Article:
Lunny, Carole, Tricco, Andrea C, Veroniki, Areti Angeliki et al. (5 more authors) (2021) A methodological review to develop a list of bias items used to assess reviews incorporating network meta-analysis : protocol and rationale. BMJ Open. e045987. ISSN 2044-6055

https://doi.org/10.1136/bmjopen-2020-045987

The development of a risk of bias tool to assess reviews incorporating network meta-analysis: Full protocol, design and rationale

Lunny C, Cochrane Hypertension Review Group and the Therapeutics Initiative, University of British Columbia, Canada
Tricco AC, Knowledge Translation Program, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, 209 Victoria Street, East Building, Toronto, ON, M5B 1T8, Canada
Veroniki A, School of Education, University of Ioannina, Ioannina, Greece
Dias, S, Centre for Reviews and Dissemination, University of York, York, UK
Hutton, B, Ottawa Hospital Research Institute, Ottawa, Canada. Ottawa University, School of Epidemiology and Public Health, Ottawa, Canada
Salanti G, Institute of Social and Preventive Medicine, University of Bern, Switzerland.
Wright J, Cochrane Hypertension Review Group and the Therapeutics Initiative, University of British Columbia, Canada
White IR, MRC Clinical Trials Unit at UCL, London, UK
Whiting P, Population Health Sciences, Bristol Medical School, University of Bristol

Abstract

Background: Network meta-analysis (NMA) is an extension of pairwise meta-analysis which aims to simultaneously synthesise evidence (of effects) from multiple studies on healthcare interventions of interest. Our aim is to develop a tool to assess the degree to which the methods lead to a risk of bias in the review conclusions. Our specific objectives are to: (i) conduct a methodological review to generate a list of potential items for inclusion in such a tool; (ii) based on the findings of the methodological review, decide on the structure of the tool; (iii) conduct a Delphi process to refine the tool; and (iv) pilot test the tool.

Methods: A steering group of experts in tool development, bias and NMAs was convened. We will follow the methods proposed by Whiting (2013) to develop the tool. For the methodological review, we will include tools, scientific papers and editorial standards that present items related to bias, reporting, or methodological quality, or articles that assess the methodological quality of reviews with NMA. We will search MEDLINE, the Cochrane library, and difficult to locate/unpublished literature. Once all items have been extracted, we will combine conceptually similar items, classifying them as referring to bias or to other aspects of quality (e.g. reporting). When relevant, items related to reporting will be re-worded into items related to bias in NMA review conclusions, and then re-worded as signalling questions. The steering group will review and refine the list of items. Feedback from a larger expert group will be obtained via a Delphi survey. Participants will be asked to rate whether items should be included. All agreed-upon items, additional or aggregated items, will be included in a second and possibly a third round of the Delphi survey (depending on the level of agreement). An explanation and elaboration guidance statement will be written for each item included in the final tool version. The tool will be piloted.

Discussion: Patients, healthcare providers and policy makers need the highest quality evidence to make decisions about which treatments should be used in healthcare practice. Being able to critically appraise the findings of systematic reviews that include NMA is central to informed decision-making in patient care. Our research will develop the first tool for assessing bias in the findings of reviews with NMA.

Systematic review registration: Open Science Framework https://osf.io/ncg9t/

Abstract word count: 380

Keywords: indirect comparison, mixed treatment comparison, network meta-analysis, multiple treatment comparison, quality, risk of bias, critical appraisal, tool, checklist, standard

1.0 BACKGROUND

Reviews with network meta-analysis (NMA; i.e., multiple treatment comparisons, indirect comparisons, mixed treatment effects) have gained popularity due to their ability to provide comparative effectiveness of multiple treatments for the same condition [1]. We adopt a broad definition of NMAs, specifically, a review that aims to, or intends to, simultaneously synthesise more than two heath care interventions of interest, irrespective of study design. Reviews that intend to compare multiple treatments with an NMA but then find that the assumptions are violated (e.g. a disconnected network, or studies are too heterogeneous to combine), and that NMA is not feasible will also be included in our definition.

Reviews with NMA have also grown in number. Between 1997 and 2015, 771 NMAs were published in 336 journals from 3459 authors and 1258 institutions in 49 countries [2]. More than three-quarters (n = 625; 81%) of these NMAs were published in the last 5-years. Many organisations such as the National Institute for Health and Care Excellence (NICE) in the UK, the World Health Organization (WHO), and the Canadian Agency for Drugs and Technologies in Health (CADTH) conduct NMAs as they represent the best available evidence to inform clinical practice guidelines [3-5].

Evidence shows that biased results from poorly designed and reported studies can mislead decision-making in healthcare at all levels [6-9]. If a review is at risk of bias and inappropriate methods are used, the validity of the findings can be compromised [10-12]. Evaluating how well a review has been conducted is essential to determining whether the findings are relevant to patient care and outcomes. Several empirical studies have shown that bias can obscure the real effects of a treatment [13-16]. Being able to appraise reviews with NMA is central to informed decision-making in patient care.

A well-conducted review draws conclusions that are appropriate to the evidence reviewed, and can therefore be free of bias even when the primary studies included in the review have high risk of bias. The systematic procedures that are required to conduct a systematic review (e.g. double and independent data extraction and comparison of the extractions) help mitigate the risk of bias. However, bias can also be introduced when interpreting the reviews findings. For example, review conclusions may not be supported by the evidence presented, the relevance of the included studies may not have been considered by review authors, and reviewers may inappropriately emphasise results on the basis of their statistical significance [17].

Tools are available for most study designs to make quality assessment easier for a knowledge user (e.g. healthcare practitioners, policy-makers, citizens, media outlets [18]). Many tools and checklists can be used either when conducting a systematic review, or when knowledge users want to assess the reporting or methodological quality of the conclusions of a review. The methodological quality of studies (i.e., how well the study is conducted) is often confused with reporting quality (i.e., how well authors report their methodology and results). A risk of bias assessment is an assessment of review limitations, which focus on the potential of those methods to bias the study findings [23].

Currently, several checklists exist for critically appraising reviews with NMA: e.g. PRISMA statement extension for reviews incorporating network meta-analysis (PRISMA-NMA) [19] and the National Institute for Health and Care Excellence Decision Support Unit checklist (NICE-DSU) [20] for assessing reporting quality, International Society for Pharmacoeconomics and Outcomes Research (ISPOR) checklist [21] for assessing credibility and applicability; and Dias 2018 [22] for assessing validity. These tools were designed with different purposes; some for assessing reporting quality in reviews with NMA, some for assessing conduct, applicability or validity. These review-level tools are not to be confused with tools to assess the individual primary studies included in systematic reviews (e.g. Cochrane risk of bias tool for randomised controlled trials [23]). A table comparing the different tools is found in **Table 1**, indicating whether an equivalent tool for reviews including NMAs exists.

More than 40 tools have been identified [24, 25] for critically appraising the quality of reviews. AMSTAR (A MeaSurement Tool to Assess the methodological quality of systematic Reviews; [26]) and the OQAQ (Overview Quality Assessment Questionnaire [27]) have been identified as the most commonly used, and they follow a simple checklist format for critically appraising reviews [25, 28]. AMSTAR has been recently updated to AMSTAR 2, which aims to evaluate how reviews are planned and conducted [26, 29].

Table 1. Tools and checklists to aid systematic review conduct, or to assess the reporting or methodological quality of a review

Tool purpose	Examples of tools or checklists	Description of an example tool	Available tool for reviews with NMA
Guidance for conducting systematic reviews	MECIR [30]	Detailed methodological guidance for systematic reviews of effectiveness (including integration of qualitative data), diagnostic test accuracy, individual patient data reviews, overviews and reviews in public health and health promotion	Not at present, but in process with Cochrane
Assess the quality of published reviews	AMSTAR-2 [26, 29], OQAQ [27]	**AMSTAR-2** is a critical appraisal tool to assess the quality of conduct of reviews of randomised controlled trials of interventions	No
Assess the risk of bias of published reviews	ROBIS [23]	**ROBIS** is a tool for assessing the risk of bias in reviews. It is aimed at four broad categories of reviews mainly within health care settings: interventions, diagnosis, prognosis, and etiology.	Not at present, but in process **RoB-NMA tool**
Assess the certainty in evidence and the strength of recommendations in health care	GRADE [33]	**GRADE approach** defines the certainty of a body of evidence as the extent to which one can be confident that a pooled effect estimate is close to the true effect of the intervention. Five domains assessed: risk of bias, inconsistency, indirectness, imprecision, and publication bias.	GRADE-NMA [34, 35], CINeMA, [36], Threshold method [37]
Guidelines for the complete reporting published reviews	PRISMA [38]	**PRISMA** is an evidence-based minimum set of items for reporting in reviews and meta-analyses. **PRISMA** focuses on the reporting of reviews evaluating randomized trials, but can also be used as a basis for reporting reviews of other types of research, particularly evaluations of interventions.	PRISMA-NMA [19]

AMSTAR-2: A Measurement Tool to Assess Systematic Reviews 2; CINeMA: Confidence in Network Meta-Analysis; Cochrane RoB 2: Cochrane risk-of-bias tool for randomized trials Version 2.0; GRADE: (Grading of Recommendations Assessment, Development and Evaluation; GRADE-NMA: (Grading of Recommendations Assessment, Development and Evaluation for Network Meta-Analysis; MECIR: Methodological Expectations of Cochrane Intervention Reviews; N/A: not applicable; OQAQ: Overview Quality Assessment Questionnaire ; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT: Randomised Controlled Trial; ROBINS-I: Risk Of Bias In Non-randomized Studies - of Interventions.

Only ROBIS (Risk Of Bias In Systematic reviews) is designed to assess the risk of bias in reviews with or without pairwise meta-(a description of ROBIS is found in **Appendix A**) [32]. The ROBIS tool involves assessment of methodological features in reviews known to increase the risk of bias categorised into four domains (study eligibility criteria; identification and selection of studies; data collection and study appraisal; and synthesis and findings). Assessors using the ROBIS tool use signalling questions to make judgements for the domain-based risk of bias, supported by quotes from the review manuscript and protocol, and rationale for their judgments. Domain-based assessment tools require a careful reading and thoughtful analysis of the study to adequately rate risk of bias, instead of identifying keywords reported in the article, as usually made in a checklist type of assessment. No tool currently exists specifically to assess the risk of bias in reviews with NMA.

As systematic reviews with NMA contain many of the same steps as conducting a systematic review with pairwise meta-analysis, ROBIS can potentially be used as a starting point in the development of an extension for assessing the risk of bias in reviews with NMA, either as an extension to the ROBIS tool or as a stand alone tool. PRISMA-extension checklists for different review types have been developed as adaptations of the original PRISMA checklist for systematic reviews (e.g. [19]). After a methodological review of items related to risk of

bias in NMA, the steering committee will make a collective decision about whether the risk of bias tool should take the form of an extension to ROBIS, or whether it should be a new stand-alone tool.

2.0 OBJECTIVES AND PROJECT PROTOCOL

The aim is to develop a tool to assess the risk of bias in reviews with NMA.

Our specific objectives are to:

1. Conduct a methodological review to develop a list of items relating to risk of bias in NMAs;
2. Based on the findings of the methodological review, decide if we should use the ROBIS tool and extend it to reviews with NMA, or develop a stand alone tool;
3. Conduct a multi-round Delphi process to select and define the items, and compile the items into a tool;
4. Pilot test and then refine the draft tool with different user groups.

The protocol for this project will be registered in the Open Science Framework (https://osf.io/ncg9t/), and we will publish the methodological review in a peer reviewed journal. We will follow the methodology proposed by Whiting [40] to develop the tool.

The development of the tool will be multi-staged. In the first stage, a methodological review to identify items related to bias in reviews with NMA will be conducted; second, the steering group will make conceptual decisions about the type of tool that will be developed, and refine the items from the methodological review; third, expert opinion will be obtained through a Delphi survey to select and define the items, and compile the items into a tool; and finally pilot test and refine the tool.

3.0 MANAGEMENT OF THE PROJECT

3.1 Assemble Team

A steering group was convened of nine experts in NMA and risk of bias tool development (**Appendix 2)** [32]. The steering group will be responsible for the management of the project and will have executive power over all decisions related to the new tool. The steering group will meet through videoconferencing monthly, or as needed, to manage the project.

3.2 Conceptual decisions to be made by the steering group

For the purpose of this project, we will adopt a broad definition of reviews with NMA, namely: A review that aims to, or intends to, simultaneously synthesise more than two healthcare interventions of interest, irrespective of study design. Reviews that intend to compare multiple treatments with an NMA but then find that the assumptions are violated (e.g. a disconnected network, or studies are too heterogeneous to combine), and that NMA is not feasible will also be included in our definition. Our RoB NMA tool will address the degree to which the methods lead to a risk of bias in the review conclusions. Important concepts and definitions are found in **Appendix 3.**

The steering group will initially make several conceptual decisions about the structure of the tool, including whether to adopt or develop alternatives to the following concepts:

- Domain based structure, and categorisation of domains (e.g. eligibility criteria, data collection);
- Domain level ratings judged as “low,” “high,” or “unclear”;
- Signaling questions answered as “Yes”, “Probably Yes”, “Probably No”, “No” and “No Information”, with “Yes” indicating low concerns; and
- Include assessment of relevance/applicability.

Domain based tools based on signalling questions reduce the practice of assigning quality or summary scores to individual studies. Signalling questions are used to flag potential for bias in reviews with NMA.

3.3 Delphi group of experts

A larger group of experts will be recruited that will provide broader input through a Delphi web survey on the conceptual decisions, the definitions, whether we should develop a new tool or adapt from ROBIS, and the items to be included. This Delphi group will be a group of experts in NMA and/or risk of bias tool development (see section 6.0 for more details).

4.0 OBJECTIVE 1 METHODS

The objective of this first methodological review is to develop a list of items relating to risk of bias in NMA. We will follow the methodology proposed by Sanderson and Page [7, 41] for creating systematically developed lists of quality items.

4.1 Eligibility criteria

We will include studies of two types. Study type 1 are tools, papers or editorial standards that present and describe items related to bias, reporting, or methodological quality of reviews with NMA. Items related to reporting will be retained because they can potentially be translated into a risk of bias item. For example, in the PRISMA-P guideline [42], one item asks whether study characteristics were used as criteria for eligibility. Full reporting of all outcomes in a protocol may prevent the introduction of bias into the study selection process of a published systematic review. Study type 2 are studies that assess the methodological quality of reviews with NMA.

Study Type 1 (i.e., tools, papers, editorial standards) will meet any of these inclusion criterion:

- Tools, checklists, scales, instruments or standards describing items related to risk of bias or methodological quality in reviews with NMA (e.g. Dias 2018 [22]); tools that only assess general aspects of reviews without focusing specifically on NMA will be excluded (e.g. AMSTAR [26]), AMSTAR 2 [26, 29] or ROBIS [32]).
- Articles or reports identifying or addressing sources of bias and variation in NMA and published after PRISMA-NMA in 2014. Studies of any design, including reviews, and any topic area are eligible.
- Articles, reports or webpages describing editorial standards for reviews with NMA (e.g. similar to the Cochrane MeCIR (Methodological standards for the conduct of new Cochrane Intervention Reviews) standards for reviews [30]);
- Tools or papers describing items related to reporting quality in reviews with NMA (e.g. PRISMA-NMA [39]).

Study Type 2 (i.e., cross-sectional studies comparing quality of NMAs) will meet any of these inclusion criterion:

- Papers assessing the methodological quality (or risk of bias) of reviews with NMA (i.e. a sample of NMAs are assessed for methodological quality; e.g. Chambers 2015 [43]) using criteria that focus specifically on aspects of NMA not just on general aspects of systematic reviews.;

We will include papers with any publication status and in any language, and where the co-authors are not fluent in the language, Google Translate will be used.

If through our main search, we identify a systematic review encompassing the eligible reports, or one aspect of the eligible studies, we will use the results of the review and only include primary studies published subsequent to the review. For example, a review by Laws et al. in 2019 [5] identified all guidance documents for conducting an NMA from countries throughout the world. We therefore would not search for guidance documents published before the last search date of this review.

4.2 Search strategy

A systematic search strategy will be adapted by two methodologists (CL, PW) without limitations to publication type, status, language, or date to identify existing tools. Two search algorithms for risk of bias/quality assessment tools were adapted [32, 44], and combined with a search strategy used by Zarin et al. [45] for retrieval of reviews with NMA. An information specialist will check the search strategy for MEDLINE Ovid and assess the strategy using the PRESS (Peer Review Electronic Search Strategies) guidance [46].

We will search Ovid MEDLINE, the Cochrane library, the following grey literature databases: the EQUATOR Network (http://www.equator-network.org/reportingguidelines/), Dissertation Abstracts, websites of evidence synthesis organisations (Campbell Collaboration Cochrane Multiple Treatments Methods Group, CADTH, NICE-DSU, Pharmaceutical Benefits Advisory Committee, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, European Network for Health Technology Assessment, Guidelines International Network, ISPOR, International Network of Agencies for Health Technology Assessment, and JBI), and methods collections (e.g. Cochrane Methodology Register, AHRQ Effective Health Care Program). We will validate the MEDLINE strategy by using the PubMed IDs of ten included studies (identified by experts prior to our eligibility screening) and evaluating whether the strategy identified the PMIDs (**Appendix 4**). The full search strategies for all databases and websites can be found in **Appendix 4**. To identify other potentially relevant studies, we will examine the reference lists of included studies.

We will search the reference section of a bibliometric study of reviews with NMAs [47] and extract the name of the journals that publish NMAs. We will then contact their editors in chief and ask if they have any in-house editorial standards for reviews with NMA. Journal editors will be asked to participate in the Delphi survey outlined below.

We will ask members of the steering committee, who are experts in methods for NMA, to identify articles missed by our search. We will contact authors of abstracts to retrieve the full report or poster.

4.3 Process for screening, data extraction and analysis

The eligibility criteria will be piloted by reviewers independently on a sample of 25 citations (titles and abstracts) retrieved from the search to ensure consistent application. After high agreement (>70%) is achieved, the Covidence [48] web-based tool (https://www.covidence.org) will be used by two reviewers to independently screen the citations based on the eligibility criteria. Disagreements in coding of titles/abstracts will be discussed until consensus is reached. A third reviewer (CL) will arbitrate if disagreements cannot be resolved.

After a pilot test on 25 articles, the full text of potentially eligible studies will be retrieved and independently screened by two reviewers. Disagreements will be resolved through discussion, and if necessary by a third senior author (CL).

The data extraction form will be piloted by reviewers independently on a sample of five included papers to ensure consistent coding. Two independent authors will extract data on the characteristics of the studies and items. Any disagreements by duplicate reviewers will be arbitrated by a third senior author.

4.4 Data extraction

The characteristics of the sources will be extracted including the type of article (coded as per our inclusion criteria), all listed authors, the corresponding authors email, year of publication, commissioning organisation, purpose of the tool or paper, and methods used to develop the tool. We will also extract the name of the tool, the intended audience (e.g., researcher, peer reviewer, etc.), and whether items were presented as statements, questions or prompts. We will attempt to retrieve the emails of the authors of the included studies

using Google, as these will then be used for recruitment of the members of the Delphi committee. We will use a free online email checker (https://snov.io/) to verify whether the email addresses are active.

Data will be extracted on items, criteria and guidance that are potentially relevant to the risk of bias or quality of reviews with NMAs. Items and their guidance will be initially extracted verbatim from both the tool and the accompanying "explanation and elaboration" document (e.g. instruction manual) (if available) [41].

A table of tool characteristics will be developed with the following headings: (first author, year); number of items; type of tool (tool, scale, checklist, or domain-based tool); whether the tool is designed specific topic areas (specify); domains within the tool; whether the item relates to reporting or methodological quality (or other concepts such as precision, acceptability); how items and domains within the tool are rated; methods used to develop the tool (e.g. review of items, Delphi study, expert consensus meeting); and the availability of an "explanation and elaboration" [7]. Items retained will be referenced to their corresponding "explanation and elaboration" document, and the guidance on the item will be extracted and tabulated. A sample data extraction tool is found in **Appendix 5**).

Two seminal tools were chosen for extraction first because (a) they have the most comprehensive list of items, and (b) they were rigorously developed: the NICE ISPOR [21] and the PRISMA NMA extension [19] checklists. The items from these two tools will be mapped to the ROBIS tool. Once the items from these two tools are extracted and mapped to ROBIS, a new source will be reviewed one at a time based on year of publication (newest first) [41]. It is hypothesised that old tools would contain outdated methods, and are not as comprehensive. Data will be extracted using Microsoft Excel.

4.5 Data analysis

Once the items have been compiled and mapped to ROBIS items and domains, the following steps will be used when analysing items:

1. Split items so that each item covers a single concept

Two or more concepts grouped in one item will be split so that each item covers a single concept. A rationale as to why the item was split will be described. For example, PRISMA-NMA item 15 ("Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies)") will be split into two items because this item is represented by two items in ROBIS in the synthesis and findings domain, namely "4.5 Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?" and "4.6 Were biases in primary studies minimal or addressed in the synthesis?".

2. Group similar items

Items that are conceptually similar will be grouped together and noted with the source. If items are worded vaguely or are unexplained, we will use an iterative process to interpret the item and ensure there is a mutual understanding of the item between authors when coding. The process will be iterative, and if any gaps in items related to bias in reviews of NMA are identified, a new item will be inferred.

3. Omit duplicate items (but keep these in a column in the table for transparency)
4. Map to ROBIS domains

Items will be mapped to ROBIS, and to the items within the domains (study eligibility criteria; identification and selection of studies; data collection and study appraisal; and synthesis and findings). Items that do not clearly map to the existing ROBIS domains will be listed separately and grouped by similar concept. New domains may be created if items do not fit well into the established ROBIS domains.

5. New items not fitting into ROBIS domains will be grouped by concept

We will classify items as relating to bias or other aspect of quality (e.g. reporting). When relevant, items related to reporting will be reworded into items related to bias in NMA review conclusions. Depending on the type of tool we develop, we may reword items to match ROBIS semantics, so that they are questions where an answer of "yes" suggests absence of bias. The final list of items deemed unique (i.e. same conceptual or methodological issue) will be retained. Items will then be reworded as signalling questions so that each item is phrased so "yes" is good. We will provide examples to illustrate the items, and write short descriptions for the items.

We will count the number of sources and unique items included. We will summarise the characteristics of included tools in tables and figures [7]. We will calculate the median and interquartile range (IQR) number of items across all tools and tabulate the frequency of different biases identified in the tools.

5.0 DECISION ON DEVELOPMENT OF A NEW TOOL OR A ROBIS EXTENSION

5.1 Objective

Based on the findings of the methodological review, the steering group will decide if a new tool should be developed, or if we should use the ROBIS tool and extend it to reviews with NMA.

5.2 Steering group examination of the results

Results of the methodological review will be presented to the steering group in a virtual meeting. The Steering group will evaluate and review the potential list of items for inclusion in the Delphi survey and their mapping to the original ROBIS tool.

Based on the findings, the group will make several conceptual decisions:

1. What type of tool should be developed: an extension to ROBIS or a stand alone tool?
2. If we decide to extend ROBIS, do we include a single domain with all items related to NMA, or modify existing domains (considering modifying all domains or just the synthesis and findings domain)?
3. What is our definition of a review with network meta-analysis/indirect comparisons?
4. What items need to be added, grouped, omitted, split, or reworded?
5. Will the tool include all primary study designs?

The decision on the structure of the new tool is likely to be influenced by how closely the proposed items map to the existing ROBIS tool. The Steering committee will also discuss the wording and meaning of items, and review their categorisation by domain. The final list of items will then be drafted and grouped by domains.

6.0 EXPERT FEEDBACK AND REFINEMENT OF THE ITEMS

6.1 Delphi group process

Feedback from the Delphi group will be obtained via a web survey. The Delphi method is a consensus group method used to synthesize expert opinions when evidence is lacking, limited, or contradictory. The Delphi method has the capacity to include a large number of participants who are geographically dispersed, requires minimal support structure (making it relatively inexpensive), and avoids undue dominance by particular individuals through anonymity [49].

6.2 Recruitment of the Delphi group

To obtain a minimum of 22 experts, the median for creating consensus-based standards [50], a sample of 50 experts or more will be asked to participate.

Experts will be identified using a purposive sampling strategy. An email list of authors of quality tools and reviews with NMA will be extracted as part of the methodological review. We will contact key organizations developing methods for reviews with NMA (e.g. Cochrane Multiple Treatments Methods Group **Appendix 4**).

We will recruit experts by using the academic Twitter handles of @carole_lunny, @Drug_Evidence, @ATricco, @SPORAlliance, @cochranemthds, and @CochraneStats and by asking the steering group for suggested expert contacts. We will aim to include experts from as many different countries as possible. Email addresses will be collected from personal contact lists and publicly available sources (e.g., organizational websites). All individuals will be assured confidentiality of their responses.

Ethics approval will be sought from the University of British Columbia Behavioural Research Ethics Board. Informed consent will be obtained from all participants during online registration, by requesting that participants indicate consent by clicking on the consent box (**Appendix 5**: Introduction to the survey and participant consent form). All participants will be given the opportunity to withdraw from the survey at any time.

6.3 Delphi survey methods

The UBC Survey Tool provided by Qualtrics (https://it.ubc.ca/services/teaching-learning-tools/survey-tool) will be used as the survey platform. It is an easy-to-use, top-tier survey tool platform that offers a wide range of features. It complies with the BC Freedom of Information and Protection of Privacy Act (FIPPA) because the survey data is kept secure and is stored and backed up in Canada. To anonymise responses, the Qualtrics Emailer will be used with the option to "anonymize responses" ticked. This will allow us to track responses, and send selective reminders; however, no personal data or IP addresses will be recorded under this option.

Round 1 will collect demographic information (occupation/field and place of employment) and participants' self-rated level of expertise in participating in this process. The survey will provide definitions of all important concepts. Items will be grouped by domain to mirror the structure of the new tool.

Participants of the Delphi survey will rate items based on importance [51]. The importance of each item will be rated on a 5-point Likert scale of 1 (not important - should be dropped as an item to consider) to 5 (very important – must be included) or unable to score [51, 52]. There is current academic debate as to whether a 5-point Likert scale is as effective as a 7-point Likert scale [53]. A 5-point Likert scale was chosen for this survey because it is faster to complete. If participants do not provide a rating, the item will be recorded as missing and no imputation of missing values will be conducted.

Participants will also be asked to comment on whether they prefer to modify or reword the items [41]. In addition, survey respondents will be given the opportunity to propose additional items throughout the survey. Free text comment boxes will allow experts to provide additional comments. The responses and comments will be used to refine the tool during each round. Non-responders or those failing to complete each round will be sent a maximum of three email reminders, at one week intervals, per survey round [52].

The Delphi process will continue until a high level of agreement is reached or three survey rounds have been completed [49]. A high level of agreement is defined if there is at least 70% agreement on each item's importance (e.g., at least 70% of participants scored 4 or above on the 5-point Likert scale [52].

We will use the decision criteria in **Table 2** [54] for item inclusion, exclusion, and further consideration after each round. Participants will be provided with their previous rating of each item, group summary ratings (medians, IQRs and frequency distributions) and anonymized free text comments after the first round [55]. In a second survey, all agreed upon items and any aggregated or new items from the first round, will be included in the second survey. For example, if in the round 1 survey several comments were made about an item(s) covering multiple concepts, a proposal will be made to split items. Experts who complete the first survey will be eligible to participate in the follow-up surveys.

Table 2. Decision criteria for inclusion, exclusion, and further consideration of potential items*

Scenario (rounds 1 and 2)	Handling of information
Item scored 4-5 (moderately to very important) by ≥70% of participants with no suggested changes to wording or content	Consensus achieved for inclusion in NMA tool. Further consideration in a subsequent Delphi round not needed.
Item scored 4-5 (moderately to very important) by ≥70% of participants with minor suggested changes to wording	Consensus achieved for inclusion. Further consideration in a subsequent Delphi round not needed. Minor modifications in wording to be decided by the steering committee.
Item scored 4-5 (moderately to very important) by ≥70% of participants with suggested changes to content (major changes in wording)	Include in following Delphi round.
Item scored 3 (somewhat important to neutral) by ≥70% of participants (regardless of wording or content changes)	Include in following Delphi round.
Item scored 1 or 2 (not important) by ≥70% of participants	Do not include in NMA tool.
Item not achieving consensus criterion.	Include in following Delphi round.
New items nominated by participants	Include in subsequent round. Follow decision criteria scenarios above.

*Table adapted from Stevens et al. 2018 [54]

Each survey round will take two weeks to complete, followed by a one-week period for analysis. For three potential rounds, it is estimated that it will take 9 weeks to send out the surveys, collect data, modify the next stage survey, and finalize the items. However, this will depend on the number of survey rounds needed. All Delphi participants will have the opportunity to be acknowledged with group authorship if they provide feedback on the overall tool and fulfill ICMJE criteria.

An overall response rate will be calculated as well as summary statistics for each item. We will report the number of participants after each round. We will record the geographic background and expertise of the participants [49, 55]. The qualitative data from the free text questions will be analysed through thematic analysis. If there is very little change after round 2 a third round will not be undertaken.

6.4 Steering committee refinement of the items and tool development

The results of the Delphi process will be presented to the Steering committee for final discussion and consensus on what changes and edits should be made. The Steering committee has the final decision in any proposed changes suggested by the Delphi group to items. All final items will be organised by domains and organised into a tool structure.

6.5 Elaboration document

Once the steering committee has made the final changes, an explanation and elaboration will be written for each item. Co-authors will be assigned one or more items to draft the explanation and elaboration and to find an example that best illustrated reporting of that item(s). The document will then be circulated to the steering committee for commenting.

7.0 PILOT TEST THE TOOL

The piloting of the tool will be done online, or in person if appropriate. During the meeting, we will present the structure of the tool and train participants about each of the items and their meaning. Participants will be 10 experts and 10 non-experts at the Knowledge Translation Program at St. Michael's Hospital in Toronto, Canada, and 10 experts and 10 non-experts at the University of British Columbia in Vancouver, Canada who will not be involved in in the previous Delphi process. Before the study, a self-assessment questionnaire on expertise and experience in quality assessment will be sent to participants (covering working experience (in years), number of

reviews assessed using ROBIS, number of reviews assessed using AMSTAR, AMSTAR-2, R-AMSTAR or OQAQ, number of reviews with NMA assessed using PRISMA-NMA or ISPOR, and number of reviews assessed with any other tool for NMAs).

We will use articles collected as part of an NMA database (Petropoulou et al [56] and Zarin et al [57]) to draw a current random sample of 5 to 10 reviews with NMAs published from 2012 onwards. This database contains published NMAs in peer reviewed journals, as well as a smaller percentage of NMAs from NICE and other health technology assessment (HTA) agencies. To ensure an adequate sample of NMAs from HTA agencies, we will also draw a sample of 5 recent NMAs from NICE guidelines and other agencies conducting heath technology assessments from 2012 to 2020. We will describe and tabulate the characteristics of these reviews (first author, year of publication, medical classification of the intervention being studies, number of included RCTs, sample size, primary outcome).

The first phase will be for participants to gain experience using the tool. The groups will be asked to assess 3 reviews with NMA independently and then compare their results with a partner. Participants will be asked to comment on the understandability of the items and explanations and whether they need to be reworded, as well as on the item structure (i.e. the need to add, re-group or group, omit, or split items). If there are serious differences in the application of the tool in this first phase, exploration of misinterpretation of the items will be recorded, and additional assessments will be undertaken. Decision rules on interpretation of the items in the tool will be noted by the facilitators.

When this first process is complete, they will then be asked to assess two more reviews with NMA using the tool, and each assessment will be timed. We will collapse “yes” and “probably yes” answers before the analysis. Inter-rater reliability testing will be used to compare agreement between the participants' assessments. Scores will be compared between raters and the consensus of reviewer pairs across each domain and item will be calculated using the kappa statistic. Interpretation of the Cohen's kappa (κ) is found in Table 3, and we judge that any kappa below 0.60 indicates inadequate agreement among the raters [58]. We also calculated inter-rater reliability as a mean of all the tool’s items for the Cohen’s kappa (κ) using the method for nominal scaled data.

Table 3: Interpretation of Cohen’s kappa (κ)

Value of Kappa	Level of Agreement	% of data that are reliable
0–.20	None	0–4%
.21–.39	Minimal	4–15%
.40–.59	Weak	15–35%
.60–.79	Moderate	35–63%
.80–.90	Strong	64–81%
Above .90	Almost Perfect	82–100%

* Table from McHugh 2012 [58]

In Table 3 above, the column “% of data that are reliable” corresponds to the squared kappa, an equivalent of the squared correlation coefficient, which is directly interpretable [58]. Squaring the kappa translates conceptually to the amount of accuracy (i.e. the reverse of error) in the data due to agreement between raters.

The time to complete each tool will start when the reviewer begins reading the SR and applying the tool (which may occur simultaneously) and will end when the appraisal is fully complete [59]. The time to reach consensus for each tool and each SR will start once the reviewers convene and will end when agreement is established.

If possible, we will perform this piloting exercise as well with groups at the Cochrane Colloquium, the Guidelines International Conference, the Evidence Based Medicine Live conference, and the Society for Research Synthesis Methodology conference. The Steering committee will discuss the result of the pilot exercise and come to consensus on what changes should be made.

8.0 DISSEMINATION

The Steering committee will develop an integrated knowledge translation plan, such as launching the tool on websites, twitter campaign, emailing NMA authors directly, and LinkedIn. The tool will be summarised into an executive summary using the Strategy for Patient-Oriented Research (SPOR) Evidence Alliance and Cochrane methods and the final tool will be presented in workshop format at the Cochrane Colloquium, the Guidelines International Conference, the Evidence Based Medicine Live conference, and the Society for Research Synthesis Methodology conference. The tool will be presented in a short and long form video, and we will provide 1-page tip sheets and further examples of how to answer each item. We also plan to publish three papers in peer reviewed journals, namely (a) a methodological review of items for assessing the risk of bias in network meta-analyses provides, (b) a Delphi process to develop the tool, with pilot testing of face and content validity of the tool, and (c) a paper outlining the tool with its accompanying elaboration and guidance document.

9.0 CONCLUSION

Patients, healthcare providers and policy makers need the highest quality evidence to make decisions about which treatments should be used in healthcare practice. Being able to critically appraise the findings of reviews with NMA is central to informed decision-making in patient care. Our research aims to develop the first tool for assessing bias in the findings of reviews with NMA.

List of abbreviations
AMSTAR-2: A Measurement Tool to Assess Systematic Reviews 2; CINeMA: Confidence in Network Meta-Analysis; Cochrane RoB 2: Cochrane risk-of-bias tool for randomized trials Version 2.0; GRADE: (Grading of Recommendations Assessment, Development and Evaluation; GRADE-NMA: (Grading of Recommendations Assessment, Development and Evaluation for Network Meta-Analysis; MECIR: Methodological Expectations of Cochrane Intervention Reviews; N/A: not applicable; NMA: network meta-analysis; OQAQ: Overview Quality Assessment Questionnaire ; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT: Randomised Controlled Trial; ROBINS-I: Risk Of Bias In Non-randomized Studies - of Interventions.

Declarations
Ethics approval will be sought from the University of British Columbia Behavioural Research Ethics Board. Informed consent will be obtained from all participants during online registration, by requesting that participants indicate consent by clicking on the consent box (**Appendix 4**: Introduction to the survey and participant consent form). All participants will be given the opportunity to withdraw from the survey at any time.

Consent for publication
Not applicable

Availability of data and materials
Not applicable

Competing interests
The authors declare that they have no competing interests.

Funding
No grants or funding were successful for this important study. Andrea Tricco currently holds a Tier 2 Canada Research Chair in Knowledge Synthesis and an Ontario Ministry of Research and Innovation Early Researcher Award. Brian Hutton has previously received honoraria from Eversana Incorporated for the provision of methodologic advice related to the conduct of systematic reviews and meta-analysis. IW was supported by the Medical Research Council Programme MC_UU_12023/21.

Authors' contributions
CL conceived of the study; all authors contributed to the design of the study; PW and AC revised the manuscript; all authors edited the manuscript; and all authors read and approved the final manuscript.

References

1. Ioannidis, J.P., *Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses.* Cmaj, 2009. **181**(8): p. 488-493.
2. Li, L., et al., *The global research collaboration of network meta-analysis: a social network analysis.* PloS one, 2016. **11**(9).
3. Kanters, S., et al., *Use of network meta-analysis in clinical guidelines.* Bulletin of the World Health Organization, 2016. **94**(10): p. 782-784.
4. Eiring, Ø., et al., *Rapid methods including network meta-analysis to produce evidence in clinical decision support: a decision analysis.* Systematic reviews, 2018. **7**(1): p. 168.
5. Laws, A., et al., *A Comparison of National Guidelines for Network Meta-Analysis.* Value in Health, 2019. **22**(10): p. 1178-1186.
6. Egger, M., et al., *Bias in meta-analysis detected by a simple, graphical test.* Bmj, 1997. **315**(7109): p. 629-634.
7. Page, M.J., J.E. McKenzie, and J.P.T. Higgins, *Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review.* BMJ Open, 2018. **8**(3): p. e019703.
8. Page, M.J., et al., *Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions.* Cochrane Database of Systematic Reviews, 2014(10).
9. Petticrew, M., *Why certain systematic reviews reach uncertain conclusions.* Bmj, 2003. **326**(7392): p. 756-758.
10. Greco, T., et al., *The attractiveness of network meta-analysis: a comprehensive systematic and narrative review.* Heart, lung and vessels, 2015. **7**(2): p. 133.
11. Jansen, J.P. and H. Naci, *Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers.* BMC medicine, 2013. **11**(1): p. 159.
12. Li, T., et al., *Network meta-analysis-highly attractive but more methodological research is needed.* BMC medicine, 2011. **9**(1): p. 79.
13. Dechartres, A., et al., *Association between analytic strategy and estimates of treatment outcomes in meta-analyses.* Jama, 2014. **312**(6): p. 623-630.
14. Hart, B., A. Lundh, and L. Bero, *Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses.* Bmj, 2012. **344**: p. d7202.
15. Klassen, T.P., M.L. Lawson, and D. Moher, *Language of publication restrictions in systematic reviews gave different results depending on whether the intervention was conventional or complementary.* Journal of clinical epidemiology, 2005. **58**(8): p. 769-776. e2.
16. Mhaskar, R., et al., *Critical appraisal skills are essential to informed decision-making.* Indian journal of sexually transmitted diseases and AIDS, 2009. **30**(2): p. 112.
17. Whiting, P., et al., *ROBIS: Tool to assess risk of bias in systematic reviews-Guidance on how to use ROBIS.* Available at)(Accessed March 26, 2018) http://www. bristol. ac. uk/media-library/sites/social-community-medicine/robis/robisguidancedocument. pdf View in Article, 2016.
18. Canadian Institutes of Health Research. *Knowledge User Engagement*. 2016.
19. Hutton, B., et al., *The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations.* Annals of Internal Medicine, 2015. **162**(11): p. 777-84.
20. Dias, S., et al., *NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials.* 2011.
21. Jansen, J.P., et al., *Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report.* Value in Health, 2014. **17**(2): p. 157-173.
22. Dias, S., et al., *Chapter 8 Validity of network meta-analyses. In: Network meta-analysis for decision-making*. 2018: John Wiley & Sons.

23. Higgins, J.P., et al., *The Cochrane Collaboration's tool for assessing risk of bias in randomised trials.* Bmj, 2011. **343**: p. d5928.
24. Lunny, C., et al., *Toward a comprehensive evidence map of overview of systematic review methods: paper 2-risk of bias assessment; synthesis, presentation and summary of the findings; and assessment of the certainty of the evidence.* Syst Rev, 2018. **7**(1): p. 159.
25. Whiting, P., et al., *Chapter 4. Phase 2: Review of existing quality assessment tools for systematic reviews.* , in *Evidence to inform the development of ROBIS, a new tool to assess the risk of bias in systematic reviews*. 2013: Available from http://www.robis-tool.info [accessed 11/06/2020]. p. p. 22–35.
26. Shea, B.J., et al., *Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews.* BMC medical research methodology, 2007. **7**(1): p. 10.
27. Oxman, A.D. and G.H. Guyatt, *Validation of an index of the quality of review articles.* J Clin Epidemiol, 1991. **44**(11): p. 1271-8.
28. Pieper, D., et al., *Overviews of reviews often have limited rigor: a systematic review.* J Clin Epidemiol, 2012. **65**(12): p. 1267-73.
29. Shea, B.J., et al., *AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both.* bmj, 2017. **358**: p. j4008.
30. Chandler, J., et al., *Methodological standards for the conduct of new Cochrane Intervention Reviews.* Sl: Cochrane Collaboration, 2013.
31. Higgins, J.P., et al., *A tool to assess the quality of a meta-analysis.* Research Synthesis Methods, 2013. **4**(4): p. 351-366.
32. Whiting, P., et al., *ROBIS: A new tool to assess risk of bias in systematic reviews was developed.* J Clin Epidemiol, 2016. **69**: p. 225-34.
33. Guyatt, G., et al., *GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables.* Journal of clinical epidemiology, 2011. **64**(4): p. 383-394.
34. Puhan, M.A., et al., *A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis.* Bmj, 2014. **349**: p. g5630.
35. Brignardello-Petersen, R., et al., *Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis.* Journal of clinical epidemiology, 2018. **93**: p. 36-44.
36. Nikolakopoulou, A., et al., *CINeMA: An approach for assessing confidence in the results of a network meta-analysis.* PLoS Medicine / Public Library of Science, 2020. **17**(4): p. e1003082.
37. Phillippo, D.M., et al., *Threshold analysis as an alternative to GRADE for assessing confidence in guideline recommendations based on network meta-analyses.* Annals of internal medicine, 2019. **170**(8): p. 538-546.
38. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* PLoS med, 2009. **6**(7): p. e1000097.
39. Hutton, B., F. Catala-Lopez, and D. Moher, *The PRISMA statement extension for systematic reviews incorporating network meta-analysis: PRISMA-NMA.* Med Clin (Barc), 2016. **147**(6): p. 262-266.
40. Whiting, P., et al., *A proposed framework for developing quality assessment tools.* Syst Rev, 2017. **6**(1): p. 204.
41. Page, M.J., et al., *Mapping of reporting guidance for systematic reviews and meta-analyses generated a comprehensive item bank for future reporting guidelines.* J Clin Epidemiol, 2020. **118**: p. 60-68.
42. Moher, D., et al., *Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement.* Systematic reviews, 2015. **4**(1): p. 1.
43. Chambers, J.D., et al., *An assessment of the methodological quality of published network meta-analyses: a systematic review.* PloS one, 2015. **10**(4).
44. Thaler, K., et al., *Inadequate use and regulation of interventions against publication bias decreases their effectiveness: a systematic review.* J Clin Epidemiol, 2015. **68**(7): p. 792-802.
45. Zarin, W., et al., *Characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review.* BMC medicine, 2017. **15**(1): p. 3.

46. McGowan, J., et al., *PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement.* Journal of Clinical Epidemiology, 2016. **75**: p. 40-46.
47. Ban, J.K., et al., *History and publication trends in the diffusion and early uptake of indirect comparison meta-analytic methods to study drugs: animated coauthorship networks over time.* BMJ open, 2018. **8**(6): p. e019110.
48. Innovation, V.H., *Covidence systematic review software*. 2016, Veritas Health Innovation Melbourne, VIC.
49. Humphrey-Murto, S., et al., *The Use of the Delphi and Other Consensus Group Methods in Medical Education Research: A Review.* Academic Medicine, 2017. **92**(10): p. 1491-1498.
50. Moher, D., et al., *Guidance for developers of health research reporting guidelines.* PLoS medicine, 2010. **7**(2).
51. Tricco, A.C., et al., *An international survey and modified Delphi approach revealed numerous rapid review methods.* Journal of clinical epidemiology, 2016. **70**: p. 61-67.
52. Pandor, A., et al., *Delphi consensus reached to produce a decision tool for SelecTing Approaches for Rapid Reviews (STARR).* Journal of Clinical Epidemiology, 2019. **114**: p. 22-29.
53. Preston, C.C. and A.M. Colman, *Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences.* Acta psychologica, 2000. **104**(1): p. 1-15.
54. Stevens, A., et al., *Developing PRISMA-RR, a reporting guideline for rapid reviews of primary studies (Protocol)*. 2018, EQUATOR Network: Oxford, UK: https://www.equator-network.org/wp-content/uploads/2018/02/PRISMA-RR-protocol.pdf.
55. Boulkedid, R., et al., *Using and Reporting the Delphi Method for Selecting Healthcare Quality Indicators: A Systematic Review.* PLOS ONE, 2011. **6**(6): p. e20476.
56. Petropoulou, M., et al., *Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015.* J Clin Epidemiol, 2017. **82**: p. 20-28.
57. Zarin, W., et al., *Characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review.* BMC Med, 2017. **15**(1): p. 3.
58. McHugh, M.L., *Interrater reliability: the kappa statistic.* Biochemia medica: Biochemia medica, 2012. **22**(3): p. 276-282.
59. Gates, M., et al., *Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers.* Journal of Clinical Epidemiology, 2020. **125**: p. 9-15.

Appendix 1: Description of the ROBIS tool

The ROBIS tool was targeted to three specific user groups: (a) overviews of reviews authors; (b) guideline developers; and (c) review authors who may want to assess risk of bias in their review once it is complete or at the protocol stage to minimise the risk of bias. ROBIS may also be helpful for anyone who wants to assess the risk of bias in a review.

According to the ROBIS guidance document, ROBIS assesses both the risk of bias in a review and (where appropriate) the relevance of a review to the research question of interest [29]. Specifically, it addresses:
(i) the degree to which the review methods minimise the risk of bias in the pooled effect estimates and review conclusions, and
(ii) the extent to which the research question addressed by the review matches the research question being addressed by its user (e.g. an overview author or guideline developer).

Briefly, ROBIS is completed in 3 phases: (1) assess relevance (optional), (2) identify concerns with the review process, and (3) judge the overall risk of bias in the review The ROBIS tool contains four domains, namely study eligibility criteria, identification and selection of studies, data collection and study appraisal, and synthesis and findings (Figure 1) [29].

Figure 1: Structure of the ROBIS tool

ROBIS rating of items

Each domain contains signalling questions with answers being: Yes, Probably Yes, Probably No, No, or No Information. Having answered the signalling questions in each domain, reviewers can make a judgement that indicates if they have concerns about the methods relating to each domain. Responses are low, high or unclear concerns about the methods reported. In the third phase, reviewers then make an overall judgement of the risk of bias in the review, by assessing the interpretation of the findings. Specifically, whether the concerns identified were addressed; whether the relevance of studies was considered; and whether the authors avoided emphasising results based on statistical significance.

Appendix 2: Steering committee members

Chair, Carole Lunny
Jim Wright
Andrea Tricco
Penny Whiting
Julian Higgins
Sofia Dias
Ian R. White
Georgia Salanti
Areti Angeliki Veroniki
Brian Hutton

Appendix 3: Definitions

Definition of systematic review: A systematic review attempts to collate all study-specific evidence that fits pre-specified eligibility criteria in order to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimising bias, thus providing more reliable findings from which conclusions can be drawn and decisions made [22].

Definition of a systematic review with NMA: "Network meta-analysis" (NMA) is a term that encompasses both indirect and mixed treatment comparisons [19]. A systematic review with NMA simultaneously compares multiple interventions that form a connected network within the same model. The direct treatment effects of each intervention compared to with a common comparator to obtain an indirect estimate. In mixed or multiple treatment comparisons, both direct and indirect information is available to inform the effect size estimates for at least some of the comparisons; visually, this is shown by closed loops in a network plot [19]. Closed loops are not required to be present for every comparison under study.

Definition of a systematic review with pairwise meta-analysis: Pairwise meta-analysis is the statistical synthesis used in a systematic review to pool the effect estimates of primary studies comparing one treatment and a comparison treatment or control.

Definition of risk of bias
The ROBIS guidance defines bias as occurring if systematic flaws or limitations in the design or conduct of a review distort the results [23].

Definition of tools, checklists, scaled and domain based tools
A tool is defined as any structured instrument aimed at aiding the user to assess quality or susceptibility to bias [24]. To be defined as a scale, a numeric score was ascribed to each item and a summary score calculated [25]. To be defined as a checklist, it had to include multiple questions, but without the intention to ascribe a numerical score to each response or to calculate a summary score [9]. Domain-based tools are designed to assess risk of bias or quality within specific domains [9].

Appendix 4: Search strategies

Validation set of ten included studies
("31563261" or "12609941" or "26030634" or "27201949" or "24671099" or "25269948" or "23804511" or "21085712" or "29051107" or "24636374").ui

Ovid MEDLINE (R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions(R) 1946
((network meta-analys?s or NMA or ((indirect or mixed) adj3 comparison)) adj5 (tool? or instrument? or checklist? or check list? or scale? or measure? or assess? or compar*)).ab,ti.

EBM Reviews - Cochrane database of Systematic Reviews (2005 –)
network meta-analys?s or NMA or ((indirect or mixed treatment? or treatment?) adj3 comparison) AND (tool? or instrument? or checklist? or check list? or scale? or assess? or Validity or bias$ or apprais$ or quality)

The EQUATOR Network (http://www.equator-network.org/reportingguidelines/)
Study type: Systematic reviews and contl F "network"

ProQuest Dissertations & Theses Global
TI(Network Meta-Analysis) AND AB(tool)

Cochrane Comparing Multiple Interventions Methods Group (https://methods.cochrane.org/cmi/welcome)
https://methods.cochrane.org/cmi/relevant-publications-and-resources
(Cochrane Chapter on NMA; and MECIR considerations for NMA – in development)

Campbell Collaboration (https://campbellcollaboration.org/)

Joanna Briggs Institute (https://joannabriggs.org/)

EBM Reviews - Cochrane Methodology Register (includes Cochrane Colloquium abstracts) (3rd Quarter 2012) (includes Cochrane Colloquium abstracts)
((network meta-analys?s or NMA or ((indirect or mixed) adj3 comparison)) adj5 (tool? or instrument? or checklist? or check list? or scale? or measure? or assess? or compar* or valid$ or invalid or bias$ or apprais$ or quality)).ab,ti.

Scientific Resource Center Methods library of the AHRQ Effective Health Care Program
http://www.refworks.com/refworks2/?site=027181135918800000%2F57381342557464357%2FSRC+Methods+Library
network meta-analysis

International Network of Agencies for Health Technology Assessment: https://www.inahta.org/
Searched for "network meta-analysis" and "mixed treatment comparison"

Pharmaceutical Benefits Advisory Committee: https://www.pbs.gov.au/info/industry/listing/participants/pbac
Searched for "network meta-analysis" and "mixed treatment comparison"

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen: https://www.iqwig.de/en/home.2724.html
Searched for "network meta-analysis" and "mixed treatment comparison"

European Network for Health Technology Assessment: https://eunethta.eu/methodology-guidelines/

Searched for “network meta-analysis” and “mixed treatment comparison”

Guidelines International Network: https://g-i-n.net/home
Searched for “network meta-analysis”, and “mixed treatment comparison”

International Society for Pharmacoeconomics and Outcomes Research: https://www.ispor.org/
Searched for “network meta-analysis” n=1037, and “mixed treatment comparison”
and https://tools.ispor.org/peguidelines/

National Institute for Health and Care Excellence Decision Support Unit: http://nicedsu.org.uk/multivariate-meta-analysis-tsd/
Searched for “network meta-analysis” n=1, and “mixed treatment comparison”

Canadian Agency for Drugs and Technologies in Health: https://www.cadth.ca/
Searched study type “reports”, then for “network meta-analysis” n=900, and “mixed treatment comparison”

Appendix 4: Introduction to the survey and participant consent form

Introduction to the survey

Thank you for agreeing to take part in this study. Your input will help us to develop a domain based risk of bias tool to assess risk of bias in reviews with network meta-analysis (NMA). We want to identify what items should be included in this tool. To generate a potential list of items for inclusion, we conducted a methodological review of items related to quality and risk of bias, which is available as a protocol and a full manuscript preprint prior to full publication (https://XXXX).

We want to get the views of a wide range of reviewers and users of reviews with NMA on what they believe is most important items to include in the tool. We will use this information to further develop and refine the tool. The data gathered within this survey will not be not be subject to any public disclosure.

This study will involve two to three survey rounds during November and December 2020 to ask your opinion about the most important items to be included in the tool. Each survey will take approximately XX minutes. You will be asked to give your opinion on the importance of XX items. You will also be given the opportunity to edit, omit, split or group items, suggest new items, or make wording changes. We know we will not get consensus on every item and are interested in the written comments you make so we can understand more about why people have different views. Please feel free to provide feedback in the comment box provided. You will be asked to rate the importance of each item presented. If you feel unable to give a rating, please select "unable to score". There are no right or wrong responses. We are only interested in your opinion.

All participants who complete the survey will receive a copy of the results. Once we have received responses from all participants, we will collate and summarize the findings and formulate a brief second questionnaire. You will receive this early next month. The identity of all participants will remain confidential at all times.

If you have any questions, please contact Dr Carole Lunny at carole.lunny@XXX

Questionnaire

Statement of Consent

I have read the foregoing information and I consent voluntarily to be a participant in this study. Your information will be used only to identify participants for round 2 and 3 of the survey

Name of Participant	
Email address	

○ Yes, I consent
○ No, I do not consent (will terminate the survey)

----------Begin Survey----------

Appendix 5
Data extraction Form template