ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions

Jonathan AC Sterne, Miguel A Hernán, Barnaby C Reeves, Jelena Savović, Nancy D Berkman, Meera Viswanathan, David Henry, Douglas G Altman, Mohammed T Ansari, Isabelle Boutron, James R Carpenter, An-Wen Chan, Rachel Churchill, Jonathan J Deeks, Asbjørn Hróbjartsson, Jamie Kirkham, Peter Jüni, Yoon K Loke, Therese D Pigott, Craig R Ramsay, Deborah Regidor, Hannah R Rothstein, Lakhbir Sandhu, Pasqualina L Santaguida, Holger J Schünemann, Beverly Shea, Ian Shrier, Peter Tugwell, Lucy Turner, Jeffrey C Valentine, Hugh Waddington, Elizabeth Waters†, George A Wells, Penny F Whiting, Julian PT Higgins

Jonathan AC Sterne, professor, School of Social and Community Medicine, University of Bristol, Bristol, BS8 2PS, UK.

Miguel A Hernán, professor, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; and Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; and Harvard-Massachusetts Institute of Technology Division of Health Sciences and Technology, Boston, Massachusetts, USA.

Barnaby C Reeves, professorial research fellow, School of Clinical Sciences, University of Bristol, Bristol, BS2 8HW, UK.

Jelena Savović, research fellow, School of Social and Community Medicine, University of Bristol, Bristol, BS8 2PS, UK; and The National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West (NIHR CLAHRC West) at University Hospitals Bristol NHS Foundation Trust, Bristol, BS1 2NT, UK.

Nancy D Berkman, senior health policy research analyst, Program on Health Care Quality and Outcomes, Division of Health Services and Social Policy Research, RTI International, Research Triangle Park, NC 27709, USA.

Meera Viswanathan, director, RTI-UNC Evidence-based Practice Center, RTI International, 3040 Cornwallis Rd, Research Triangle Park, NC 27709, USA.

David Henry, professor, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada.

Douglas G Altman, professor, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Windmill Road, Oxford, OX3 7LD, UK.

Mohammed T Ansari, adjunct professor, School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8M5, Canada.

Isabelle Boutron, professor, METHODS Team, Centre of Epidemiology and Statistics Sorbonne Paris Cité Research, INSERM UMR 1153, University Paris Descartes, Paris, France.

James Carpenter, professor, Department of Medical Statistics, London School of Hygiene and Tropical Medicine and MRC Clinical Trials Unit at UCL, Kingsway, London, UK.

An-Wen Chan, Phelan Scientist, Women's College Research Institute, Dept of Medicine, University of Toronto, Canada.

Rachel Churchill, professor, Centre for Reviews and Dissemination, University of York, York, YO10 5DD, UK.

Jonathan J Deeks, professor, Institute of Applied Health Research, University of Birmingham, Birmingham, B15 2TT, UK.

Asbjørn Hróbjartsson, professor, Center for Evidence-Based Medicine, University of Southern Denmark & Odense University Hospital, 5000 Odense C, Denmark.

Jamie Kirkham, lecturer, Department of Biostatistics, University of Liverpool, Liverpool, L69 3GL, UK.

Peter Jüni, professor, Applied Health Research Centre (AHRC), Li Ka Shing Knowledge Institute of St. Michael’s Hospital, and Department of Medicine, University of Toronto, Toronto, Ontario, Canada.

Yoon Loke, professor, Norwich Medical School, University of East Anglia, Norwich, NR4 7TJ, UK.

Terri Pigott, professor, School of Education, Loyola University Chicago, Chicago, IL 60611, USA.

Craig Ramsay, professor, Health Services Research Unit, University of Aberdeen, Aberdeen, AB25 2ZD, UK.

Deborah Regidor, senior consultant, Evidence Services, Kaiser Permanente, Care Management Institute, Oakland, CA 94612, USA.

Hannah R Rothstein, professor, Department of Management, Zicklin School of Business, Baruch College—CUNY, New York, NY 10010, USA.

Lakhbir Sandhu, resident, Division of General Surgery, University of Toronto, Toronto, Canada.

Pasqualina L Santaguida, assistant professor, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, L8S 4K1, Canada.

Holger J Schünemann, professor, Departments of Clinical Epidemiology and Biostatistics and of Medicine, Cochrane Applicability and Recommendations Methods (GRADEing) Group, MacGRADE center, Ontario, L8N 4K1, Canada.

Beverly Shea, Ottawa Hospital Research Institute, Center for Practice Changing Research and School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8M5, Canada.

Ian Shrier, investigator, Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, Quebec, Canada.

Peter Tugwell, professor, Department of Medicine and School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada.

Lucy Turner, senior research associate, Ottawa Hospital Research Institute, Ottawa, ON, Canada.

Jeffrey C Valentine, associate professor, University of Louisville, Louisville, KY 40292, USA.

Hugh Waddington, lecturer, International Initiative for Impact Evaluation, London School of Hygiene and Tropical Medicine, and London International Development Centre, London, UK.

Elizabeth Waters, professor†, The Jack Brockhoff Child Health & Wellbeing Program, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC 3010, Australia.

George Wells, professor, School of Epidemiology, Public Health and Preventive Medicine and Director, Cardiovascular Research Methods Centre, University of Ottawa Heart Institute, Ottawa, Ontario, K1Y 4W7, Canada.

Penny F Whiting, senior research fellow, School of Social and Community Medicine, University of Bristol, Bristol, BS8 2PS, UK; and The National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West (NIHR CLAHRC West) at University Hospitals Bristol NHS Foundation Trust, Bristol, BS1 2NT, UK.

Julian PT Higgins, professor, School of Social and Community Medicine, University of Bristol, Bristol, BS8 2PS, UK.

*Correspondence to*: Prof Jonathan AC Sterne, School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK. Email: jonathan.sterne@bristol.ac.uk.

†Deceased 22 September 2015.

# Standfirst

Non-randomized studies of the effects of interventions are critical to many areas of health care evaluation, but their results may be biased. It is therefore important to understand and appraise their strengths and weaknesses. We developed ROBINS-I (“Risk Of Bias In Non-randomized Studies - of Interventions”), a new tool for evaluating risk of bias in estimates of the comparative effectiveness (harm or benefit) of interventions from studies that did not use randomization to allocate units (individuals or clusters of individuals) to comparison groups. The tool will be particularly useful to those undertaking systematic reviews that include non-randomized studies.

# Summary Box

* Non-randomized studies of the effects of interventions are critical to many areas of health care evaluation, but are subject to confounding and a range of other potential biases.
* We developed, piloted and refined a new tool, ROBINS-I, to assess Risk Of Bias In Non-randomized Studies - of Interventions”.
* The tool views each study as an attempt to emulate (mimic) a hypothetical pragmatic randomized trial, and covers seven distinct domains through which bias might be introduced.
* We use “signalling questions” to help users of ROBINS-I to judge risk of bias within each domain.
* The judgements within each domain carry forward to an overall risk of bias judgement across bias domains for the outcome being assessed

# Introduction

Non-randomized studies of the effects of interventions (NRSI) are critical to many areas of health care evaluation. Designs of NRSI that can be used to evaluate the effects of interventions include observational studies such as cohort studies and case-control studies in which intervention groups are allocated during the course of usual treatment decisions, and quasi-randomized studies in which the method of allocation falls short of full randomization. Non-randomized studies can provide evidence additional to that available from randomized trials about long-term outcomes, rare events, adverse effects and populations that are typical of real world practice.1 2 The availability of linked databases and compilations of electronic health records has enabled NRSI to be conducted in large representative population cohorts.3 For many types of organizational or public health interventions, NRSI are the main source of evidence about the likely impact of the intervention because randomized trials are difficult or impossible to conduct on an area-wide basis. Therefore systematic reviews addressing the effects of health-related interventions often include NRSI. It is essential that methods are available to evaluate these studies, so that clinical, policy and individual decisions are transparent and based on a full understanding of the strengths and weaknesses of the evidence.

Many tools to assess the methodological quality of observational studies in the context of a systematic review have been proposed.4 5 The Newcastle-Ottawa6 and Downs-Black7 tools have been two of the most popular: both were on a shortlist of methodologically sound tools5 but each includes items relating to external as well as internal validity, and a lack of comprehensive manuals means that instructions may be interpreted differently by different users.5

In the last decade, major developments have been made in tools to assess study validity. A shift in focus from methodological quality to risk of bias has been accompanied by a move from checklists and numeric scores towards domain-based assessments in which different types of bias are considered in turn. Examples are the Cochrane Risk of Bias tool for randomized trials,8 the QUADAS 2 tool for diagnostic test accuracy studies9 and the ROBIS tool for systematic reviews.10 However, there is no satisfactory domain-based assessment tool for NRSI.4

In this paper we describe the development of ROBINS-I (“Risk Of Bias In Non-randomized Studies - of Interventions”), which is concerned with evaluating risk of bias in estimates of the effectiveness or safety (benefit or harm) of an intervention from studies that did not use randomization to allocate interventions.

# Development of a new tool

We developed the tool over a period of three years, largely by expert consensus, and following the seven principles we previously described for assessing risk of bias in clinical trials.8 A core group co-ordinated development of the tool, including recruitment of collaborators, preparation and revision of documents, and administrative support. An initial scoping meeting in October 2011 was followed by a survey of Cochrane Review Groups in March 2012 to gather information about the methods they were using to assess risk of bias in NRSI. A meeting in April 2012 identified the relevant bias domains and established working groups focusing on each of these. We agreed at this stage to use the approach previously adopted in the QUADAS-2 tool, in which answers to “signalling questions” help reviewers judge the risk of bias within each domain.9 We distributed briefing documents to working groups in June 2012, specifying considerations for how signalling questions should be formulated and how answers to these would lead to a risk of bias judgement. We also identified methodological issues that would underpin the new tool: these are described below.

After collation and harmonization by the core group of the working groups’ contributions, all collaborators considered draft signalling questions and agreed on the main features of the new tool during a two-day face-to-face meeting in March 2013. A preliminary version of the tool was piloted within the working groups between September 2013 and March 2014, using NRSI in several review topic areas. Substantial revisions, based on results of the piloting, were agreed by leads of working groups in June 2014. Further piloting then took place, along with a series of telephone interviews with people using the tool for the first time that explored whether they were interpreting the tool and the guidance as intended. We posted version 1.0.0, along with detailed guidance, at [www.riskofbias.info](http://www.riskofbias.info) in September 2014. We explained the tool during a three-day workshop involving members of Cochrane Review Groups in December 2014, and applied it in small groups to six papers reporting NRSI. Further modifications to the tool, particularly regarding wording, were based on feedback from this event and from subsequent training events conducted during 2015.

# Methodological issues in assessing risk of bias in non-randomized studies

## The target trial

Evaluations of risk of bias in the results of NRSI are facilitated by considering each NRSI as an attempt to emulate (mimic) a “target” trial. This is the hypothetical pragmatic randomized trial, conducted on the same participant group and without features putting it at risk of bias, whose results would answer the question addressed by the NRSI.11 12 Such a “target” trial need not be feasible or ethical: for example it could compare individuals who were and were not assigned to start smoking. Description of the target trial for the NRSI being assessed includes details of the population, experimental intervention, comparator and outcomes of interest. Correspondingly, we define bias as a systematic difference between the results of the NRSI and the results expected from the target trial. Such bias is distinct from issues of generalizability (applicability or transportability) to types of individuals who were not included in the study.

## The effect of interest

In the target trial, the effect of interest will typically be that of either:

1. assignment to intervention at baseline (start of follow up), regardless of the extent to which the intervention was received during the follow-up (sometimes referred to as the “intention-to-treat” effect); or
2. starting and adhering to the intervention as indicated in the trial protocol (sometimes referred to as the “per-protocol” effect).

For example, in a trial of cancer screening, our interest might be in the effect of either sending an invitation to attend screening, or of responding to the invitation and undergoing screening.

Analogues of these effects can be defined for NRSI. For example, the intention-to-treat effect in a study comparing aspirin with no aspirin can be approximated by the effect of being prescribed aspirin or (if using dispensing rather than prescription data) the effect of starting aspirin (this corresponds to the intention-to-treat effect in a trial in which participants assigned to an intervention always start that intervention). Alternatively, we might be interested in the effect of starting and adhering to aspirin.

The type of effect of interest influences assessments of risk of bias related to deviations from intervention. When the effect of interest is that of assignment to (or starting) intervention, risk of bias assessments generally need not be concerned with post-baseline deviations from interventions.13 By contrast, unbiased estimation of the effect of starting and adhering to intervention requires consideration of both adherence and differences in additional interventions (“co-interventions”) between intervention groups.

## Domains of bias

We achieved consensus on seven domains through which bias might be introduced into a NRSI (see Table 1 and supplementary information on bmj.com). The first two domains, covering confounding and selection of participants into the study, address issues before the start of the interventions that are to be compared (“baseline”). The third domain addresses classification of the interventions themselves. The other four domains address issues after the start of interventions: biases due to deviations from intended interventions, missing data, measurement of outcomes and selection of the reported result.

For the first three domains, risk of bias assessments for NRSI are mainly distinct from assessments of randomized trials because randomization, if properly implemented, 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 proved a challenge to development of ROBINS-I. The same terms are sometimes used to refer to different types of bias in randomized trials and NRSI literature,13 and different types of bias are often described by a host of different terms: those used within ROBINS-I are shown in the first column of Table 1.

# The risk of bias tool, ROBINS-I

The full ROBINS-I tool is shown in Table 2, Table 3 and Table 4.

## Planning the risk of bias assessment

It is very important that experts in both subject matter and epidemiological methods are included in any team evaluating a NRSI. The risk of bias assessment 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. This will be based on experts’ knowledge of the literature: the team should also address whether conflicts of interest might affect experts’ judgements.

The research question is conceptualized by defining the population, experimental intervention, comparator and outcomes of interest (Table 2 Stage I). The comparator could be “no intervention”, “usual care” or an alternative intervention. It is important to consider in advance the confounding factors and co-interventions that have the potential to lead to bias. Relevant confounding domains are the prognostic factors that predict whether an individual receives one or the other intervention of interest. Relevant co-interventions are those that individuals might receive with or after starting the intervention of interest and that are both related to the intervention received and prognostic for the outcome of interest. Both confounding domains and co-interventions are likely to be identified through the expert knowledge of 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 help in identification of prognostic factors that influence treatment decisions.

## Assessing a specific study

The assessment of each NRSI included in the review involves the following six steps (Table 2 Stage II). Steps 3 to 6 should be repeated for each key outcome of interest:

1. specify the research question through consideration of a target trial;
2. specify the outcome and result being assessed;
3. for the specified result, examine how the confounders and co-interventions were addressed;
4. answer signalling questions for the seven bias domains;
5. formulate risk of bias judgements for each of the seven bias domains, informed by answers to the signalling questions; and
6. formulate an overall judgement on risk of bias for the outcome and result being assessed.

Examination of confounders and co-interventions involves determining whether the important confounders and co-interventions were measured or administered in the study at hand, and whether additional confounders and co-interventions were identified. An online supplement provides a structured approach to assessing the potential for bias due to confounding and co-interventions and includes the full tool with the signalling questions to be addressed within each bias domain.

The signalling questions are broadly factual in nature and aim to facilitate judgements about the risk of bias. The response options are: ‘Yes’; ‘Probably yes’; ‘Probably no’; ‘No’; and ‘No information’. Some questions are only answered if the response to a previous question is ‘Yes’ or ‘Probably yes’ (or ‘No’ or ‘Probably no’). Responses of ‘Yes’ are intended to have similar implications to responses of ‘Probably yes’ (and similarly for ‘No’ and ‘Probably no’), but allow for a distinction between something that is known and something that is likely to be the case. Free text should be used to provide support for each answer, using direct quotations from the text of the study where possible.

Responses to signalling questions provide the basis for domain-level judgements about risk of bias, which then provide the basis for an overall risk of bias judgement for a particular outcome. The use of the word “judgement” to describe this process is important and reflects the need for review authors to consider both the severity of the bias in a particular domain and the relative consequences of bias in different domains.

The categories for risk of bias judgements are ‘Low risk’, ‘Moderate risk’, ‘Serious risk’ and ‘Critical risk’ of bias. Importantly, ‘Low risk’ corresponds to the risk of bias in a high quality randomized trial. Only exceptionally will an NRSI be assessed as at low risk of bias due to confounding. Criteria for reaching risk of bias judgements for the seven domains are provided in Table 3 and Table 4. 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 judgements within each domain carry forward to an overall risk of bias judgement for the outcome being assessed (across bias domains, that is), as summarized in Table 5. The key to applying the tool is to make domain-level judgements about risk of bias that have the same meaning 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 risk of bias for a particular outcome is relatively straightforward. For instance, a ‘Serious risk’ of bias in one domain means the effect estimate from the study is at serious risk of bias or worse, even if the risk of bias is judged to be lower in the other domains.

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 predict 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 as an increase or decrease in the effect estimate and not in relation to the null. 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.

# Discussion

We developed a tool for assessing risk of bias in the results of non-randomized studies of interventions that addresses weaknesses in previously available approaches.4 Our approach builds on recent developments in risk of bias assessment of randomized trials and diagnostic test accuracy studies.8 9 Key features of ROBINS-I include specification of the target trial and effect of interest, use of signalling questions to inform judgements of risk of bias, and assessments within seven bias domains.

The ROBINS-I tool was developed through consensus among a group that included both methodological experts and systematic review authors and editors, and was substantially revised based on extensive piloting and user feedback. It includes a structured approach to assessment of risk of bias due to confounding that starts at the review protocol stage. Use of ROBINS-I requires that review groups include members with substantial methodological expertise and familiarity with modern epidemiological thinking. We tried to make ROBINS-I as accessible and easy to use as possible, given the requirement for comprehensive risk of bias assessments that are applicable to a wide range of study designs and analyses. An illustrative assessment using ROBINS-I can be found at [www.riskofbias.info](http://www.riskofbias.info); detailed guidance and further training materials will also be available.

ROBINS-I separates relatively factual answers to signalling questions from more subjective judgements about risk of bias. We hope that the explicit links between answers to signalling questions and risk of bias judgements will improve reliability of the domain-specific and overall risk of bias assessments14. Nonetheless, we expect that the technical difficulty in making risk of bias judgements will limit reliability. Despite this, ROBINS-I provides a comprehensive and structured approach to assessing non-randomized studies of interventions. It should therefore facilitate debates and improve mutual understanding about the ways in which bias can influence effects estimated in NRSI, and clarify reasons for disagreements about specific risk of bias judgements. Note that the tool focusses specifically on bias and does not address problems related to imprecision of results, for example when statistical analyses fail to account for clustering or matching of participants.

We developed the ROBINS-I tool primarily for use in the context of a systematic review. Broader potential uses include the assessment of funding applications and peer review of journal submissions. Furthermore, ROBINS-I may be used to guide researchers about issues to consider when designing a primary study to evaluate the effect(s) of an intervention.

Figure 1 summarizes the process of assessing risk of bias using the tool in the context of a systematic review of NRSI. To draw conclusions about the extent to which observed intervention effects might be causal, the studies included in the review 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. Syntheses of findings across studies through meta-analysis must consider the risks of bias in the studies available. We recommend against including studies assessed as at ‘Critical risk’ of bias in any meta-analysis, and advocate caution for studies assessed as at ‘Serious risk’ of bias. Subgroup analyses (in which intervention effects are estimated separately according to risk of bias), meta-regression analyses and sensitivity analyses (excluding studies at higher risk of bias) might be considered, either within specific bias domains or overall. Risk of bias assessments might alternatively be used as the basis for deriving adjustments for bias through prior distributions in Bayesian meta-analyses.15 16

The GRADE system for assessing confidence in estimates of the effects of interventions currently assigns a starting rating of ‘Low certainty, confidence or quality’ to non-randomized studies, a downgrading by default of two levels.17 ROBINS-I provides a thorough assessment of risk of bias in relation to a hypothetical randomized trial, and ‘Low risk’ of bias corresponds to the risk of bias in a high quality randomized trial. This opens up the possibility of using the risk of bias assessment, rather than the lack of randomization per se, to determine the degree of downgrading of a study result, and means that results of NRSI and randomized trials could be synthesized if they are assessed to be at similar risks of bias. However in general we advocate analysing these study designs separately and focusing on evidence from NRSI when evidence from trials is not available.

Planned developments of ROBINS-I include further consideration of the extent to which it works for specific types of NRSI, such as self-controlled designs, controlled before-and-after studies, interrupted time series studies, and studies based on regression discontinuity and instrumental variable analyses. We also plan to develop interactive software to facilitate use of ROBINS-I. Furthermore, the discussions that led up to the tool will inform a reconsideration of the tool for randomized trials, particularly in the four post-intervention domains.8

The role of NRSI in informing treatment decisions remains controversial. Because randomized trials are expensive, time consuming and may not reflect real-world experience with healthcare interventions, research funders are enthusiastic about the possible use of observational studies to provide evidence about the comparative effectiveness of different interventions,18 and encourage use of large, routinely collected datasets assembled through data linkage.18 However, fear that evidence from NRSI may be biased, based on misleading results of some NRSI19 20 has led to caution in their use in making judgements about efficacy. There is greater confidence in the capacity of NRSI to quantify uncommon adverse effects of interventions.21 We believe that evidence from NRSI should complement that from randomized trials, for example in providing evidence about effects on rare and adverse outcomes and long-term effects to be balanced against the outcomes more readily addressed in randomized trials.22

# Funding

Development of ROBINS-I was funded by a Methods Innovation Fund grant from Cochrane and Medical Research Council grant MR/M025209/1. Sterne and Higgins are members of the Medical Research Council Integrative Epidemiology Unit at the University of Bristol which is supported by the Medical Research Council and the University of Bristol (grant MC\_UU\_12013/9). This research was partly funded by NIH grant P01 CA134294. Sterne was supported by National Institute for Health Research Senior Investigator award NF-SI-0611-10168. Savović and Whiting were supported by National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West (NIHR CLAHRC West). Reeves was supported by the National Institute for Health Research Bristol Biomedical Research Unit in Cardiovascular Disease. None of the funders had a role in the development of the ROBINS-I tool, although employees of Cochrane contributed to some of the meetings and workshops. The views expressed are those of the authors and not necessarily those of Cochrane, the NHS, the NIHR or the Department of Health.

# Acknowledgements

We are grateful to Professor Jan Vandenbroucke for his contributions to discussions within working groups and during face to face meetings. Professor David Moher and Dr Vivan Welch contributed to the grant application and to initial discussions. We thank all those individuals who contributed to development of ROBINS-I through discussions during workshops and training events, and through their work on piloting.

# Contributions of authors

JACS, BCR, JS, LT, YKL, EW, CRR, PT, GAW and JPTH conceived the project. JACS, JPTH, BCR, JS and LT oversaw the project. JACS, DH, JPTH, IB and BRR led working groups. MAH developed the idea of the target trial and its role in ROBINS-I. NDB and MV undertook cognitive testing of previous drafts of the tool. All authors contributed to development of ROBINS-I and to writing associated guidance. JACS, MAH and JPTH led on drafting the manuscript. All authors reviewed and commented on drafts of the manuscript.

# Competing interests

All authors have completed the ICMJE uniform disclosure form at <http://www.icmje.org/coi_disclosure.pdf> and declare: grants from Cochrane, MRC and NIHR, during the conduct of the study. Dr. Carpenter reports personal fees from Pfizer, grants and non-financial support from GSK and grants from Novartis, outside the submitted work. Dr. Reeves is a co-convenor of the Cochrane Non-Randomized Studies Methods Group. The authors report no other relationships or activities that could appear to have influenced the submitted work.

# Provenance

The authors are epidemiologists, statisticians, systematic reviewers, trialists and health services researchers, many of whom are involved with Cochrane systematic reviews, methods groups and training events. Development of ROBINS-I was informed by relevant methodological literature, previously published tools for assessing methodological quality of non-randomized studies, systematic reviews of such tools and relevant literature, and by the authors’ experience of developing tools to assess risk of bias in randomized trials, diagnostic test accuracy studies and systematic reviews. All authors contributed to development of ROBINS-I and to writing associated guidance. All authors reviewed and commented on drafts of the manuscript. Jonathan Sterne will act as guarantor.

# References

1. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;**312**(7040):1215-8.

2. Feinstein AR. An additional basic science for clinical medicine: II. The limitations of randomized trials. Ann Intern Med 1983;**99**(4):544-50.

3. Strom B. Overview of automated databases in pharmacoepidemiology. In: Strom BL KS, Hennessy S., ed. Pharmacoepidemiology. 5th Edition ed: Wiley, 2012.

4. Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;**36**(3):666-76.

5. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;**7**:27.

6. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2008. http://www.ohri.ca/programs/clinical\_epidemiology/oxford.asp (accessed 1/03/2016).

7. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Commun H 1998;**52**(6):377-84.

8. Higgins JPT, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;**343**:d5928.

9. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;**155**(8):529-36.

10. Whiting P, Savović J, Higgins JPT, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2015;**69**:225-34.

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

12. Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016(published online Mar 18).

13. Mansournia MA, Higgins JPT, Sterne JAC, et al. Biases in randomized trials: a conversation between trialists and epidemiologists. Epidemiology (in press).

14. Hartling L, Hamm MP, Milne A, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013;**66**(9):973-81.

15. Turner RM, Spiegelhalter DJ, Smith GCS, et al. Bias modelling in evidence synthesis. J Roy Stat Soc A 2009;**172**(1):21-47.

16. Wilks DC, Sharp SJ, Ekelund U, et al. Objectively measured physical activity and fat mass in children: a bias-adjusted meta-analysis of prospective studies. PloS One 2011;**6**(2):e17205.

17. Guyatt GH, Oxman AD, Schunemann HJ, et al. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011;**64**(4):380-2.

18. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. Ann Intern Med 2009;**151**(3):203-5.

19. Bjelakovic G, Nikolova D, Gluud LL, et al. Mortality in randomized trials of antioxidant supplements for primary and secondary prevention: systematic review and meta-analysis. JAMA 2007;**297**(8):842-57.

20. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 2008;**19**(6):766-79.

21. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med 2011;**8**(5):e1001026.

22. Schunemann 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. Res Synth Methods 2013;**4**(1):49-62.

Table 1. Bias domains included in ROBINS-I.

|  |  |  |
| --- | --- | --- |
| **Domain** | **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 | 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 | 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 | 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. |
| *Post-intervention* |  | Post-intervention domains for which there is substantial overlap with assessments of randomized trials |
| Bias due to deviations from intended interventions | 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 | 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  | 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 | Selective reporting of results in a way that depends on the findings and prevents the estimate from being included in a meta-analysis (or other synthesis). |

Table 2. The Risk Of Bias In Non-randomized Studies – of Interventions (ROBINS-I) assessment tool

# ROBINS-I tool (Stage I): 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

|  |
| --- |
|  |

# ROBINS-I tool (Stage II): For each study

## 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.

|  |
| --- |
|  |

## 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”.

## 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 |

## Risk of bias assessment

Responses underlined in green 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Bias domain** | **Signalling questions** | **Elaboration** | **Response options** |
| Bias due to confounding | 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 | 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 |
| **If Y/PY to 1.1**: determine whether there is a need to assess time-varying confounding: |  |
| 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**, proceed 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 |
| **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 adjusted 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 adjusted for measured validly and reliably by the variables available in this study? | See 1.5 above. | NA / Y / PY / PN / N / NI |
| **Risk of bias judgement** | See Table 3 | Low / Moderate / Serious / Critical / NI |
| Optional: What is the predicted direction of bias due to confounding? | 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. | Favours experimental / Favours comparator / Unpredictable |
| Bias in selection of participants into the study | 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?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? | 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 / NINA / 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 3 | 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 |
| Bias in classification of interventions  | 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 3 | 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 |
| Bias due to deviations from intended interventions | **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 |
| 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 4 |  |
| 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. |  |
| Bias due to missing data | 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 4 | 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 |
| Bias in measurement of outcomes  | 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 4 | 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 |
| Bias in selection of the reported result | 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 4 | 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 |
| Overall bias | **Risk of bias judgement** | See Table 5 | Low / Moderate / Serious / Critical / NI |
| Optional: What is the overall predicted direction of bias for this outcome? |  | Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable |

| Table 3. Reaching risk of bias judgements in ROBINS-I: pre-intervention and at-intervention domains |
| --- |
| **Judgement** | **Bias due to confounding** | **Bias in selection of participants into the study** | **Bias in classification of interventions** |
| Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain) | No confounding expected. | (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. | (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) 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. | (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. | (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) 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. | (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. | (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); | (i) Confounding inherently not controllable*or*(ii) The use of negative controls strongly suggests unmeasured confounding. | (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. | (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 information on whether confounding might be present. | No information is reported about selection of participants into the study or whether start of follow up and start of intervention coincide. | No definition of the intervention or no explanation of the source of information about intervention status is reported. |

|  |
| --- |
| Table 4. Reaching risk of bias judgements in ROBINS-I: post-intervention domains |
| **Judgement** | **Bias due to deviations from intended intervention** | **Bias due to missing data** | **Bias in measurement of outcomes** | **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) | **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. | (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. | (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. | 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): | **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. | (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. | (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. | (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); | **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. | (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. | (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. | (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); | **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. | (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. | The methods of outcome assessment were so different that they cannot reasonably be compared across intervention groups. | (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. | No information is reported on whether there is deviation from the intended intervention. | No information is reported about missing data or the potential for data to be missing. | No information is reported about the methods of outcome assessment. | There is too little information to make a judgement (for example, if only an abstract is available for the study). |

Table 5. Interpretation of domain-level and overall risk of bias judgements in ROBINS-I

|  |  |  |  |
| --- | --- | --- | --- |
| **Judgement**  | **Within each domain** | **Across domains** | **Criterion** |
| Low risk of bias  | The study is comparable to a well-performed randomized trial with regard to this domain | 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 is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial  | The study provides 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 in this domain | 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 in this domain to provide any useful evidence on the effects of intervention | 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  | No information on which to base a judgement about risk of bias for this 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*). |

Figure 1. Summary of the process of assessing risk of bias in a systematic review of non-randomized studies of interventions (NRSI)



Online supplement

The seven domains of bias addressed in the ROBINS-I assessment tool

## Confounding

(Related terms: Selection bias as it is sometimes used in relation to clinical trials; Allocation bias; Case-mix bias; Channelling bias.)

In contrast to randomized trials, the characteristics of participants in NRSI will typically differ between intervention groups. The assessment of risk of bias arising from uncontrolled confounding is therefore a major component of ROBINS-I. Confounding of intervention effects occurs when one or more prognostic variables (variables that predict the outcome of interest) also predict whether an individual receives one or the other of the interventions of interest.

Baseline confounding, which occurs when one or more prognostic variables predicts the intervention received at start of follow up, is likely to be an issue in most NRSI. 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. Appropriate methods to control for measured confounders include stratification, regression, matching, standardization, g-estimation, and inverse probability weighting. They may control for individual variables or for the estimated propensity score.

ROBINS-I also addresses time-varying confounding.1 This only needs to be considered in studies that partition follow up time for individual participants into time spent in different intervention groups. 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. For example, CD4 cell count measured after start of antiretroviral therapy (a post-baseline prognostic variable) might influence switches between the regimens of interest.2 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.2-3 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.

## Selection bias

(Related terms: Selection bias as usually used in relation to observational studies and sometimes used in relation to clinical trials; Inception bias; Lead-time bias; Immortal time bias.)

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 type of bias is called selection bias2 and is distinct from confounding, although the terms are sometimes confused. As an example, studies of folate supplementation to prevent neural tube defects were biased because they were restricted to live births.4 The bias arises because stillbirths and therapeutic abortions (which were excluded from the sample), are related to both the intervention and the outcome.2 4 Another example is that the apparently increased risk of venous thromboembolism with the newer oral contraceptive progestogens when investigated in NRSI.5 6 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 the so-called “immortal time bias”.5 7

Our use of the term selection bias refers 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).8 For example, restricting the study to individuals free of comorbidities may limit the generalizability of its findings to clinical practice, where comorbidities are common.9 However it does not bias the estimated effect of intervention, compared to a target trial in which participants were free of comorbidities.

## Bias in measurement of interventions

(Related terms: Misclassification bias; Information bias; Recall bias; Measurement bias; Observer bias.)

Misclassification of assignment to intervention is seldom a problem in randomized trials, but misclassification of intervention received may occur in NRSI. For example, the absence of a record of vaccination does not guarantee that no vaccination was administered. 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. It is therefore important that, wherever possible, interventions are defined and categorized without knowledge of subsequent outcomes. A well-known example is recall bias in a case-control study, whereby knowledge of case-control status may affect recall of previous intervention. Differential misclassification can also occur in cohort studies, if information (or availability of information) on intervention status is influenced by outcomes. For example, in a cohort study of elderly people in which the outcome is dementia, some participants may have had mild cognitive impairment at inception. The recall of previous interventions at study inception may be affected in these participants.

## Bias due to deviations from intended interventions

(Related terms: Performance bias; Time-varying confounding)

This domain (sometimes known as “performance bias”)10 relates to biases that arise when there are systematic differences between the care provided to experimental intervention and comparator groups, beyond the assigned interventions. Bias may occur when these differences arise because of knowledge of the intervention applied and the expectation of finding a difference between experimental intervention and comparator consistent with the hypothesis being tested in the study. Deviations from intended interventions may arise because an intervention was not implemented successfully (for example if laboratory errors meant that the drugs administered did not have the intended formulation), because participants did not adhere to interventions, or because important co-interventions were not balanced between intervention groups.

Whether non-adherence or co-interventions lead to bias depends on the effect of interest: they will not lead to bias in the effect of assignment to intervention. By contrast, bias will arise if we are interested in the effect of starting and adhering to intervention, which is often the case when investigating adverse effects of drugs. 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.

Note that some deviations from intervention happen during usual clinical care (for example, cessation of a drug intervention because of acute toxicity), and can be considered to be part of the intended intervention.

## Bias due to missing data

(Related terms: Attrition bias; Selection bias as it is sometimes used in relation to observational studies)

Reasons for missing data include attrition (loss to follow up), missed appointments, incomplete data collection and participants being excluded from analysis by primary investigators. In NRSI, data may be missing for interventions received, confounders or outcome measurements. Differences between intervention groups in the extent of and reasons for missing data are key.11 If the proportion of missing data is low and the reasons for missing data are similar across intervention groups, then the risk of bias is likely to be low. As the proportion of missing data rises, differences in treatment response between available and missing participants may increase the potential for bias.

## Bias in measurement of outcomes

(Related terms: Detection bias; Recall bias; Information bias; Misclassification bias; Observer bias; Measurement bias)

Bias may be introduced if outcomes are misclassified or measured with error.12 Misclassification or measurement error is non-differential if it is unrelated to the intervention received.13 Random errors that do not depend on the intervention or the outcome (non-differential measurement errors) will not necessarily cause bias. Differential measurement errors (those related to intervention status) will bias the estimated effect of intervention-outcome relationship.11 This is sometimes referred to as detection bias.10 Detection bias can arise when outcome assessors are aware of intervention status, particularly when the outcome is subjective, if different methods (or intensities of observation) are used to assess outcomes in different intervention groups, or if measurement errors are related to intervention status or effects (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, and is much less common in NRSI than in randomized trials.

## Bias in selection of the reported result

(Related terms: Outcome reporting bias; Analysis reporting bias)

Selective reporting will lead to bias if it is based on the direction, magnitude or statistical significance of intervention effect estimates.14 This last bias domain includes three types of selective reporting. Selective outcome reporting occurs when an effect estimate for a particular outcome measurement is selected from among multiple measurements, for example a measurement made at one of a number of time points or based on one of multiple pain scales. Selective analysis reporting occurs when the reported results are selected from intervention effects estimated in multiple ways, such as analyses of both change scores and post-intervention scores adjusted for baseline, or multiple analyses with adjustment for different sets of potential confounders. Finally, there may be selective reporting of a subgroup of participants, selected from a larger cohort, for which results are reported on the basis of a more interesting finding.

## References for online supplement

1. Hernán MA, Robins JM. Longitudinal causal inference. In: Wright JD, ed. International Encyclopedia of the Social & Behavioral Sciences. Oxford, UK: Elsevier, 2015:340-44.

2. 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.

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

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

5. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 2003;**158**(9):915-20.

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

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

8. Schunemann 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. Res Synth Methods 2013;**4**(1):49-62.

9. Hernán MA. Levels of selection and selection bias. Proposal of a vote of thanks to Keiding and Louis. J Roy Stat Soc A (in press).

10. Higgins JPT, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;**343**:d5928.

11. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;**338**:b2393.

12. Hernán MA, Cole SR. Invited Commentary: Causal diagrams and measurement bias. Am J Epidemiol 2009;**170**(8):959-62; discussion 63-4.

13. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;**359**(9302):248-52.

14. Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PloS One 2008;**3**(8):e3081.