A systematic review: Tools for assessing methodological quality of human observational studies

Zhicheng Wang[1,2], Kyla Taylor[3], Margaret Allman-Farinelli[1,4], Bruce Armstrong[5], Lisa Askie[1,6], Davina Ghersi[6,7], Joanne E. McKenzie[8], Susan L. Norris[9], Matthew J. Page[8], Andrew Rooney[3], Tracey Woodruff[10], Lisa A. Bero*[1,2]

[1] Charles Perkins Centre, The University of Sydney, Sydney, Australia. [2] Sydney Pharmacy School, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia. [3] Office of Health Assessment and Translation, National Toxicology Program, National Institute of Environmental Health Sciences, NIH, DHHS, North Carolina, USA. [4] School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia. [5] School of Population and Global Health, The University of Western Australia, Perth, Australia. [6] NHMRC Clinical Trials Centre, The University of Sydney, Sydney, Australia [7]National Health and Medical Research Council, Canberra, Australia. [8] School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia. [9]World Health Organization, Geneva, Switzerland. [10]Department of Obstetrics, Gynecology and Reproductive Sciences & the Institute for Health Policy Studies, University of California, San Francisco

* Corresponding author Lisa.bero@sydney.edu.au

D17, The Hub, 6th floor, Charles Perkins Centre, The University of Sydney, NSW, 2006

Running title: Tools for assessing methodological quality of human observational studies

Keywords: human observational studies, risk of bias, study quality, environmental exposure, environmental health, epidemiology

Preamble:

As part of a Guideline Developers Network initiative, the National Health and Medical Research Council (NHMRC) and Professor Lisa Bero (University of Sydney) held a one-day workshop on 17 April 2018 for organisations involved in developing guidance in public and environmental health. The workshop was held at the Charles Perkins Centre, University of Sydney and focused on improving methods in guideline development and strategies for risk communication. Objectives of the workshop included in identifying common challenges experienced by developers and possible ways to overcome them. Participants from the workshop were presented with the findings of a systematic review of tools for assessing risk of bias in human observational studies of exposures. This review identified key domains that are included in tools and participants discussed these domains. Participants agreed that these key domains should be considered when selecting a risk of bias tool.

Please cite as:

Wang Z, Taylor K, Allman-Farinelli M, Armstrong B, Askie L, Ghersi D, McKenzie J, Norris S, Page M, Rooney A, Woodruff T, Bero L. A systematic review: Tools for assessing methodological quality of human observational studies. NHMRC. 2019. Available at https://nhmrc.gov.au/guidelinesforguidelines/develop/assessing-risk-bias

Acknowledgements:
We would like to thank Nicholas Chartres (NC) for his help in the data extraction process

Grant information:
This study was supported by the University of Sydney as a Sydney Policy Lab Project. The funder had no role in the design, conduct or publication of the review.

Abbreviations: COI – conflict of interest, N/A – not applicable, RCT - randomized controlled trial, AHRQ - Agency for Healthcare Research and Quality

BOX 1: Definitions

Quality: study characteristics, including but not limited to, whether it was peer-reviewed, characteristics pertaining to internal or external validity, completeness of study reporting, best research practices (e.g. ethical approval) and / or risk of bias.

Risk of bias: study characteristics related to systematic error; to assess the risk of a study over- or underestimating outcome effects due to factors in the design, conduct, or analysis of the study (1).

Internal validity: the validity of the inferences drawn as they pertain to the members of the source population.

External validity: the validity of the inferences as they pertain to people outside the source population.

Reporting: study characteristics describing how clearly and completely the details and procedures of a study are described (e.g. “was the objective of the study clearly stated”).

Precision: study characteristics related to random error (e.g. sample size).

Outline of section headers

Abstract
Introduction
Methods
Results
Discussion
Limitations
Conclusions
References

Abstract

Background: Observational studies are the primary human study type for evaluating potential harm from environmental exposures, and therefore critical for developing public health guidance. Although there are tools for assessing methodological quality in observational studies, there is no consensus on optimal tools or key elements to include when assessing quality. This study aims to help decision makers in selecting tools to assess risk of bias in observational studies of exposures.

Objective: To identify, describe, categorize elements into domains, and evaluate published tools designed to assess methodological quality or risk of bias in observational human studies assessing the effects of exposures.

Study appraisal and synthesis: Data was extracted on the characteristics, development and testing of each tool. The categories of items were classified into 9 domains informed by the literature.

Results: We identified 62 tools with 17 categories of similar or overlapping items.

Conclusions: Our review highlights the need for a common tool for assessing risk of bias in human observational studies of exposures. Absent that common tool, this review provides a clear cross-walk across available tools and identifies four main take home messages: (1) the tool should have clear definitions for each item and be transparent regarding the empirical or theoretical basis for each domain, (2) tools should include questions addressing 9 domains: Selection, Exposure, Outcome assessment, Confounding, Loss to follow-up, Analysis, Selective reporting, Conflicts of interest and Other, (3) the ratings for each domain should be reported, rather than an overall score, (4) the tool should be rigorously and independently tested for usability and reliability.

Systematic Review Registration: PROSPERO: CRD42018094173

Introduction

Systematic reviews are increasingly used to inform health policy and practice. It is important to evaluate methodological quality, generally in terms of risk of bias in the individual studies that are included in systematic reviews and this is an essential step in the systematic review process (1). One widely used tool for evaluating the risk of bias in randomized controlled trials (RCTs) is the Cochrane Risk of Bias tool (2). Another tool exists for evaluating risk of bias in observational studies of interventions (3). However, there has been less systematic development of, and little consensus on, optimal tools for assessing the risk of bias in observational studies of exposures that are not controlled by the investigator. For example, neither the National Health and Medical Research Council and World Health Organization recommend a specific tool for assessing risk of bias in human observational studies of exposures (4, 5). Assessing risk of bias is a challenge for systematic reviews examining the health effects of exposures not controlled by the investigators. These types of studies address public health questions such as whether exposure to a chemical is associated with harmful developmental outcomes, whether a dietary pattern is associated with an adverse cardiovascular health effect, or whether exposure to air pollution is associated with asthma.

Currently, when guideline developers such as Australia's National Health and Medical Research Council (NHMRC) or the World Health Organization conduct a systematic review to evaluate the human evidence for environmental or public health guidelines, the relevant human studies are mostly observational in design (e.g. cohort, case control, or cross-sectional studies). Observational designs are used because real-world exposures, such as diet, water

contaminants, or environmental health hazards, are not assigned by an investigator. In observational studies, the participants' exposure is a consequence of personal choice or life circumstances and, for multiple practical or ethical reasons, not at all under the investigator's control. Thus, some of the questions in tools developed for interventions studies, such as RCTs, are not relevant for assessing studies of exposures.

Although there are tools that purport to assess risk of bias of bias in observational studies of exposures, they commonly also evaluate other study features (e.g. reporting, precision). Box 1 lists definitions used in this review. Risk of bias is defined as a measure of whether features of the design, conduct or analysis of a study may cause systematic error in the study's results (2). Many tools for observational studies are described as assessing "quality" which may include items about risk of bias, reporting, internal or external validity, and best research practices (i.e., ethics approval). Some of the tools include domains that are not related to risk of bias or are missing key domains of assessment and, therefore, could result in misjudging risk of bias in studies (6). For the purposes of this review, we are interested in tools described as assessing methodological quality or risk of bias. Because some tools are broadly described as assessing "quality" when they assess risk of bias, other features such as reporting or precision, or combinations of these features, a search for these tools must be broad in scope.

Existing tools for assessing the quality of human observational studies examining effects of exposures differ in their content, reliability and usability (7-9). Some of the tools have been developed to assess specific study topics (e.g. occupational exposure, nutrition) or study designs (e.g. case-control, cohort, cross-sectional). It is not unusual for systematic reviewers to develop *ad hoc* tools that meet the need of a specific review (e.g. (10), NutriGrade (11)). The lack of consensus on the essential items in a risk of bias tool and the large variety of such tools leads to confusion among end-users. Uncertainties about which tool to use or how to use a tool

could produce inconsistency across systematic review findings, delay evaluation of observational studies and, ultimately, inhibit their use in decision making.
This review aims to identify, describe and evaluate published tools designed to assess methodological quality or risk of bias in observational human studies of exposure effects.

Exposure in this review is defined as any exposure that is not controlled by the investigator and could include exposure to chemical, biological or physical stressors (e.g. air pollution or physical activity). This review expands and updates published reviews of tools for assessing observational studies (6-9). We describe the development and evaluation of the tools, the items included in each tool, and group the items into common risk of bias domains.

We undertook this research in partnership with the National Health and Medical Research Council (NHMRC), Research Translation Department, World Health Organization, Guidelines Secretariat, and Office of Health Assessment and Translation, National Toxicology Program, National Institute of Environmental Health Sciences. We jointly applied for funding from the Sydney Policy Lab, a competition designed to support and deepen partnerships between academics and policy makers. Preliminary findings of the review were presented at an NHMRC Guideline Developer Network meeting. By providing a better understanding of the available tools, this review can aid decision makers in selecting tools to assess risk of bias in observational studies of exposures. Findings of the review can also be used to support the development and critical evaluation of a user-friendly tool that contains items with clear empirical or conceptual justifications.

Methods

Our protocol is registered in PROSPERO: CRD42018094173.

We conducted a systematic review to identify, describe and evaluate published tools purportedly designed to assess methodological quality or risk of bias in human observational studies of the effects of exposures.

Inclusion and exclusion criteria

Tools were included if they met the following criteria:

- were published in English;
- were described as assessing the quality or risk of bias of human observational studies of exposures; including, but not limited to, study designs described as “observational,” case control, cohort, cross-sectional, longitudinal, time-series, and ecological;
- assessed risk of bias in both randomized trials and observational studies, but only information relevant to risk of bias in observational studies were extracted by the review team; and
- were published after 1997 (we included tools from the past 20 years as newer tools are generally more comprehensive and focus more on risk of bias than quality).

We included tools that were described as:

- “Domain-based tools”, where individual items are grouped into categories of bias such as performance or detection bias (2);
- “Scales”, where each item has a numeric score, and an overall summary score is calculated; or

- “Checklists”, which include multiple questions but do not provide any overall summary rating or score (12).

We excluded:

- tools for assessing RCTs only;
- tools for assessing observational studies of interventions only;
- duplicate publications of specific tools (we generally used the latest publication providing a full description of the tool and its application; updates of an existing tool were considered new tools only if they underwent a substantial update (e.g. Cochrane risk of bias tool for randomized trials (2008) versus RoB 2.0 (2016) would be considered different tools);
- commentaries, expert opinions, and letters about tools;
- tools to assess studies of diagnostic accuracy of tests; and
- tools that were used as part of a research study (e.g. a systematic review), but were not referenced or published, and the study was not specifically about the tool or development of the tool.

Search and screening

We searched for publications of tools or about tools using Ovid MEDLINE® in July 2017. Because previous reviews encompassed searches from 1997 to 2005 (8,9), our search was limited to articles published from 1 January 2005 to 1 July 2017. Search concepts included: risk of bias, quality, internal validity, and critical appraisal; and various types of observational studies (e.g. case-control, cohort, cross-sectional, longitudinal, time-series, and ecological). More detail on the search strategy is found in Supplemental file 1.

We searched for previous reviews of relevant tools (6-9). These were thoroughly reviewed by the authors and tools from these reviews that met our inclusion criteria were included in this review.

Given that many quality or risk of bias tools for observational studies have not been published in the peer-reviewed literature, we supplemented our electronic search. A grey literature search was conducted on organizational websites, including: the Joanna Briggs Institute and Critical Appraisal Skills Programme (CASP) and supplemented by searching for tools using search engine Google. Tools identified by the authors were also considered for inclusion.

All searches were conducted between April 1 and July 30, 2017. Two coders (ZW, KT) independently screened tools of inclusion according to the inclusion and exclusion criteria. Any discrepancies were resolved by a third coder (LB).

Data Extraction

The information about each tool was extracted and managed on REDcap (13). One researcher extracted information from all of the tools; two other researchers split the tools randomly into two groups, with each researcher extracting information from their assigned group of tools. Thus, information was independently extracted by two coders for each tool. Any discrepancies between the two coders were discussed and consensus reached.

Information was extracted using an *a priori* protocol which included:

Descriptive items:

- Identifying information (reference, authors of tool, name of tool)
- The number of items in each tool
- Number of domains for methodological quality
- Whether a quality score is calculated
- Whether a quality rating is used (e.g. high, low, moderate)

- Type of study(ies) for which the tool was originally designed. Type of study design was extracted verbatim from the tool and classified.
- Public health or clinical topic of studies for which the tool was developed. Topic was extracted verbatim from the tool (e.g. nutrition, pollution, chemical exposure, etc.)

Development of the tool:

- Whether / how the tool was tested, verbatim description of how the tool was tested
- Whether validity of the tool was reported to have been tested
- Whether reliability of the tool was reported to have been tested
- Purported applicability to different study designs
- Conflict of interest of the developers of the tool (as declared in the paper)
- Sponsor for study (as declared in the paper)
- Methods used to develop the tool (e.g. systematic review of existing tools, Delphi survey, face-to-face consensus meeting, or a combination of methods)
- Accessibility of tool (e.g. cost, published or available through open access, regional, etc.)

I**tems related to methodological quality or risk of bias:**

Items of a tool were defined as the individual questions asked (e.g. Was exposure accurately measured? Was confounding adequately controlled?). Two coders (ZW, KT) extracted items that appeared to be related to methodological quality or risk of bias.

We extracted items relate to precision (e.g. “was a sample size calculated?”) We did not extract items that were clearly designed to assess only the reporting of a study, for example, “Are the baseline characteristics of included patients reported? (14)”. Items that were found in fewer than 3 tools were also not extracted because these most often consisted of items that were specific to the topic being studied. For example, Roth 2014, a tool for assessing studies of

neurodevelopment, had an item: “Was the physical measurement and/or neurodevelopment assessment procedure appropriate and clear?”(15)

Our information extraction form was pre-tested on 20 tools. Coders extracted information from the same 20 tools in two rounds and compared the information that was extracted. Any discrepancies between the coders were resolved with discussion. Once a consensus was reached between the two coders, adjustments to the form were made to make sure that the coders extracted the same items and categorized them consistently.

Synthesis

The information on tool characteristics and development are summarised in evidence tables.

Classification of items

Two coders (ZW, KT) collated the items extracted from each tool into categories of items that had similar meaning. For example, the following 3 questions were grouped under the item “Blinding of the research staff.”

1. “Were the outcome criteria objective and applied in a blinded fashion?”(16)
2. “Was the outcome assessor not blinded to the intervention or exposure status of participants?”(17)
3. “Was an attempt made to blind research staff to the activity levels or characteristics of the participants to avoid biasing the results?”(18)

A third coder (MP) reviewed the categorization of each item. Any discrepancies were discussed with a fourth coder (LB) until consensus was achieved.

The frequency of each item and item category is reported. We further categorized each item category into domains that we termed “quality” domains. Domains were defined as those that may include a number of item categories. For example, “Lost to follow-up” included items about:

1. Adequacy of length of follow-up
2. Amount of loss to follow-up
3. Handling of loss to follow-up

As there is no gold standard for risk of bias domains or common terminology in observational studies, the review team categorized the “quality” domains basing these categorizations on previously published literature that described domains and were applicable to observational studies of exposures (2, 19-21). The 9 “quality” domains (Selection, Exposure, Outcome assessment, Confounding, Lost to follow-up, Analysis, Selective reporting, and Conflict of interest) were derived from the Cochrane risk of bias (RoB) tool (1), the Navigation Guide (21) and National Research Council (NRC) Review of the Environmental Protection Agency’s integrated risk information systems (IRIS) process (19)

Results

We included 62 tools (Figure 1). Citations for the tools are in Table 3. One tool was still under development at the time of data analysis, but we included the latest version made available to us and its user manual (22). Reasons for exclusion during screening are shown in the Flowchart (Figure 1)

The characteristics of the tools are summarised in Table 1. The number of items in the tools ranged from 5 to 53. Tools were accompanied by instruction manuals ranging from 1 to 56 pages. The tools were developed for a variety of clinical or public health topics, and most were

designed to assess multiple observational study designs (68%, N=42). Almost half of the tools calculated a quality score (44%, 27/62) and 40% derived a rating (e.g. “high” quality) (N= 25).

Table 2 summarises data on the development of the tools. Twenty-three (37%) tools did not describe the method used for their development. The methods used to develop the content of the tools included systematic review of existing tools, Delphi survey, consensus meetings, and expert consultation. The most commonly used method was expert consultation (45%, N=28), which was sometimes use in combination with other methods. Twenty-one tools (34%) declared conflicts of interest of developers or funding sources.

We found that 28 tools stated that they were tested during their development, although some did not describe how. We did not develop an a priori list of ways that a tool could be tested, but rather reported how the studies described the testing. Twelve tools were said to have been tested for validity and 22 for reliability but did not describe how they defined “validity” or “reliability”.

The tools were accessible by subscription journals (48%), non-journal related weblinks (39%), and open access journals (11%). The tools available via weblinks usually did not include as much information regarding their development and testing as did tools presented in journal articles.

Table 3 summarises the tool items and domains. We grouped the questions in the tools into 17 item categories, and grouped these into 9 domains: Selection, Exposure, Outcome assessment, Confounding, Loss to follow-up, Analysis, Selective reporting, Conflicts of interest and Other. The most frequently occurring items related to assessing risk of bias in selection and outcome assessment. The accuracy (reliability or validity) of exposure assessment, appropriateness of

statistical analysis, and efforts to minimise risk of bias related to confounding or loss to follow-up were also frequently assessed in the tools. Items related to selective outcome reporting and conflicts of interest were less common.

Many tools that assess aspects of risk of bias are described by their authors as “quality” assessment tools. These were not excluded from our literature search as this may have resulted in overlooking tools that included items relevant to assessing risk of bias. However, some of the items that were frequently included in the tools purported to assess “quality” did not specifically address risks of bias and were, therefore, not included in our categorization of items and domains. For example, the adequacy of sample size or power calculations was assessed in 25 tools (e.g. “Was the study sufficiently powered to detect an effect?”). This item assesses precision, rather than risk of bias. Items related to precision were extracted, however they were not included in a “quality” domain. Two frequently occurring items focused on the clarity of the study, rather than assessment of bias in the design. Twenty-one tools included an item asking about the clarity of the objectives or hypotheses (e.g. “Is the hypothesis / aim / objective clearly described?”). Seven tools included an item assessing whether the conclusions were supported by the results or methods (e.g. “Are the conclusions of the study supported by results?”). Different response options were offered for the questions in different tools. For example, some tools provided definitions of the possible answers for each question. One tool included a question with four response options: Question, “Were outcome data complete without attrition or exclusion from analysis? (23)”; Response options “definitely low risk of bias,” “probably low risk of bias,” “probably high risk of bias,” and “definitely high risk of bias”. Other tools had only a yes or no option for a similar question. Some tools requested that the user provide explanations for their rating, while others did not.

Discussion

We identified 62 tools for assessing methodological quality or risk of bias of human observational studies of effects of exposures. Almost half of the tools calculated a quality score, although such scores are not recommended for use in meta-analyses (6).

Overall, the methods used to develop the tools were poorly described or based mostly on consensus approaches. In addition, descriptions of how the tools were tested were vague and terms such as “validity” and “reliability” were not defined. We did not develop an a priori list of ways that a tool could be tested, but rather reported how the studies described their tool testing process. There are many ways the quality of a tool can be tested, including but not limited to, test-retest reliability, inter-rater agreement reliability, face validity, content validity, internal consistency, criterion validity, respondent burden and usability. How a particular tool performs in each of these tests would have different implications to the reliability, validity and usability of the tools.

None of the included tools reported on their usability. However, many of the tools included over 20 individual items to be coded and had lengthy instruction manuals, suggesting that they could be time consuming to apply.

We found great variability in the way a particular question about a quality or risk of bias item was asked in different tools. For example, questions about the accuracy of outcome measurement included, “Are objective, suitable and standard criteria used for measurement of the health outcomes,”(24) “Was the outcome accurately measured to minimise bias,”(25) “Were the risk factors and outcome variables measured correctly using instruments/measurements

that had been trialed, piloted or published previously,"(26) and "How objective were outcome measures? If significant judgement required, were independent adjudicators used?"(27). Although these questions were all meant to assess the same issue (accuracy of outcome measurement), the usability and accuracy of the assessment may differ depending on how the different questions are interpreted by users of the tool. The variability in how the questions were asked was magnified by the different response options offered for the questions and variability in whether justifications for the ratings were requested. All of these variations in how quality or risk of bias can be assessed could confuse users and obscure the primary meaning of each domain.

Empirical and theoretical support for risk of bias domains

The 9 domains into which items were categorized: Selection, Exposure, Outcome assessment, Confounding, Loss to follow-up, Analysis, Selective reporting, Conflicts of interest and Others were derived from the Cochrane risk of bias (RoB) tool (1), the Navigation Guide (21) and National Research Council (NRC) Review of the Environmental Protection Agency's integrated risk information systems (IRIS) process (19).

Some of the 9 domains have been developed based on empirical evidence about aspects of observational exposure study designs that can affect study outcomes. While risk of bias related to appropriateness of blinding, for example, can be detected in a single study, the risk that this bias will affect the outcomes of a research area can only be detected by examining a body of evidence. "Meta-research" or "meta-epidemiological" research can be used to assess the influence of different study characteristics on effect estimates (28, 29). Most meta-research has focused on assessing bodies of RCTs. For example, meta-research shows that intervention

effect estimates are exaggerated in trials with unblinded (versus blinded) assessment of subjective outcomes (30). Thus, there is empirical evidence to support assessing blinding of outcome assessors as a risk of bias item. There is also empirical evidence supporting the inclusion of items about random sequence generation and allocation concealment (which protect against selection bias) (30), selective outcome reporting (31) and funding (32). This empirical evidence from meta-research on trials indirectly supports the inclusion of similar items in a risk of bias tool for observational studies. However, more research on the influence of specific design features of observational studies of exposures on outcomes is needed to strengthen the empirical foundation for a specific item in a risk of bias tool.

Other domains in our categorization are supported by the conceptual framework underlying the design of observational studies. Confounding, for example, is such an important source of bias in epidemiology that adequate identification of and accounting for confounders should be a primary consideration when evaluating an observational study (19). The accuracy of exposure measurement is also a major factor that can affect the risk of bias in observational studies of exposures (33).

Previous studies have examined and described risk of bias tools individually without synthesizing the items within these tools into categories or domains (6-9). By synthesizing the items into domains, we highlight the domains that are empirically or conceptually supported. We found that questions addressing these domains were asked in different ways, suggesting that further work is needed to create consistent questions that are understandable to users. These domains could be a guide for how tools could be developed in the future.

Limitations

There was great variability in how each item was worded in the different tools and, therefore, others may have grouped the items into different categories. To address this limitation, four authors grouped the individual questions from each tool into item categories, and then domains. We achieved consensus on each item and used domains with empirical or theoretical support that had been derived from prior studies (1, 19-21).

Another limitation is that the tools often did not clearly define or differentiate between items related to study “quality”, “risk of bias” and “reporting” as defined in Box 1. The quality of a research study can be broadly defined as including but not limited to; whether it was peer-reviewed, how completely the study was reported, factors pertaining to both internal and external validity, and best research practices. Risk of bias is the risk of a study over- or underestimating outcome effects due to factors in the design, conduct, or analysis of the study (1). Study reporting can be defined as how clearly and completely the details and procedures of a study are described; for example, “was the objective of the study clearly stated.” We dealt with the limitation that tools had inconsistent definitions of research study features by including as many items as possible that seemed to assess methodological quality or risk of bias. Although we limited our search to tools published in English, it is unclear whether we have missed tools. Previous reviews also had this limitation (9) or did not indicate how many included tools were not published in English (6,8).

Conclusion

Our review provides guidance for decision makers or systematic review practitioners in selecting a tool for assessing bias in human observational studies of exposures for specific projects. This review provides a clear cross-walk across existing tools (available as an online tool on Tableau). We suggest that users should select a tool that contains the nine domains we identified in the synthesis. In addition, users should consider selecting tools that have been tested by potential users, are less complex and publicly available.

Our review highlights the need for continued development of a tool for assessing risk of bias in human observational studies of effects of exposures that will be widely and consistently applied by systematic reviewers, guideline developers, and those conducting environmental hazard and risk assessment. The tool should have clear definitions for each item. Our systematic review suggests that a tool with items addressing 9 domains: Selection, Exposure, Outcome assessment, Confounding, Loss to follow-up, Analysis, Selective reporting, Conflicts of interest and Other should be considered. The ratings for each domain should be reported, rather than an overall score. Such a tool, if developed, should include items based on direct or indirect empirical evidence or theoretical considerations. The tool should be rigorously and independently tested for usability and reliability among stakeholders who need to apply a tool for assessing risk of bias in observational studies of exposures.

References

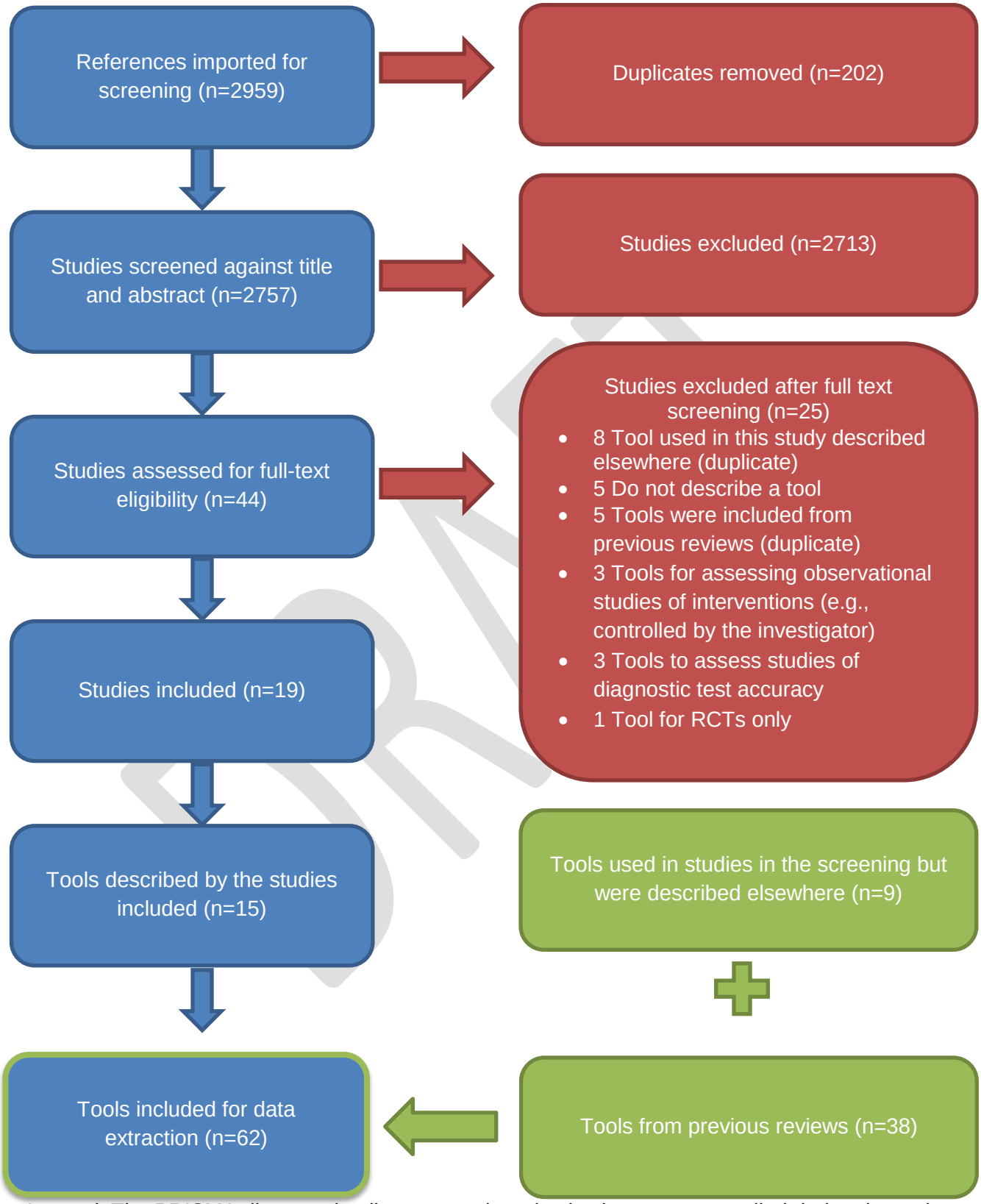
1. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions: John Wiley & Sons; 2011.
2. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343.
3. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. Bmj. 2016;355:i4919.
4. National Health Medical Research Council (NHMRC). Resources for guideline developers 2018 [Available from: https://www.nhmrc.gov.au/guidelines-publications/information-guideline-developers/resources-guideline-developers]
5. World Health Organization. WHO handbook for guideline development: World Health Organization; 2014.
6. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technology Assessment (Winchester, England). 2003;7(27):iii-x, 1-173.
7. Rooney AA, Cooper GS, Jahnke GD, Lam J, Morgan RL, Boyles AL, et al. How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. Environment International. 2016;92-93:617-29.
8. Sanderson S, Tatt ID, Higgins JP. 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.

9. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. Journal of Clinical Epidemiology. 2010;63(10):1061-70.
10. Hoy D, Brooks P, Woolf A, Blyth F, March L, Bain C, et al. Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. Journal of Clinical Epidemiology. 2012;65(9):934-9.
11. Schwingshackl L, Knüppel S, Schwedhelm C, Hoffmann G, Missbach B, Stelmach-Mardas M, et al. Perspective: NutriGrade: A Scoring System to Assess and Judge the Meta-Evidence of Randomized Controlled Trials and Cohort Studies in Nutrition Research. Adv Nutr. 2016;7(6):994-1004.
12. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Controlled Clinical Trials. 1995;16(1):62-73.
13. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377-81.
14. Cowan JB, Mlynarek RA, Nelissen RG, Pijls BG, Gagnier JJ. Evaluation of Quality of Lower Limb Arthroplasty Observational Studies Using the Assessment of Quality in Lower Limb Arthroplasty (AQUILA) Checklist. Journal of Arthroplasty. 2015;30(9):1513-7.
15. Roth N, Wilks MF. Neurodevelopmental and neurobehavioural effects of polybrominated and perfluorinated chemicals: A systematic review of the epidemiological literature using a quality assessment scheme. Toxicology Letters. 2014;230(2):271-81.
16. Carneiro AV. Critical appraisal of prognostic evidence: practical rules. Rev Port Cardiol. 2002;21(7-8):891-900.

17. Viswanathan M, Berkman ND, Dryden DM, Hartling L. Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank. Rockville, MD: Agency for Healthcare Research and Quality; 2013.
18. Hagstromer M, Ainsworth BE, Kwak L, Bowles HR. A checklist for evaluating the methodological quality of validation studies on self-report instruments for physical activity and sedentary behavior. Journal of Physical Activity & Health. 2012;9 Suppl 1:S29-36.
19. National Research Council. Review of EPA's integrated risk information system (IRIS) process: National Academies Press; 2014.
20. Bero LA. Why the Cochrane risk of bias tool should include funding source as a standard item. Cochrane Database of Systematic Reviews. 2013;12:ED000075.
21. Woodruff TJ, Sutton P. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environmental Health Perspectives. 2014;122(10):1007-14.
22. Morgan R, In collaboration with University of Bristol (UK) McMaster University (Canada) and the Environmental Protection Agency (USA). The ROBINS-E tool (Risk Of Bias In Non-randomized Studies - of Exposures) - version July 2017 2017 [Available from: http://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-e/.
23. National Toxicology Program (NTP) Office of Health Assessment and Translation. OHAT Risk of Bias Rating Tool for Human and Animal Studies 2014 [Available from: https://ntp.niehs.nih.gov/ntp/ohat/pubs/riskofbiastool_508.pdf.
24. Loney PL, Chambers LW, Bennett KJ, Roberts JG, Stratford PW. Critical appraisal of the health research literature: prevalence or incidence of a health problem. Chronic Diseases in Canada. 1998;19(4):170-6.
25. Critical Appraisals Skills Programme. CASP Cohort Study Checklist 2018 [Available from: https://casp-uk.net/casp-tools-checklists/.

26. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open. 2016;6(12):e011458.
27. Effective Practice, Informatics & Quality Improvement (EPIQ). GATE CAT workbooks Case control studies 2017 [Available from: https://www.fmhs.auckland.ac.nz/en/soph/about/our-departments/epidemiology-and-biostatistics/research/epig/2017-evidence-based-practice-and-cats.html.
28. Bero L. Meta-research matters: Meta-spin cycles, the blindness of bias, and rebuilding trust. PLoS Biol. 2018;16(4):e2005972.
29. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med. 2002;21(11):1513-24.
30. Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PLoS One. 2016;11(7):e0159267.
31. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA. 2004;291(20):2457-65.
32. Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev. 2012;12:Mr000033.
33. Baker DB, Nieuwenhuijsen MJ. Environmental epidemiology : study methods and application. Oxford ; New York: Oxford University Press; 2008. xiv, 398 p. p.

Figure 1. PRISMA Chart



Legend: The PRISMA diagram details our search and selection process applied during the review.

BOX 1: Definitions

Quality: study characteristics, including but not limited to, whether it was peer-reviewed, characteristics pertaining to internal or external validity, completeness of study reporting, best research practices (e.g. ethical approval) and / or risk of bias.

Risk of bias: study characteristics related to systematic error; to assess the risk of a study over- or underestimating outcome effects due to factors in the design, conduct, or analysis of the study (1).

Internal validity: the validity of the inferences drawn as they pertain to the members of the source population

External validity: the validity of the inferences as they pertain to people outside the source population

Reporting: study characteristics describing how clearly and completely the details and procedures of a study are described (e.g. “was the objective of the study clearly stated.”)

Precision: study characteristics related to random error (e.g. sample size)

Table 1. Characteristics of the tools (n = 62)

Tool name	First author and year	Number of items	Number of domains	Quality Score?	Quality rating?	Topic of study	Type of studies for which the tool was designed
	Al-Jader 2002	9	N/A	Yes	No	Epidemiological surveys of genetic disorders	Cross sectional study, Epidemiological surveys
AQUILA (assessment of quality in lower limb arthroplasty)		6	N/A	No	No	Lower limb arthroplasty	Cohort study, Case series
AXIS	Downes 2016	20	5	No	No	Topic is not specific	Cross sectional study
	Boulware 2002	13	N/A	Yes	Yes	Behavioural interventions for hypertension	RCT, Cohort study, Case control, Case series, Cross sectional study
	Brown 2013	10	N/A	Yes	No	ACL Reconstruction	RCT, Cohort study, Case control, Case series
	Carneiro 2002	8	N/A	No	No	Topic is not specific, Prognostic evidence	Cohort study, Case control

CASP Case Control Checklist		11	N/A	No	No	Topic is not specific	Case control
CASP Cohort Study Checklist		12	N/A	No	No	Topic is not specific	Cohort study
Centre for evidence-based medicine (CEBM) prognosis tool		7	N/A	No	No	Topic is not specific	Prognostic studies
	Coleman 2000	10	N/A	Yes	No	Patellar tendinopathy	RCT, Cohort study
	Downs and Black 1998	27	N/A	Yes	No	Topic is not specific, Health care interventions	RCT, Cohort study, Case control, Case series, Cross sectional study
Effective Public Health Practice Project (EPHPP) tool		20	6	No	Yes	Topic is not specific, Public health	RCT, Cohort study, Case control, Case series,
EPOC Quality criteria for controlled before and after (CBA) designs		7	NA	Yes	No	Topic is not specific	Controlled before and after (CBA)

EPOC Quality criteria for interrupted time series (ITSs)		7	NA	Yes	No	Topic is not specific	Interrupted time series
	Garcia-Alvarez 2009	14	N/A	Yes	No	Nutrient intake adequacy assessment	Cross sectional study, Prevalence study
GATE CAT workbooks Case control studies		14	4	No	No	Topic is not specific	Case control
GATE CAT workbooks Prognostic studies		16	4	No	No	Topic is not specific	Prognostic cohort study
	Genaidy 2007	43	N/A	Yes	No	Topic is not specific	RCT, Cohort study, Case control, Case series, Cross sectional study
	Giannakopoulos 2012	11	3	Yes	No	Disorders with non-standardised examination and	Prevalence study

						diagnostic protocols	
GRACE Checklist Dreyer 2014		11	N/A	No	No	Topic is not specific, Comparative effectiveness	Cohort study, Case control, Case series, Cross sectional study
Hagströmer-Bowles Physical Activity/Sedentary Behavior Questionnaire Checklist (HBQC)	Hagstromer 2012	22	N/A	Yes	No	Self-Report instruments for physical activity and sedentary behavior	Instrument validation studies
HEB wales Critical appraisal checklist		11	N/A	No	No	Topic is not specific, Public health	Cross sectional study
	Hoy 2012	10	N/A	No	Yes	Low back and neck pain	Prevalence study
Joanna Briggs RAPid appraisal protocol (prognostic study)		10	N/A	No	Yes	Topic is not specific, Prognostic study	Prognostic study

Joanna Briggs RAPid appraisal protocol (risk study)		10	N/A	No	Yes	Topic is not specific, Risk study	Cohort Study, Case control
	Kreif 2013	6	N/A	No	No	Topic is not specific, Cost-effectiveness analyses	Cohort study, Case control, Case series, Cross sectional study
	Littenberg 1998	5	N/A	Yes	No	Closed fractures of the tibial shaft	RCT, Cohort study, Case control, Case series
	Loney 1998	8	N/A	Yes	No	Dementia	Cross sectional study, Prevalence study
	Macfarlane 2001	29	N/A	Yes	No	Oro-facial pain	Cohort study, Case control, Cross sectional study

	MacLehose 2000	43		Yes	No	Topic is not specific	RCT, Cohort study, Case control, Case series, Cross sectional study, Prevalence study
	Manchikanti 2002	6	6	Yes	No	Medial branch neurotomy	Cohort study, Case control, Case series, Cross sectional study,
	Manchikanti 2014	17	n/a	Yes	No	Interventional pain management	Cohort Study, Case control, Case series, Cross sectional study, Prospective, controlled study
	Manterola 2006	6	N/A	Yes	No	Human therapy studies in surgical publications	RCT, Cohort study, Case control, Case series, Cross sectional study, Multicenter clinical trial

	Moga 2012	18	N/A	No	No	Topic is not specific	Case series
Newcastle Ottawa scale case control	Wells 2000	8	3	No	Yes	Topic is not specific	Case control
Newcastle Ottawa scale cohort	Wells 2000	8	3	No	Yes	Topic is not specific	Cohort study
NICE Quality appraisal checklist 2012		21	N/A	No	Yes	Topic is not specific, Guideline development	Cohort study, Case control, Case series, Cross sectional study, Quantitative studies reporting correlations and associations
OHAT Risk of Bias Rating Tool for Human and Animal Studies 2015		11	6	No	Yes	Topic is not specific, Human and animal Studies	RCT, Cohort study, Case control, Case series, Cross sectional study

National Toxicology Program Report on Carcinogens tool 2015		5	5	No	Yes	Carcinogens	Cohort study, Case control, Case series, Cross sectional study
	Ohadike 2016	23	N/A	Yes	Yes	Studies aimed at creating gestational weight gain charts	Cohort study, Case control, Case series, Cross sectional study
	Pavia 2006	38	N/A	Yes	No	Association between fruit and vegetable consumption and oral cancer	Cohort study, Case control
Quality Criteria Checklist: Primary Research of the Academy of Nutrition and Dietetics		53	N/A	No	Yes	Topic is not specific	RCT, Cohort study, Case control, Case series, Cross sectional study
	Rangel 2003	22	8	Yes	Yes	Paediatric Surgery	Cohort study, Case control, retrospective clinical study

RoBANS	Kim 2013	6	6	No	Yes	Topic is not specific	Cohort study, Case control, Case series, Cross sectional study
ROBINS-E		35	7	No	Yes	Topic is not specific	Cohort study, Case control, Case series, Cross sectional study
MetaQAT	Rosella 2016	9	n/a	No	No	Topic is not specific, Public health	RCT, Cohort study, Case control, Case series, Cross sectional study, Systematic reviews and meta-analyses, Economic evaluation studies, Mixed methods research, Qualitative research

Systematic Appraisal of Quality for Observational Research (SAQOR) Ross 2011		19	6	No	Yes	Psychiatry	Cohort Study, Case control, Case series, Cross sectional study
	Roth 2014	15	N/A	Yes	Yes	Neurodevelopmental and neurobehavioural effects of polybrominated and perfluorinated chemicals	Cohort study, Case control, Case series, Cross sectional study
	Scholten 2003	16	6	Yes	No	Whiplash-associated disorders	Cohort study
	Shamliyan 2011 (risk factors of diseases)	22	N/A	No	Yes	Topic is not specific	Risk factor study
	Shamliyan 2011 (Incidence)	13	N/A	No	Yes	Topic is not specific	Cross sectional study, Prevalence study

SIGN50 Methodology Checklist 3: Cohort studies		26	5	No	Yes	Topic is not specific	Cohort study
SIGN50 Methodology Checklist 4: Case-Control Studies		23	5	No	Yes	Topic is not specific	Case control
MINORS	Slim 2003	12	N/A	Yes	No	Surgical studies	Cohort study, Case control, Case series, Cross sectional study, Prevalence study,
the quality of genetic studies (Q-Genie) tool	Sohani 2015	11	N/A	No	Yes	Genetic studies	Genetic association studies
	Tooth 2005	33	N/A	Yes	No	Topic is not specific	Cohort study, Case series
University of Montreal Critical Appraisal Worksheet		30	N/A	No	No	Topic is not specific	Not described

	van der Windt 2000	25	N/A	Yes	No	Occupational risk factors for shoulder pain	Cohort study, Case control, Cross sectional study
AHRQ	Viswanathan 2013	16	N/A	No	Yes	Topic is not specific	Cohort Study, Case Control, Case Series, Cross sectional study
The Navigation Guide	Woodruff 2014	8	N/A	No	Yes	Environmental health science	Cohort study, Case control, Case series, Cross sectional study, Ecological
	Zaza 2000	23	6	No	Yes	Community preventive services	RCT, Cohort study, Case control, Case series, Cross sectional study

Integrated quality criteria for review of multiple study designs (ICROMS)	Zingg 2016	32	N/A	Yes	No	Topic is not specific, Public Health	RCT, Cohort study, Case control, Case series, Qualitative, Before and after studies

Abbreviations: N/A – not applicable, RCT - randomized controlled trial, AHRQ - Agency for Healthcare Research and Quality

Table 2. Development of the tools (n = 62)

Tool name	First author and year	Tool tested	How the tool was tested	Validity tested	Reliability tested	COI declaration	Study Sponsor	Methods to develop the tool	Accessibility
	Al-Jader 2002	Yes	Reproducibility tested twice. Feasibility of the scoring system.	No	Yes	Not declared	No	Not Described	Subscription journal
AQUILA (assessment of quality in lower limb arthroplasty)		No				No, Declared that there was no conflict	Yes	Delphi Survey, Expert Consultation	Open access journal
AXIS	Downes 2016	No				No, Declared that there was no conflict	Yes	Systematic review of existing tools, Delphi Survey, Expert Consultation	Open access journal
	Boulware 2002	Yes	Interrater agreement for different items	No	Yes	Not declared	Yes	Expert Consultation, Other	Subscription journal
	Brown 2013	No				No, Declared that there was no conflict	No	Consensus Meeting, Expert Consultation, Other	Subscription journal
	Carneiro 2002	No				Not declared	No	Not Described	Subscription journal
CASP Case Control Checklist		Yes	Experts piloted checklist.	No	No	Not declared	N/A	Expert Consultation	Non-journal Web link
CASP Cohort Study Checklist		Yes	Experts piloted checklist.	No	No	Not declared	N/A	Expert Consultation	Non-journal Web link
Centre for evidence-based medicine (CEBM) prognosis tool		N/A				Not declared	N/A	Not Described	Non-journal Web link
	Coleman 2000	No				Not declared	No	Not Described	Subscription journal

	Downs and Black 1998	Yes	Face and content validity assessed by three experienced reviewers, tested for internal consistency, test retest and inter-rater reliability, criterion validity, and respondent burden.	Yes	Yes	No, Declared that there was no conflict	No	Other	Subscription journal
Effective Public Health Practice Project (EPHPP) tool		No				Not declared	N/A	Not Described	Non-journal Web link
EPOC Quality criteria for controlled before and after (CBA) designs		N/A				Not declared	N/A	Not Described	Non-journal Web link
EPOC Quality criteria for interrupted time series (ITSs)		N/A				Not declared	N/A	Not Described	Non-journal Web link
	Garcia-Alvarez 2009	No				No, Declared that there was no conflict	Yes	Expert Consultation	Subscription journal
GATE CAT workbooks Case control studies		N/A				Not declared	N/A	Not Described	Non-journal Web link
GATE CAT workbooks Prognostic studies		N/A				Not declared	N/A	Not Described	Non-journal Web link
	Genaidy 2007	Yes	Piloted by a team of epidemiologists/physicians/biostatisticians. Revised version was evaluated for criterion validity and reliability.	Yes	Yes	Not declared	No	Expert Consultation, Other	Subscription journal
	Giannakopoulos 2012	Yes		No	Yes	No, Declared that there was no conflict	No	Other	Subscription journal

GRACE Checklist Dreyer 2014		Yes	Tested on observational studies of comparative effectiveness and ratings were compared with A) systematic reviews B) Single Expert Review C) Concordant Expert Review-quality assessments from 2 experts	Yes	No	Yes, Conflicts declared	Yes	Expert Consultation, Other	Subscription journal
Hagströmer-Bowles Physical Activity/Sedentary Behavior Questionnaire Checklist (HBQC)	Hagstromer 2012	Yes	Tested for interrater reliability and feasibility with 6 raters.	No	Yes	Not declared	No	Other	Subscription journal
HEB wales Critical appraisal checklist		N/A				Not declared	N/A	Not Described	Non-journal Web link
	Hoy 2012	Yes	Pretested by one author. In stage 2, 12 studies assessed by three authors. Agreement was examined. For stage 3, six researchers assessed four to six randomly selected studies each and ratings were compared with an experienced rater.	No	Yes	Not declared	No	Expert Consultation, Other	Subscription journal
Joanna Briggs RAPid appraisal protocol (prognostic study)		N/A				Not declared	N/A	Not Described	Non-journal Web link
Joanna Briggs RAPid appraisal protocol (risk study)		N/A				Not declared	N/A	Not Described	Non-journal Web link
	Kreif 2013	Yes	Three independent reviewers piloted the tool on 15 studies.	No	No	No, Declared that there was no conflict	Yes	Expert Consultation, Other	Subscription journal

	Littenberg 1998	No				Not declared	Yes	Not Described	Subscription journal
	Loney 1998	No				Not declared	No	Other	Open access journal
	Macfarlane 2001	Yes	Two assessors independently evaluated all the articles independently, and the results were compared using the kappa statistic.	No	Yes	Not declared	No	Other	Subscription journal
	MacLehose 2000	Yes	Each paper was reviewed by 3 raters, assessed for distribution of k statistics and the percentage agreement.	No	Yes	No, Declared that there was no conflict	Yes	Other	Non-journal Web link
	Manchikanti 2002	No				Not declared	No	Other	Open access journal
	Manchikanti 2014	Yes	Inter-rater agreement calculated for each item.	No	Yes	Yes, Conflicts declared	No	Other	Open access journal
	Manterola 2006	Yes	Face and content validity, and construct validity for extreme groups. Inter-observer reliability.	Yes	Yes	Not declared	Yes	Not Described	Subscription journal
	Moga 2012	Yes	Two reviewers tested for reliability	No	Yes	No, Declared that there was no conflict	Yes	Delphi Survey, Other	Non-journal Web link
Newcastle Ottawa scale case control	Wells 2000	N/A				Not declared	N/A	Not Described	Non-journal Web link
Newcastle Ottawa scale cohort	Wells 2000	N/A				Not declared	N/A	Not Described	Non-journal Web link
NICE Quality appraisal checklist 2012		N/A				Not declared	Yes	Other	Non-journal Web link
OHAT Risk of Bias Rating Tool for Human and Animal Studies 2015		No				Not declared	N/A	Other	Non-journal Web link
National Toxicology		N/A				Not declared	N/A	Not Described	Non-journal Web link

Program Report on Carcinogens tool 2015									
	Ohadike 2016	Yes	Methodological quality criteria developed and validated	Yes	No	No, Declared that there was no conflict	No	Other	Subscription journal
	Pavia 2006	No				No, Declared that there was no conflict	No	Not Described	Subscription journal
Quality Criteria Checklist: Primary Research of the Academy of Nutrition and Dietetics		N/A				Not declared	N/A	Not Described	Non-journal Web link
	Rangel 2003	Yes	Inter-rater reliability; 10 studies assessed by 6 independent reviewers. Examined the extent of agreement between reviewers for each item.	No	Yes	Not declared	Yes	Consensus Meeting, Other	Subscription journal
RoBANS	Kim 2013	Yes	A validation process with 39 NRSs examined the reliability (interrater agreement), validity (the degree of correlation between the overall assessments of RoBANS and Methodological Index for Nonrandomized Studies [MINORS]), face validity with eight experts, and completion time for the RoBANS approach.	Yes	Yes	No, Declared that there was no conflict	Yes	Expert Consultation, Other	Subscription journal
ROBINS-E		N/A				Not declared	N/A	Not Described	Other
MetaQAT	Rosella 2016	Yes	Piloted within several scientific teams. A systematic process was designed to test validity.	Yes	No	No, Declared that there was no conflict	No	Expert Consultation, Other	Subscription journal

Systematic Appraisal of Quality for Observational Research (SAQOR) Ross 2011		Yes	Feasibility testing with several studies selected at random. A research team member not involved in the tool development assessed inter-rater reliability.	No	Yes	No, Declared that there was no conflict	Yes	Expert Consultation, Other	Subscription journal
	Roth 2014	No				No, Declared that there was no conflict	Yes	Other	Subscription journal
	Scholten 2003	Yes	Inter-observer agreement was derived by kappa statistics	No	Yes	Not declared	N/A	Systematic review of existing tools	Subscription journal
	Shamliyan 2011 (risk factors of diseases)	Yes	Pilot test of the checklists. Experts each evaluated 10 articles to test reliability and discriminant validity	Yes	Yes	Not declared	Yes	Systematic review of existing tools, Expert Consultation	Subscription journal
	Shamliyan 2011 (Incidence)	Yes	Pilot test of the checklists. Experts each evaluated 10 articles to test reliability and discriminant validity	Yes	Yes	Not declared	Yes	Systematic review of existing tools, Expert Consultation	Subscription journal
SIGN50 Methodology Checklist 3: Cohort studies		No				Not declared	N/A	Not Described	Non-journal Web link
SIGN50 Methodology Checklist 4: Case-Control Studies		N/A				Not declared	N/A	Not Described	Non-journal Web link
MINORS	Slim 2003	Yes	Articles were assessed by two independent reviewers with different methodological expertise for test-retest reliability. External validity of MINORS was assessed by comparing the MINORS scores with a	Yes	Yes	Not declared	No	Systematic review of existing tools, Consensus Meeting, Expert Consultation	Open access journal

			selected group of the 15 best-scored comparative studies from the sample of 80 described previously.						
the quality of genetic studies (Q-Genie) tool	Sohani 2015	Yes	Validity and reliability tested using a sample of thirty studies randomly selected from a previously conducted systematic review	Yes	Yes	No, Declared that there was no conflict	No	Expert Consultation, Other	Open access journal
	Tooth 2005	Yes	Percentage agreement for two independent raters was calculated. The raters resolved most coding discrepancies by consensus.	No	Yes	Not declared	No	Consensus Meeting, Expert Consultation	Subscription journal
University of Montreal Critical Appraisal Worksheet		N/A				Not declared	N/A	Not Described	Non-journal Web link
	van der Windt 2000	No				Not declared	Yes	Consensus Meeting, Other	Subscription journal
AHRQ	Viswanathan 2013	No				No, Declared that there was no conflict	Yes	Expert Consultation, Other	Non-journal Web link
The Navigation Guide	Woodruff 2014	No				No, Declared that there was no conflict	Yes	Systematic review of existing tools, Expert Consultation	Non-journal Web link
	Zaza 2000	Yes	Pilot-tested for clarity and reliability of responses between reviewers	Yes	Yes	Not declared	No	Systematic review of existing tools, Expert Consultation	Subscription journal
Integrated quality criteria for review of multiple study	Zingg 2016	No				No, Declared that there was no conflict	Yes	Expert Consultation, Other	Subscription journal

designs (ICROMS)									

Legend: Characteristics in this table was derived from information reported in the tools (see Methods). Abbreviations: N/A – no statement regarding this tool characteristic, COI – conflict of interest

Table 3. "Methodological quality" or risk of bias domains and items in the tools for observational studies (n = 62)

Domain	Item category	Example of question	Count*
Selection	Sample representative of target population	Were participants representative of the target population?	27
	Comparability of exposure and comparison groups	Were the comparison groups (exposed/unexposed, cases/controls) recruited from comparable populations?	10
	Appropriateness of eligibility criteria	Were inclusion/exclusion criteria applied equally to all study groups?	8
	Recruitment time frame	Were participants in different groups recruited over the same period of time?	12
	Non-response rate	Is the information regarding the number of patients who were ineligible or who refused to participate adequately reported?	10
Exposure	Validity and reliability of exposure measurement	Was exposure status measured in a standardised, valid and/or reliable way?	19
Outcome assessment	Accuracy of outcome measurement	Were the outcome measures accurate (valid and reliable)?	33
	Blinding of the research staff	Were outcome assessors blinded to the exposure status?	28

Confounding	Description of confounding variables	Are the distribution of confounders clearly described?	11
	Accounting for confounding	Were confounding variables taken into account in the design and/or analysis?	25
Lost to follow-up	Adequacy of length of follow-up	Was follow-up period appropriate/sufficiently long enough to allow development of the outcome?	15
	Amount of loss to follow-up	Were the numbers and reasons for participant withdrawals/drop-outs recorded?	27
	Handling of loss to follow-up	Were appropriate statistical methods used to account for missing data?	6
Analysis	Appropriate statistical methods	Were the statistical methods used to analyse the outcomes appropriate?	26
Selective reporting	Selective reporting of outcomes	Were all measured outcomes reported?	8
Conflict of interest (funding)	Conflict of interest (e.g. funding)	Were there any funding sources or conflicts of interest that may affect the authors' interpretation of the results?	7
Other	Other bias	Is the study free of other biases?	3

*Count = number of tools (out of 65) containing the item