations from intended intervention that happen during usual clinical care following the intervention (for example, cessation of a drug intervention because of acute toxicity); and

4.4.2 The importance of the nature of the effect of interest

The extent to which these considerations are associated with bias depends on the nature of the effect of interest.

Deviations of the first type, (a) above, are always of concern. For example, a study compared infection rates after insertion by cardiologists of two different permanent cardiac pacemaker devices. It was not routine to give prophylactic antibiotics at the participating institutions. Blinding was not feasible. Some cardiologists believed that device A would have a higher infection rate than device B and, as a result, administered antibiotics to patients receiving device A more often than to patients receiving device B. These deviations from intended intervention did not reflect usual clinical care (type b) – they reflected cardiologists' expectations of differences in infection rates between the two devices (type a). The result of the study is at risk of bias, whichever the effect of interest.

The importance of deviations of type (b) depends on the nature of the effect of interest. If the goal is the unbiased estimation of the effect of *assignment to* (or starting) intervention, then there will be no bias due to deviation

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

from the intended interventions for deviations of type (b). Specifically, if all deviations from intended intervention are part of usual practice, then we can still evaluate the effect of assignment to intervention, regardless of the actual implementation of the interventions.

On the other hand, if the goal is the unbiased estimation of *starting and adhering* to intervention, then all deviations from the target "protocol" will lead to bias. For example, an open-label study compared respiratory tract infection (RTI) rates after minimally invasive or open surgery for oesophageal cancer. There were two important differences between intervention groups in the delivery of co-interventions. First, one-lung mechanical ventilation (which is thought to increase respiratory complications, including RTIs) was used in the open surgery group, whereas the minimally invasive group underwent two lung ventilation. Second, epidural analgesia was used more frequently in the open surgery group: patients with epidurals are generally less mobile and thus at increased risk of developing an RTI. These deviations from the intended interventions put the result of the study at risk of bias in relation to the effect of starting and adhering to the intended interventions.

4.4.3 Types of deviations from intended intervention

Biases that arise due to deviations from intended interventions are sometimes referred to as **performance biases**. They arise, in both randomized trials and NRSI, when systematic differences between the care provided to experimental intervention and comparator groups **occur after the start of intervention**, and the participant continues (for analysis purposes) to be part of the intended intervention group.

Technical aside: Considerations of the risk of performance bias are thus distinct from confounding. Note that methods that adjust for time-varying confounding may be used to adjust both for switches between interventions of interest (addressed under confounding in ROBINS-I) and for deviations from intended interventions (addressed in this bias domain). Such methods can rely on sufficient data on predictors of switches between or deviations from interventions having been collected.

In randomized trials, performance bias can sometimes be reduced or avoided by **blinding** of participants and healthcare providers. Blinding does not generally occur in NRSI: thus both patients and healthcare providers are typically aware of the interventions that are being implemented.

Knowledge of the intervention assignment may influence the likelihood of **co-interventions** (receipt of interventions other than the studied interventions, whose frequency may differ between intervention groups), compromised fidelity of **implementation** (i.e. failure to implement some or all of the intervention as intended by the health care professionals delivering care during the trial), and **adherence** to the intervention by patients or participants. Failures in implementation or adherence include *contamination* (the inadvertent application of one of the studied interventions in participants intended to receive the other), and *switches* from the intended interventions to other interventions or to none.

4.4.4 Deviations from intended intervention when assessing the effect of starting and adhering to intervention

Consideration of co-interventions, implementation of the intervention and adherence by participants should be assessed only when interest is in the effect of starting and adhering to the intervention.

4.4.4.1 Considerations for co-interventions

Co-interventions are a potentially important source of bias. For example, consider an observational study comparing rates of post-operative infection in patients who received one of two surgical procedures A and B. If antibiotic prophylaxis was provided for patients receiving A but not those receiving B, lower rates of post-operative infection observed in patients receiving A might be attributable to antibiotic prophylaxis rather than to the surgical procedure, and there is a potential for bias. This is not the case if the specified target trial compares intervention A plus antibiotics with intervention B without antibiotics. Similarly, a "pragmatic" target trial might allow for opportunistic use of antibiotics as medically indicated, whereas an "explanatory" trial comparing the effects of A versus B alone might aim for balance in the use of antibiotics.

A co-intervention is defined as a new intervention that is **not** part of intended intervention. It is important to consider what is normal or usual practice for the intended intervention before determining the presence of co-interventions. For example, the normal administration of a drug treatment for diabetes may require monitoring to allow for adjustments to the dose or addition of another drug. These adjustments are therefore not a deviation

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

from the intended intervention. Similarly, addition of other treatments aimed at diabetes control may be pre-specified as part of usual clinical practice in the context of the intended intervention.

In some instances the protocol for the intended intervention specifies that the addition of other treatments is at the discretion of physicians, but such additions differ between the intervention groups. For example, consider a cohort study comparing rates of gastrointestinal ulcers in aspirin users and non-users. The use of proton pump inhibitors (PPIs) to prevent bleeding in those taking aspirin is part of usual practice. If their use is pre-specified in the study protocol then the comparison is of aspirin plus PPIs (as necessary) with non-use of aspirin, and the PPIs should not be considered a co-intervention. However if the study aims to compare aspirin use with no use, then PPIs may be considered a co-intervention because their greater use in the aspirin group leads to an underestimate of the effect of aspirin on gastrointestinal ulcers. Similarly, PPIs cause diarrhoea, and a higher frequency of diarrhoea in aspirin users may be due to proton pump inhibitor co-intervention, rather than the aspirin.

Review authors should make in advance a list of important co-interventions that could differ between intervention groups and could have an impact on study outcomes (see section 3.1.3). They should then consider whether they are likely to be administered in the context of each particular study.

We suggest that review authors consider whether the critical co-interventions are balanced between intervention groups. If effective co-interventions are **not** balanced, performance bias is probably present. However, if the co-interventions are balanced across intervention groups, there is still a risk that intervention groups will differ in their management or behaviour beyond the intervention comparison of interest. This will be the case if a co-intervention interacts with the experimental and comparator intervention in different ways (for example, it enhances the effect of one intervention, but has no effect on the other).

4.4.4.2 Considerations for fidelity of implementation of intended interventions

Reasons that an intervention is not implemented as intended by the study investigators include problems with (i) adherence to protocols by investigators; (ii) technical problems with the intervention if it is complex or relies on operator skill; and (iii) differences in the context of the study. Problems with fidelity of implementation can occur in one or any of the intervention groups. The implementation of the intended treatment protocol and adherence to intervention by study participants cannot always be disentangled, as one may influence the other. For example, an intervention can be administered in a manner that does not encourage adherence by study participants.

An example of unsuccessful implementation of an intervention is provided by a study comparing three complex interventions for adolescents with depression and a history of substance abuse: (1) antidepressant A combined with cognitive behavioural therapy (CBT); (2) drug B combined with CBT; and (3) CBT alone. Although therapists were sent the protocol for the CBT, there was no specific training with regards to the administration of CBT, which led to potential between-therapist differences in the content and coverage of the CBT intervention. Neither was there a checklist for the content of CBT sessions. In addition, different therapists, with differing degrees of experience and professional training with CBT, administered the CBT within each of the intervention groups. These problems mean that the study is at risk of bias due to lack of fidelity of implementation of the intended interventions.

Review authors should consider the details of the intervention with respect to how, when, where and to whom it is applied. Some key features to consider when assessing the risk of bias associated with lack of implementation fidelity with the intended intervention includes evaluation of the following:

a) ***Practitioner:*** characteristics of those administering the intervention (e.g. staff characteristics, level of expertise and training, potential therapeutic alliances) and opportunities for those implementing the intervention to modify the intended protocol (e.g. physician will decrease dose because of potential for adverse events);
b) ***Intervention:*** core components of the intervention or comparator as it was intended to be delivered within the context of the primary study (e.g. content of the intervention); this includes the complexity of the intervention (e.g. it has multiple components that may adversely affect adherence), the sequence and order of how it is delivered, the dosage, duration, or format (e.g. phone follow-up rather than face to face meeting), and operational definitions of treatment;
c) ***Context:*** characteristics of the healthcare setting (e.g. public outpatient versus hospital outpatient), organizational service structure (e.g. managed care or publically funded program), geographical setting (e.g. rural vs urban), and cultural setting and the legal environment where the intervention is implemented.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

The assessment aims to determine whether the intervention and comparator were implemented as intended by the study investigators. This should be evaluated for each intervention group. Consider also if the study design and analysis attempts to minimize the impact of inadvertent application of the unintended intervention.

Differences in how patients are monitored may affect the fidelity of the intervention. For instance, in a cohort study assessing adverse effects of spironolactone compared to its non-use, physicians who recognize that spironolactone can increase serum potassium may choose more frequent monitoring of serum potassium with subsequent dose adjustment before serum potassium reaches abnormal levels.

4.4.4.3 Considerations for adherence to intervention

Risk of bias will be higher if participants did not adhere to the intervention as intended. Lack of adherence includes imperfect compliance, cessation of intervention, crossovers to the comparator intervention and switches to another active intervention. The likelihood of non-adherence will differ according to the nature of the interventions being evaluated. For example, poor adherence to pharmaceutical interventions can be frequent, and multiple switches between interventions (or between taking and not taking a pharmaceutical) can occur within one individual. However, time-varying non-adherence is unlikely in comparisons of surgical interventions such as heart valves or joint prostheses.

Users of the tool should consider the interventions being compared. Is there a potential for people receiving one intervention to receive more or less than was intended, to stop intervention, or to switch to other interventions? Are multiple switches possible or likely? It is important to consider the overall rates of non-adherence within each group and determine if this may impact on the study outcomes. The threshold of non-adherence likely to impact the outcomes will vary with the type of intervention and the study design.

4.4.4.4 Technical note: adjustment for departures from intended interventions

In Section 4.1.9 we briefly described statistical approaches to estimate the effect of starting and adhering to the intervention, allowing for switches between the interventions of interest, using methods that adjust for time-varying confounding. Related methods can be used to allow for deviations from intended interventions. One approach is to **censor follow up at the time that the deviation occurs**. It is necessary to use statistical methods that avoid the bias (technically, this is a type of selection bias) that can result from such censoring, for example through inverse-probability weighting. For example, consider a 12 month study examining the effect of starting and adhering to use of selective serotonin-reuptake inhibitor (SSRI) A versus no intervention. If some patients switched to SSRI B during follow up then there is a risk of performance bias. This might be dealt with by censoring follow up on receipt of SSRI B, but the analysis would then need to allow for the possibility that patients who switch to SSRI B are systematically different from those who remain on SSRI A.

Alternatively, consider a study that aims to estimate the effect of starting and adhering to continuous intervention with A versus continuous intervention with B, in which no participants change from A to B or from B to A after baseline. If a greater proportion of those assigned to A also take concomitant intervention C during parts of the follow up then there is time-varying confounding by C. Providing that the prognostic factors that predict intervention with C are measured over time, methods that adjust for time-varying confounding (see section 4.1.9) can be used to adjust for the bias due to time-varying confounding that is caused by the imbalance in use of intervention C.

Typically for ethical reasons, a study protocol will permit some changes in the intervention based on participants' health status during follow up; this occurs even in trials where randomization occurs. Such changes could include an alteration of the dose or type of intervention that is provided throughout the remainder of the study. For example, in a study comparing two medications to control blood glucose in diabetic patients, there are repeated measurements of glycaemic response over time (e.g. HBA1c, blood pressure) to monitor response to intervention. If during the normal course of treatment these show poor control, then the clinician will alter the medication (for example, increasing medication dose if HBA1c is too high). However glycaemic response may also influence the primary outcome (e.g. cardiovascular mortality). No statistical adjustment is necessary if the protocol for the target trial allows for modification of dose in response to glycaemic index. However specialist statistical methods (see above) are required to deal with the problem of time-varying confounding by glycaemic index in studies estimating of the effect of continuous treatment with the initial medication dose.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

4.4.5 Risk of bias assessment for bias due to deviations from intended interventions

The signalling questions and risk of bias assessments are given in Box 7 and Table 8.

We are aware that review authors would find it extremely useful if we could provide guidance on criteria that should be used to judge co-interventions to be "balanced", or the amount of adherence to intervention to be "high". Unfortunately, we do not believe that simple guidance will be generally applicable: a small absolute difference in the numbers of patients receiving an important co-intervention might lead to substantial bias if the co-intervention strongly influenced the outcome and patients in whom the outcome occurred were usually those who received the co-intervention.

We recommend that review teams ensure that judgements of "balanced" co-intervention, "successful" implementation and lack of adherence are applied consistently across the studies included in their review.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 7: The ROBINS-I tool (Stage 2, part 6): Risk of bias due to deviations from intended interventions

Signalling questions	Elaboration	Response options
If your aim for this study is to assess the effect of assignment to intervention, answer questions 4.1 and 4.2		
4.1. Were there deviations from the intended intervention beyond what would be expected in usual practice?	Deviations that happen in usual practice following the intervention (for example, cessation of a drug intervention because of acute toxicity) are part of the intended intervention and therefore do not lead to bias in the effect of assignment to intervention. Deviations may arise due to expectations of a difference between intervention and comparator (for example because participants feel unlucky to have been assigned to the comparator group and therefore seek the active intervention, or components of it, or other interventions). Such deviations are not part of usual practice, so may lead to biased effect estimates. However these are not expected in observational studies of individuals in routine care.	Y / PY / PN / N / NI
4.2. **If Y/PY to 4.1**: Were these deviations from intended intervention unbalanced between groups *and* likely to have affected the outcome?	Deviations from intended interventions that do not reflect usual practice will be important if they affect the outcome, but not otherwise. Furthermore, bias will arise only if there is imbalance in the deviations across the two groups.	NA / Y / PY / PN / N / NI
If your aim for this study is to assess the effect of starting and adhering to intervention, answer questions 4.3 to 4.6		
4.3. Were important co-interventions balanced across intervention groups?	Risk of bias will be higher if unplanned co-interventions were implemented in a way that would bias the estimated effect of intervention. Co-interventions will be important if they affect the outcome, but not otherwise. Bias will arise only if there is imbalance in such co-interventions between the intervention groups. Consider the co-interventions, including any pre-specified co-interventions, that are likely to affect the outcome and to have been administered in this study. Consider whether these co-interventions are balanced between intervention groups.	Y / PY / PN / N / NI
4.4. Was the intervention implemented successfully for most participants?	Risk of bias will be higher if the intervention was not implemented as intended by, for example, the health care professionals delivering care during the trial. Consider whether implementation of the intervention was successful for most participants.	Y / PY / PN / N / NI

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Signalling questions	Elaboration	Response options
4.5. Did study participants adhere to the assigned intervention regimen?	Risk of bias will be higher if participants did not adhere to the intervention as intended. Lack of adherence includes imperfect compliance, cessation of intervention, crossovers to the comparator intervention and switches to another active intervention. Consider available information on the proportion of study participants who continued with their assigned intervention throughout follow up, and answer 'No' or 'Probably No' if this proportion is high enough to raise concerns. Answer 'Yes' for studies of interventions that are administered once, so that imperfect adherence is not possible. We distinguish between analyses where follow-up time after interventions switches (including cessation of intervention) is assigned to (1) the new intervention or (2) the original intervention. (1) is addressed under time-varying confounding, and should not be considered further here.	Y / PY / PN / N / NI
4.6. **If N/PN to 4.3, 4.4 or 4.5**: Was an appropriate analysis used to estimate the effect of starting and adhering to the intervention?	It is possible to conduct an analysis that corrects for some types of deviation from the intended intervention. Examples of appropriate analysis strategies include inverse probability weighting or instrumental variable estimation. It is possible that a paper reports such an analysis without reporting information on the deviations from intended intervention, but it would be hard to judge such an analysis to be appropriate in the absence of such information. Specialist advice may be needed to assess studies that used these approaches. If everyone in one group received a co-intervention, adjustments cannot be made to overcome this.	NA / Y / PY / PN / N / NI
Risk of bias judgement	See Table 8.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to deviations from the intended interventions?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 8: Reaching risk of bias judgements for bias due to deviations from intended interventions

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	**Effect of assignment to intervention:** (i) Any deviations from intended intervention reflected usual practice; *or* (ii) Any deviations from usual practice were unlikely to impact on the outcome. **Effect of starting and adhering to intervention:** The important **co-interventions** were balanced across intervention groups, and there were no deviations from the intended interventions (in terms of **implementation or adherence**) that were likely to impact on the outcome.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	**Effect of assignment to intervention:** There were deviations from usual practice, but their impact on the outcome is expected to be slight. **Effect of starting and adhering to intervention:** (i) There were deviations from intended intervention, but their impact on the outcome is expected to be slight. *or* (ii) The important co-interventions were not balanced across intervention groups, or there were deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *and* The analysis was appropriate to estimate the effect of starting and adhering to intervention, allowing for deviations (in terms of implementation, adherence and co-intervention) that were likely to impact on the outcome.
Serious risk of bias (the study has some important problems)	**Effect of assignment to intervention:** There were deviations from usual practice that were unbalanced between the intervention groups and likely to have affected the outcome. **Effect of starting and adhering to intervention:** (i) The important co-interventions were not balanced across intervention groups, or there were deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *and* (ii) The analysis was not appropriate to estimate the effect of starting and adhering to intervention, allowing for deviations (in terms of implementation, adherence and co-intervention) that were likely to impact on the outcome.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	**Effect of assignment to intervention:** There were substantial deviations from usual practice that were unbalanced between the intervention groups and likely to have affected the outcome. **Effect of starting and adhering to intervention:** (i) There were substantial imbalances in important co-interventions across intervention groups, or there were substantial deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *and* (ii) The analysis was not appropriate to estimate the effect of starting and adhering to intervention, allowing for deviations (in terms of implementation, adherence and co-intervention) that were likely to impact on the outcome.
No information on which to base a judgement about risk of bias for this domain	No information is reported on whether there is deviation from the intended intervention.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

4.5 Detailed guidance: Bias due to missing data

4.5.1 Introduction

Missing data may arise, among other reasons, through attrition (loss to follow up), missed appointments, incomplete data collection and by participants being excluded from analysis by primary investigators. In NRSI, data may be missing for baseline characteristics (including interventions received or baseline confounders), for outcome measurements, for other variables involved in the analysis or a combination of these. A general rule for consideration of bias due to missing data is that we should assume that an analysis using the data we intended to collect (were they available) would produce an unbiased effect estimate, so that we concentrate only on biases that might be introduced by the missing data.

The starting point for considering risk of bias due to missing outcome data is to clarify the nature of the comparison of interest, particularly with regard to the distinction between *assignment to* (or start of) intervention and *starting and adhering to* intervention (see section 3.2.2). For example, the "complete" data set would be different for a comparison between those who were and were not *offered* screening and a comparison between those who did and did not *attend* screening. Therefore the definition of missing data would also be different. In order to consider missing outcome data, it is therefore important that a study sample is clearly defined at the outset. This can be achieved through consideration of the target randomized trial.

4.5.2 Differential missingness

Specific considerations for missing data broadly follow those established for randomized trials and described in the existing Cochrane RoB tool for randomized trials. Differentials in missing data between intervention groups are key, along with the reasons for data being missing. If (i) the proportion of missing data and (ii) the reasons for missing data are similar across intervention groups, then there would typically be only limited bias in the effect estimate, so that risk of bias would be considered low or moderate (see section 3.4). As the proportion of missing data increases, differences in response to intervention may increase concerns about the potential for bias. While (i) can usually be established from the reported data, (ii) is typically a judgement of the review authors. Given this, balance in proportions of missingness across intervention groups alone provides only moderate reassurance about the risk of bias.

4.5.3 Adverse effects

When looking at unintended effects, an important consideration is whether the review authors are satisfied that follow-up has not systematically excluded non-trivial proportions of individuals in whom adverse effects may be prevalent. For instance, if older people drop out (or miss appointments) more, and also have more adverse events, then a large proportion of adverse events may be missing from the analysis. This will not necessarily introduce bias, although bias would result if the older people are more likely to drop out of one intervention group than the other. This might occur, for example, in a comparison of exercise versus crossword puzzles to prevent cognitive decline.

4.5.4 Risk of bias assessment for bias due to missing data

The signalling questions and risk of bias assessments are given in Box 8 and Table 9.

We are aware that review authors would find it extremely useful if we could provide guidance on the extent of missing data that should lead to the conclusion that a result is at moderate or high risk of bias. For example, a criterion of less than 80% completeness of follow up has been used as a threshold in some guidance. Unfortunately, we do not believe that a single threshold can be meaningfully defined: for example a result based on 95% complete outcome data might still be at high risk of bias if the outcome was rare and if reasons for missing outcome data were strongly related to intervention group.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 8: The ROBINS-I tool (Stage 2, part 7): Risk of bias due to missing data

Signalling questions	Elaboration	Response options
5.1 Were outcome data available for all, or nearly all, participants?	"Nearly all" should be interpreted as "enough to be confident of the findings", and a suitable proportion depends on the context. In some situations, availability of data from 95% (or possibly 90%) of the participants may be sufficient, providing that events of interest are reasonably common in both intervention groups. One aspect of this is that review authors would ideally try and locate an analysis plan for the study.	Y / PY / PN / N / NI
5.2 Were participants excluded due to missing data on intervention status?	Missing intervention status may be a problem. This requires that the *intended* study sample is clear, which it may not be in practice.	Y / PY / PN / N / NI
5.3 Were participants excluded due to missing data on other variables needed for the analysis?	This question relates particularly to participants excluded from the analysis because of missing information on confounders that were controlled for in the analysis.	Y / PY / PN / N / NI
5.4 **If PN/N to 5.1, or Y/PY to 5.2 or 5.3**: Are the proportion of participants and reasons for missing data similar across interventions?	This aims to elicit whether either (i) differential proportion of missing observations or (ii) differences in reasons for missing observations could substantially impact on our ability to answer the question being addressed. "Similar" includes some minor degree of discrepancy across intervention groups as expected by chance.	NA / Y / PY / PN / N / NI
5.5 **If PN/N to 5.1, or Y/PY to 5.2 or 5.3**: Is there evidence that results were robust to the presence of missing data?	Evidence for robustness may come from how missing data were handled in the analysis and whether sensitivity analyses were performed by the investigators, or occasionally from additional analyses performed by the systematic reviewers. It is important to assess whether assumptions employed in analyses are clear and plausible. Both content knowledge and statistical expertise will often be required for this. For instance, use of a statistical method such as multiple imputation does not guarantee an appropriate answer. Review authors should seek naïve (complete-case) analyses for comparison, and clear differences between complete-case and multiple imputation-based findings should lead to careful assessment of the validity of the methods used.	NA / Y / PY / PN / N / NI
Risk of bias judgement	See Table 9.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to missing data?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 9: Reaching risk of bias judgements for bias due to missing data

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	(i) Data were reasonably complete; *or* (ii) Proportions of and reasons for missing participants were similar across intervention groups; *or* (iii) The analysis addressed missing data and is likely to have removed any risk of bias.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) Proportions of and reasons for missing participants differ slightly across intervention groups; *and* (ii) The analysis is unlikely to have removed the risk of bias arising from the missing data.
Serious risk of bias (the study has some important problems)	(i) Proportions of missing participants differ substantially across interventions; *or* Reasons for missingness differ substantially across interventions; *and* (ii) The analysis is unlikely to have removed the risk of bias arising from the missing data; *or* Missing data were addressed inappropriately in the analysis; *or* The nature of the missing data means that the risk of bias cannot be removed through appropriate analysis.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(i) (Unusual) There were critical differences between interventions in participants with missing data; *and* (ii) Missing data were not, or could not, be addressed through appropriate analysis.
No information on which to base a judgement about risk of bias for this domain	No information is reported about missing data or the potential for data to be missing.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

4.6 Detailed guidance: Bias in measurement of outcomes

4.6.1 Introduction

Bias may be introduced if outcomes are misclassified or measured with error. Possible methods for data collection include:

- systematic assessment of patients (clinical examinations, interviews, diagnostic tests);
- administrative or in-house databases (prospective recording of data with no pre-specified purpose); and
- extraction from medical records; and
- organizational records or policy documents (e.g. for organizational or public health outcomes).

4.6.2 Differential and non-differential measurement error

Misclassification or measurement error of outcomes may be non-differential or differential. **Non-differential measurement error** is unrelated to the intervention received. It can be systematic (for example when measurement of blood pressure is consistently 5 units too high in every participant) – in which case it will not affect precision or cause bias; or it can be random (for example when measurement of blood pressure is sometimes too high and sometimes too low in a manner that does not depend on the intervention or the outcome) – in which case it will affect precision without causing bias.

Differential measurement error is measurement error related intervention status. It will bias the intervention-outcome relationship. This is often referred to as **detection bias**. Examples of situations in which detection bias can arise are (i) if outcome assessors are aware of intervention status (particularly when the outcome is subjective); (ii) different methods (or intensities of observation) are used to assess outcomes in the different intervention groups; and (iii) measurement errors are related to intervention status (or to a confounder of the intervention-outcome relationship).

Blinding of outcome assessors aims to prevent systematic differences in measurements between intervention groups. However, blinding is frequently not possible or not performed for practical reasons. It is also much less frequent in NRSI than in randomized trials.

The signalling questions include consideration of the comparability of data collection methods and of whether measurement errors may be related to intervention status. If data collection methods are very well standardized, the risk of bias may be lower. It is important also to consider the intensity of investigation across intervention groups. For example, in a study evaluating whether corticosteroid drugs are associated with an increased risk of type 2 diabetes mellitus, diagnosis of type 2 diabetes mellitus may be based on regular blood tests. If patients taking corticosteroids have more frequent blood tests than non-users of corticosteroids (possibly because of their underlying disease), then diabetes is more likely to be detected, introducing a bias against corticosteroids. Another example of detection bias despite standardized assessment of a reasonably objective outcome measures relates to the use of a regular size cuff for measuring blood pressure, which may overestimate the true blood pressure in obese patients. If intervention is also related to body mass then the measurement error will introduce bias, and this bias may be present even if outcomes are measured blind to intervention status.

4.6.3 Risk of bias assessment for bias in measurement of outcomes

The signalling questions and risk of bias assessments are given inBox 4 Box 9 and Table 10.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 9: The ROBINS-I tool (Stage 2, part 8): Risk of bias in measurement of outcomes

Signalling questions	Elaboration	Response options
6.1 Could the outcome measure have been influenced by knowledge of the intervention received?	Some outcome measures involve negligible assessor judgment, e.g. all-cause mortality or non-repeatable automated laboratory assessments. Risk of bias due to measurement of these outcomes would be expected to be low.	Y / PY / PN / N / NI
6.2 Were outcome assessors aware of the intervention received by study participants?	If outcome assessors were blinded to intervention status, the answer to this question would be 'No'. In other situations, outcome assessors may be unaware of the interventions being received by participants despite there being no active blinding by the study investigators; the answer this question would then also be 'No'. In studies where participants report their outcomes themselves, for example in a questionnaire, the outcome assessor is the study participant. In an observational study, the answer to this question will usually be 'Yes' when the participants report their outcomes themselves.	Y / PY / PN / N / NI
6.3 Were the methods of outcome assessment comparable across intervention groups?	Comparable assessment methods (i.e. data collection) would involve the same outcome detection methods and thresholds, same time point, same definition, and same measurements.	Y / PY / PN / N / NI
6.4 Were any systematic errors in measurement of the outcome related to intervention received?	This question refers to differential misclassification of outcomes. Systematic errors in measuring the outcome, if present, could cause bias if they are related to intervention or to a confounder of the intervention-outcome relationship. This will usually be due either to outcome assessors being aware of the intervention received or to non-comparability of outcome assessment methods, but there are examples of differential misclassification arising despite these controls being in place.	Y / PY / PN / N / NI
Risk of bias judgement	See Table 10.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to measurement of outcomes?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 10: Reaching risk of bias judgements for bias in measurement of outcomes

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	(i) The methods of outcome assessment were comparable across intervention groups; *and* (ii) The outcome measure was unlikely to be influenced by knowledge of the intervention received by study participants (i.e. is objective) or the outcome assessors were unaware of the intervention received by study participants; *and* (iii) Any error in measuring the outcome is unrelated to intervention status.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) The methods of outcome assessment were comparable across intervention groups; *and* (ii) The outcome measure is only minimally influenced by knowledge of the intervention received by study participants; *and* (iii) Any error in measuring the outcome is only minimally related to intervention status.
Serious risk of bias (the study has some important problems)	(i) The methods of outcome assessment were not comparable across intervention groups; *or* (ii) The outcome measure was subjective (i.e. vulnerable to influence by knowledge of the intervention received by study participants); *and* The outcome was assessed by assessors aware of the intervention received by study participants; *or* (iii) Error in measuring the outcome was related to intervention status.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	The methods of outcome assessment were so different that they cannot reasonably be compared across intervention groups.
No information on which to base a judgement about risk of bias for this domain	No information is reported about the methods of outcome assessment.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

4.7 Detailed guidance: Bias in selection of the reported result

4.7.1 Introduction

In this document we define: an **outcome domain** as a true state or endpoint of interest, irrespective of how it is measured (e.g. presence or severity of depression), an **outcome measurement** as a specific measurement made on the study participants (e.g. measurement of depression using the Hamilton rating scale 6 weeks after initiation of treatment) and an **outcome analysis** as a specific result obtained by analysing one or more outcome measurements (e.g. the difference in mean change in Hamilton rating scale scores from baseline to 6 weeks between intervention and control groups).

4.7.2 Exclusion of outcome non-reporting bias from the risk of bias tool

Selective reporting within clinical trials has to date mainly been described with respect to the failure to report, or partial reporting of, outcome domains that were measured and analysed (Kirkham et al, 2010). Outcome reporting bias (ORB) arises when the outcome domain is not reported or partially reported based on the direction, magnitude or statistical significance of its association with intervention group. The presence of such bias in one or more of the studies included in a systematic review puts the treatment effect estimate reported by the systematic review at risk of bias (usually in the direction of exaggeration of the magnitude of effect).

The proposed new structure of the RoB tool considers this kind of selective outcome reporting as analogous to publication bias. Therefore, it is proposed to appraise this kind of selective outcome reporting using a different mechanism (e.g. as part a GRADE assessment in the Summary of Findings), not as part of the RoB tool. **This is a notable departure from the current Cochrane RoB tool for randomized trials**. We therefore do not include signalling questions for selective non-reporting (or insufficient reporting) of outcome domains in this document. We recommend the Kirkham et al (2010) framework for considering this kind of selective outcome reporting.

4.7.3 Selective reporting of a result contributing to the synthesis

We consider here the selective reporting of **fully reported results**, that is results that are sufficiently reported to allow the estimate to be included in a meta-analysis (or other synthesis). This domain combines (i) **selective reporting of a particular outcome measurement** from multiple measurements assessed within an outcome domain; (ii) **selective reporting of a particular analysis** from multiple analyses of a specific outcome measurement; and (iii) **selective reporting of a subset of the participants**. These types of selective reporting put effect estimates from individual primary studies at risk of bias in the same way as other bias domains considered in the ROBINS-I tool. Selective reporting will lead to bias if it is based on the direction, magnitude or statistical significance of intervention effect estimates.

Selective outcome reporting occurs when the effect estimate for an outcome measurement was selected from among analyses of multiple outcome measurements for the outcome domain. Examples include: use of multiple measurement instruments (e.g. pain scales) and reporting only the most favourable result; reporting only the most favourable subscale (or a subset of subscales) for an instrument when measurements for other subscales were available; reporting only one or a subset of time points for which the outcome was measured.

Selective analysis reporting occurs when results are selected from intervention effects estimated in multiple ways: e.g. carrying out analyses of both change scores and post-intervention scores adjusted for baseline; multiple analyses of a particular measurement with and without transformation; multiple analyses of a particular measurement with and without adjustment for potential confounders (or with adjustment for different sets of potential confounders); multiple analyses of a particular measurement with and without, or with different, methods to take account of missing data; a continuously scaled outcome converted to categorical data with different cut-points; multiple composite outcomes analysed for one outcome domain, but results were reported only for one (or a subset) of the composite outcomes. (Reporting an effect estimate for an unusual composite outcome might be evidence of such selective reporting.)

Selection of a subgroup from a larger cohort: The cohort for analysis may have been selected from a larger cohort for which data were available on the basis of a more interesting finding. Subgroups defined in unusual ways (e.g. an unusual classification of subgroups by dose or dose frequency) may provide evidence of such selective reporting.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Selective reporting can arise for both harms and benefits, although the motivations (and direction of bias) underlying selective reporting of effect estimates for harms and benefits may differ. Selective reporting typically arises from a desire for findings to be newsworthy, or sufficiently noteworthy to merit publication, and this could be the case if previous evidence (or a prior hypothesis) is either supported or contradicted.

These types of selective reporting apply to all cohort study designs, irrespective of whether they involve clustering.

Selective reporting is more likely to arise in studies which have exploratory objectives because, by their nature, such studies often involve inspecting many associations between multiple interventions or multiple outcomes. However, an exploratory study that fully reported all associations investigated would not be at risk of selective reporting: it is selective reporting that it is the problem, not the exploratory nature of the objective *per se*.

4.7.4 Evidence of selective reporting

Papers can provide evidence of selective reporting in many ways, too numerous to catalogue. Congruence between outcome measurements and analyses specified in a protocol or statistical analysis plan, before analyses were carried out, is required in order to assign low risk of bias.

Indirect evidence that selective reporting may not be a serious problem can be gleaned from: consistency (not as strong a requirement as congruence) between the reported outcome measurements and analyses and an a priori plan, or clearly defined outcome measurements and analyses that are internally consistent across Methods and Results in the paper, and externally consistent with other papers reporting the study. To assign moderate risk of bias there should also be no indication of selection of the reported analysis from among multiple analyses and no indication of selection of the cohort or subgroups for analysis and reporting on the basis of the results.

Inconsistency (internally or externally) in outcome measurements, analyses or analysis cohorts (e.g. a large difference between the size of cohort of eligible participants and the size of the cohort analysed) should indicate a serious risk of selective reporting, especially if all reported results are statistically significant. Some circumstances increase the risk of selective reporting from among multiple analyses, e.g. substantial imbalance in prognostic variables at baseline, without describing a strategy to minimize this risk (e.g. criteria for including covariates in a multiple regression model).

Direct proof or strong suspicion of selective reporting (indicative of *critical risk of bias*, see below) can sometimes be found in the text of a paper. Examples of the kinds of statements that should cause alarm include: (a) "the results for outcome X [relevant to the systematic review outcome domain D] were more favourable than for outcome Y [also relevant to the same systematic review outcome domain D]"; (b) "various cut-off criteria for dichotomizing/classifying a continuous variable were 'tried out'"; (c) "change scores were also analysed but not reported because the effect was not significant". The specific text provoking a judgement of critical bias must be recorded in the free text box.

4.7.5 Risk of bias assessment for bias in selection of the reported result

The signalling questions and risk of bias assessments are given in Box 10 and Table 11.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Box 10: The ROBINS-I tool (Stage 2, part 9): Risk of bias in selection of the reported result

Signalling questions	Elaboration	Response options
Is the reported effect estimate likely to be selected, on the basis of the results, from...		
7.1. ... multiple outcome *measurements* within the outcome domain?	For a specified outcome domain, it is possible to generate multiple effect estimates for different measurements. If multiple measurements were made, but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
7.2 ... multiple *analyses* of the intervention-outcome relationship?	Because of the limitations of using data from non-randomized studies for analyses of effectiveness (need to control confounding, substantial missing data, etc), analysts may implement different analytic methods to address these limitations. Examples include unadjusted and adjusted models; use of final value vs change from baseline vs analysis of covariance; different transformations of variables; a continuously scaled outcome converted to categorical data with different cut-points; different sets of covariates used for adjustment; and different analytic strategies for dealing with missing data. Application of such methods generates multiple estimates of the effect of the intervention versus the comparator on the outcome. If the analyst does not pre-specify the methods to be applied, and multiple estimates are generated but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
7.3 ... different *subgroups*?	Particularly with large cohorts often available from routine data sources, it is possible to generate multiple effect estimates for different subgroups or simply to omit varying proportions of the original cohort. If multiple estimates are generated but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
Risk of bias judgement	See Table 11.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to selection of the reported result?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table 11: Reaching risk of bias judgements for bias in selection of the reported result

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	There is clear evidence (usually through examination of a pre-registered protocol or statistical analysis plan) that all reported results correspond to all intended outcomes, analyses and sub-cohorts.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) The outcome measurements and analyses are consistent with an *a priori* plan; or are clearly defined and both internally and externally consistent; *and* (ii) There is no indication of selection of the reported analysis from among multiple analyses; *and* (iii) There is no indication of selection of the cohort or subgroups for analysis and reporting on the basis of the results.
Serious risk of bias (the study has some important problems)	(i) Outcomes are defined in different ways in the methods and results sections, or in different publications of the study; *or* (ii) There is a high risk of selective reporting from among multiple analyses; *or* (iii) The cohort or subgroup is selected from a larger study for analysis and appears to be reported on the basis of the results.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(i) There is evidence or strong suspicion of selective reporting of results; *and* (ii) The unreported results are likely to be substantially different from the reported results.
No information on which to base a judgement about risk of bias for this domain.	There is too little information to make a judgement (for example if only an abstract is available for the study).

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

5 References

Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine* 2006; **119**: 166.e7-166.e16.

Cook TD, Campbell DT, Peracchio L. Quasi experimentation. In: Dunnette MD, Hough LM (Eds.), *Handbook of Industrial and Organizational Psychology* (pp. 491-567). Palo Alto, CA, US: Consulting Psychologists Press, 1990.

Davey Smith G, Phillips AN, Neaton JD. Smoking as "independent" risk factor for suicide: illustration of an artifact from observational epidemiology? *Lancet* 1992; **340**: 709-12.

Feinstein AR. An additional basic science for clinical medicine: II. The limitations of randomized trials. *Annals of Internal Medicine* 1983;99:544-50.

Hernán MA. With great data comes great responsibility. Publishing comparative effectiveness research in Epidemiology [editorial]. *Epidemiology* 2011; **22**: 290-291.

Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15(5):615-25.

Hernán MA, Hernandez-Diaz S, Werler MM, et al, Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155(2):176-84.

Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JAC, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011; **343**: d5928.

Institute of Medicine. *Ethical and Scientific Issues in Studying the Safety of Approved Drugs*. Washington, DC: The National Academies Press, 2012.

Jackson LA, Jackson ML, Nelson JC, Neuzil KM, Weiss NS. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International Journal of Epidemiology* 2006; **35**: 337–344.

Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010; **340**: c365.

Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010; **21**: 383-8.

Magid DJ, Shetterly SM, Margolis KL, Tavel HM, O'Connor PJ, Selby JV, Ho PM. Comparative effectiveness of angiotensin-converting enzyme inhibitors versus beta-blocker as second-line therapy for hypertension. *Circulation: Cardiovascular Quality and Outcomes* 2010; **3**: 453-458.

Majeed AW, Troy G, Nicholl JP, et al, Randomized, prospective, single-blind comparison of laparascopic versus small-incision cholesystectomy. *Lancet* 1996; 347: 989-94.

McMahon AJ, Russell IT, Baxter JN, et al, Laparascopic versus minilaparotomy cholesystectomy: a randomised trial. *Lancet* 1994; **343**: 135-8.

Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology* 2003;158(9):915-20.

Schünemann HJ, Tugwell P, Reeves BC, et al, Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; **4**: 49-62.

Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 3rd ed. New York: Oxford University Press, 2003.

Suissa S, Spitzer WO, Rainville B, et al, Recurrent use of newer oral contraceptives and the risk of venous thromboembolism. *Human Reproduction* 2000;15(4):817-21.

Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 2008;167(4):492-9.

Rothman KJ and Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott Williams & Wilkins 1998.

Velie EM, Shaw GM. Impact of prenatal diagnosis and elective termination on prevalence and risk estimates of neural tube defects in California, 1989-1991. *American Journal of Epidemiology* 1996;144(5):473-9.

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011; **155**: 529-36.

© 2016 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.