European Journal of Epidemiology (2020) 35:49–60
https://doi.org/10.1007/s10654-019-00576-5

GUIDELINES

A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research

Taulant Muka[1] · Marija Glisic[1,2] · Jelena Milic[3,4] · Sanne Verhoog[1] · Julia Bohlius[1] · Wichor Bramer[5] · Rajiv Chowdhury[6] · Oscar H. Franco[1]

Received: 21 June 2019 / Accepted: 29 October 2019 / Published online: 13 November 2019

Abstract

To inform evidence-based practice in health care, guidelines and policies require accurate identification, collation, and integration of all available evidence in a comprehensive, meaningful, and time-efficient manner. Approaches to evidence synthesis such as carefully conducted systematic reviews and meta-analyses are essential tools to summarize specific topics. Unfortunately, not all systematic reviews are truly systematic, and their quality can vary substantially. Since well-conducted evidence synthesis typically involves a complex set of steps, we believe formulating a cohesive, step-by-step guide on how to conduct a systemic review and meta-analysis is essential. While most of the guidelines on systematic reviews focus on how to report or appraise systematic reviews, they lack guidance on how to synthesize evidence efficiently. To facilitate the design and development of evidence syntheses, we provide a clear and concise, 24-step guide on how to perform a systematic review and meta-analysis of observational studies and clinical trials. We describe each step, illustrate it with concrete examples, and provide relevant references for further guidance. The 24-step guide (1) simplifies the methodology of conducting a systematic review, (2) provides healthcare professionals and researchers with methodologically sound tools for conducting systematic reviews and meta-analyses, and (3) it can enhance the quality of existing evidence synthesis efforts. This guide will help its readers to better understand the complexity of the process, appraise the quality of published systematic reviews, and better comprehend (and use) evidence from medical literature.

Keywords 24 Steps · Systematic review · Meta-analysis · Guideline · Evidence synthesis

Introduction

The practice of evidence-based medicine requires up-to-date syntheses of existing evidence. Systematic reviews and meta-analyses ought to be rigorous and transparent, and provide empirically derived answers to focused research questions. The publication of systematic reviews and meta-analyses has grown exponentially in recent decades [1], and they have gradually migrated to the top of the pyramid of what is considered good evidence. Nevertheless, systematic reviews and meta-analyses can be large, challenging endeavours that are sensitive to bias and errors. To provide accurate answers and limit potential pitfalls they require careful preparation and organisation. Several organised efforts such as the Cochrane collaboration (founded in 1993) have attempted to regulate and improve the quality and uniformity of systematic reviews [2]. A few guidelines and textbooks also have been published that offer comprehensive descriptions of the methodology [2–4]. However, several studies assessing the quality of published systematic reviews have shown that not all systematic reviews are truly systematic and that their quality is highly variable [5, 6].

Experienced researchers as well as those who are learning the methodology can use better guidance and training in how to conduct a systematic review and meta-analysis

Taulant Muka and Marija Glisic have contributed equally to this work.

Wichor Bramer and Rajiv Chowdhury have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10654-019-00576-5) contains supplementary material, which is available to authorized users.

✉ Taulant Muka
taulant.muka@ispm.unibe.ch

Extended author information available on the last page of the article

[1]. Here we provide a concise, 24-step guide on how to perform a systematic review and meta-analysis.

Aim and scope

We present a concise and comprehensive practical guide and a checklist with 24 steps that can help biomedical researchers conduct a systematic review and meta-analysis (Fig. 1). This guide (1) simplifies the methodology of a systematic review, (2) provides tools to conduct methodologically sound systematic reviews and meta-analyses, and (3) it can enhance the quality of existing evidence synthesis efforts. This guide can be used by anyone planning a systematic review, whether one that is solely narrative or includes a quantitative element; however, health professionals and researchers who are familiar with basic methods of research and are able to interpret basic statistical principles used in health research may benefit most from it. The supplemental material accompanying this article provides more detailed information and examples for less experienced researchers.

Step-by-step guide: the 24 steps

Step 1: Define research question

To define a research question, first establish in detail the primary and secondary aims of the study (including potential effect modifiers). The more clearly a research question focuses on, and clearly defines the science and summarizes the aim of the research project, the more it will facilitate building the search strategy, whether focused or broad, and conducting the systematic review. Developing a good research question and defining the aim of the study requires scanning the literature to identify gaps in the field. The existence of systematic reviews on similar research questions is not an obstacle to another systematic review if new analysis will close gaps and add value. Some research fields also develop rapidly; if new publications appear frequently, new and more current systematic review of the evidence or maintaining living network meta-analyses using automatized approaches may be indicated [7]. A valuable research question necessarily emerges from existing knowledge, and there are tools that may facilitate the definition and analysis of the research question. These include *PICO(S)*, used in

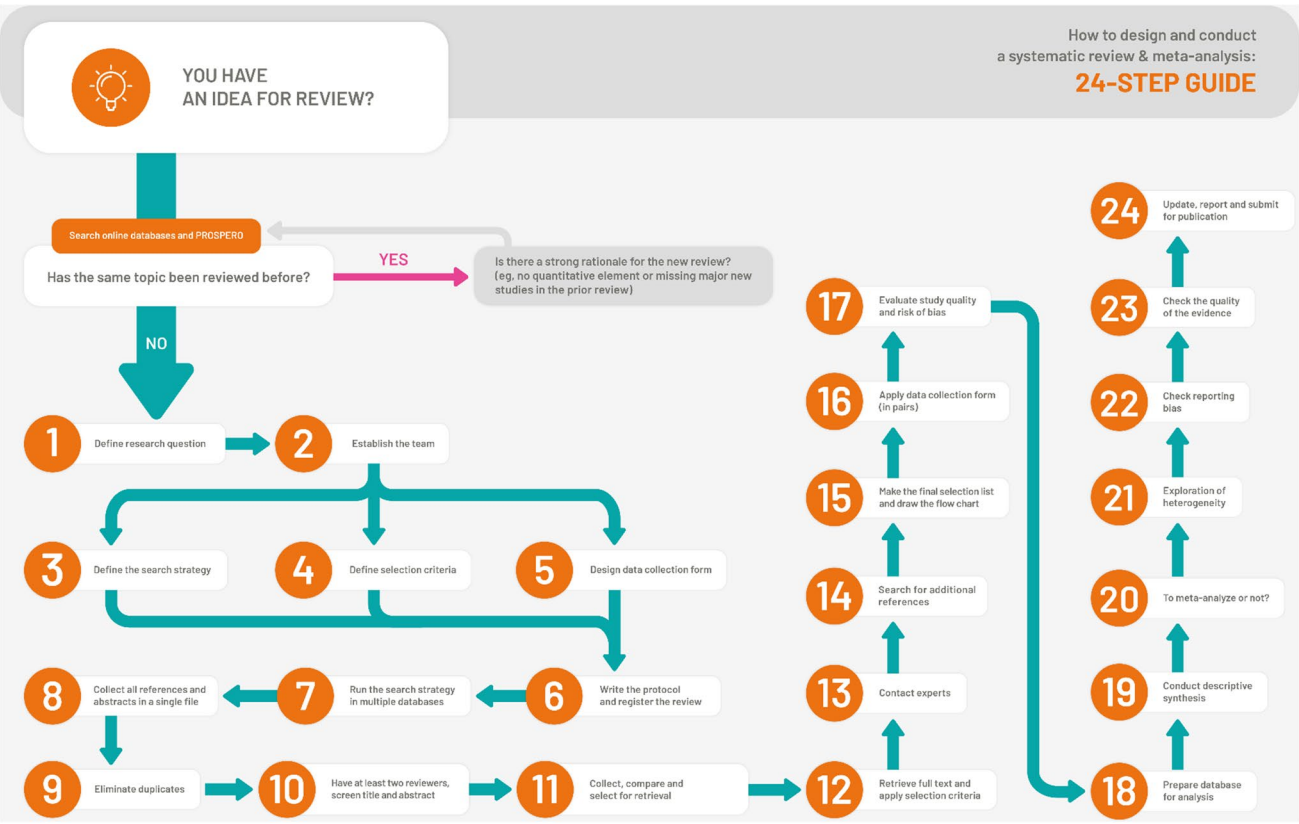

Fig. 1 24-STEP GUIDE checklist: steps to be followed to successfully design and conduct a systematic review and meta-analysis

evidence-based clinical practice [8]; *PEO* [9] and *SPICE* [10] for qualitative research questions; and *SPIDER* [11] in mixed-methods research. In **w1** we discuss in more detail how to define a research question.

Step 2: Establish the team

A well-organized and coordinated team is necessary. Many steps such as the literature search, revision process, and quality assessment require double-checking by independent reviewers, and a third independent reviewer is often needed to resolve disagreements that may arise during the study inclusion process. Choose carefully colleagues and experts that you are planning to collaborate with; you should evaluate their competence in the field and integrity [12]. The team should have members whose expertise spans searching for studies (i.e. a librarian or medical information specialist), understanding primary study methods and systematic review methods, synthesizing findings and performing meta-analysis, and knowing the area under investigation. The complexity of the question being addressed and the expected number of references also will figure in the size of the team. The makeup of the team will be established after the final search since the number of hits obtained will determine the participation of independent reviewers. Expertise should be balanced across the team members so that one group of experts is not overly influential. For example, review teams that are too dominated by clinical content experts are more likely to hold preconceived opinions related to the topic of the systematic review, spend less time conducting the review, and produce lower quality reviews [13]. Finally, a team cannot function without a team leader. The leader is not by definition a professor or the most senior member of the team. The leader coordinates the project, takes care that study protocol is followed, keeps all team members informed, and facilitates their participation in all phases of the project.

Step 3: Define the search strategy (steps 3, 4, and 5 are done in parallel)

A comprehensive search forms the foundation of any systematic review and consists of writing specific search strategies in different online databases to retrieve eligible studies. Inadequate searches or errors in search strategies may miss evidence, while untargeted, broad searches lead to superfluous articles and waste time. Missing relevant articles may bias estimates. Numerous online databases can be searched. It is not necessary to search all databases, however, no single database can encompass all medical literature. Recently, our group has shown that optimal searches should be performed by using Embase, MEDLINE, Web of Science, and Google Scholar at a minimum to ensure adequate, efficient coverage [14]. In the search it is not necessary however to retrieve the full 1000 references (the maximum number of references that is possible to download from Google Scholar) from Google Scholar, but mostly only the first 200 references have to be added from Google Scholar [14]. In case the work includes synthesis of evidence from clinical trials, Cochrane library is recommend to be searched for relevant references, although our work has shown that all included references would have been found had we not searched the Cochrane library [14]. PsycINFO and CINAHL databases should be searched if the research question is related to the field of psychiatry, psychology and/or to nursing and allied health. Research has also shown that in CINAHL, the indexing of qualitative research is better than that in Pubmed, therefore it is recommended for the search of those study types as well. Central to search quality and reproducibility is the inclusion of a librarian or search specialist [15]. Our group has established a method that describes in detail a 15-step process to develop a systematic search strategy [16]. In appendix **w2** we offer a few basic recommendations for searching databases.

Example See **w3** for an example of the search strategy for a recently published meta-analysis of intervention studies, which evaluated the association between phytoestrogen supplementation followed with regular, normocaloric diet and glucose homeostasis, and risk of type 2 diabetes in adult women. A literature search was done using five electronic databases: MEDLINE via Ovid, Embase.com, Web of Science Core Collection, Cochrane CENTRAL via Wiley, and Google Scholar. This example reflects the complexity of a search strategy and presents search syntaxes for the different medical databases [16].

Step 4: Define selection criteria (inclusion/exclusion)

Selection criteria identify relevant evidence during the screening process. The selection criteria guide the reviewers, save time, minimize mistakes, and guarantee transparency and reproducibility. They depend on the research question and incorporate study characteristics that can include study design, date of publication, and geographical location; characteristics of the study population such as age, sex, and presence of disease; characteristics of the exposure and outcome measured; and characteristics of the methods used such as type of analysis, adjustment for confounders, measure of association reported, etc. An important step in establishing the selection criteria is the evaluation of the type(s) of study design that may best answer the research question. In addition to looking for study designs that may yield the highest level of evidence it is important to think about which study designs fit the research question. After inclusion/exclusion criteria are established a so-called *checklist* should be written. A checklist guides the reviewers through the screening

process and a well written checklist will save time and minimize mistakes during screening. A sample checklist can be found in **w4**.

Example Meta-analyses that include intervention studies often pool estimates from randomized controlled trials (RCTs) that use substantially different control groups. For example, a control arm may receive placebo or a control substance, or the same treatment as the intervention arm but at a lower dose. In a recent systematic review and meta-analysis of RCTs we evaluated the association of plant-based therapies with menopausal symptoms. To maintain consistency and because of the difficulty of interpreting results without a placebo or control, we excluded head-to-head trials that compared nonhormonal therapies with estrogen or other medications, that lacked a placebo group [17].

Step 5: Design data collection form

A key step in a systematic review is the extraction of pertinent data from primary studies (and not from individual subjects) using a standardized data extraction form. Data are collected on (1) general characteristics of the study such as investigator name(s), year of the study, and funding source, (2) characteristics of the study population that may include age, sex, and ethnicity, (3) exposure or intervention, which can include assessment method, distribution in the study population, and dosage when describing drugs, (4) outcomes, (5) methods such as the type of statistical analysis that was used and factors adjusted for, and (6) results such as measures of association, stratified analyses, and distribution of results agreement. Designing the data extraction form requires careful consideration of the research question and often benefits from piloting the form in at least five studies in the field before finalization. A variety of software applications allow organization of the data extraction form, including Microsoft Access/Excel, Qualtrics, REDCap, Google Forms/Sheets, SRDR (Systematic Review Data Repository; https://srdr.ahrq.gov/home/index) etc. [18]. Further, software that could be used for reference screening such as Covidence (https://www.covidence.org) and DistillerSR (www.distillercer.com) can be also used as data extraction tools and In **w5** an example of a data extraction form is provided.

Step 6: Write the study protocol and register the review

The study protocol contains the research question, primary and secondary aims, study design, inclusion and exclusion criteria, electronic search strategy, and the analysis plan described in detail. The study protocol guides the reviewers through the screening process. When writing the protocol, relevant experts should be asked to provide feedback and make sure the protocol covers all elements. We provide an example of a study protocol in **w6**. Registering the review is recommended to avoid overlap and superfluous efforts, and to provide transparency. There are a few platforms for registration; those most often used are *Prospero* (http://www.crd.york.ac.uk/PROSPERO/) for reviews in health or social care, and *Cochrane* (http://www.cochrane.org/cochrane-reviews) for reviews regarding interventions. Instructions how to register a review at Prospero are given in **w7**.

Step 7: Run the search strategy in multiple databases

As mentioned in step 3, a literature search should include at least four online databases: Embase, MEDLINE, Web of Science, and Google Scholar. Each database has its own way of writing a search strategy.

Step 8: Collect all references and abstracts in a single file

Collect all of the research results of step 7 in EndNote or another tool such as Covidence (www.covidence.org), DistillerSR (www.distillercer.com), or Rayyan (rayyan.qcri.orgrayyan.qcri.org). If available, we recommend using EndNote, which provides support for step 9. To import the selected references from a database into an Endnote file, export the references from that database in a format recognized by EndNote. The instructions on exporting citations from the major databases into EndNote are provided in **w8**.

Step 9: Eliminate duplicates

Retrieving relevant studies from various databases generally leads to articles being identified multiple times. Removal of duplicate records will reduce the reviewers' workload when screening titles and abstracts. Due to the heterogeneous nature of articles in databases, de-duplication can be cumbersome and time-consuming. Our group has published a method using EndNote for faster yet accurate de-duplication [19]. There are other software available for deduplication, but they have not been thoroughly evaluated for accuracy.

Example Detailed instructions on how to perform de-duplication using EndNote can be found elsewhere [19].

Step 10: Have at least two reviewers screen title and abstract

The titles and/or abstracts of each reference should be screened for relevance by at least two reviewers. It is not necessary that any single person screen all references as long as each reference has been screened by two independent reviewers. For example, one person might screen all of the references, while for the second screening all of the references may be divided across other reviewers. Titles

and abstracts can be screened simultaneously, judging the relevance of the abstract if the title is found to be relevant. It is not necessary to screen title first and after that the abstract as was done in the past when screening was done on paper. In this phase, references are selected based on the selection criteria applied to the title and abstract, and not to the full text of the article. If a reference lacks an abstract, and has only the title, the reference should be included for the next step. At this phase, it is not necessary to keep track of the reason for exclusion [19]. Various software applications such as Rayyan, Covidence, and DistillerSR are available for the title and abstract screening phase [20–22]. We do not recommend the use of Excel for this purpose because it is complicated and time-consuming. Our group has developed a method for screening title and abstract using EndNote. The method is very fast with a median of 300 references screened per hour [19].

Screening of titles and abstracts can be performed in other software as well. Rayyan, Covidence distiller sr are tools that offer these services. Rayyan uses artificial intelligence to determine the highest potential references among those yet to be screened. If this can be trusted, it will reduce the time needed to screen references. However, until now it is not yet trusted, meaning that despite the relevance ranking all reference normally will have to be screened. In our experience most tools require actions for each reference to be included. Our method in Endnote allows multiple references to be excluded at once, which greatly reduces the time needed for screening.

Step 11: Collect, compare, and select for retrieval

The references screened by two independent reviewers are then collected and compared. The software tools mentioned in step 10 all have a comparison feature, and the method described in our previous research describes how this can be performed in EndNote [19]. The overlapping set of references that both reviewers have selected to include in the review are considered for the next step: retrieval of the full text. For the non-overlapping references on which the reviewers did not agree, a meeting between the two reviewers should be organized to reach a common final decision. A third independent reviewer should be assigned in cases of persisting disagreement (Fig. 2); or, the two reviewers could decide to consider each reference included by at least one of the reviewers in the next step. Usually the third reviewer should be a senior researcher with experience in the topic.

Step 12: Retrieve full text and apply selection criteria

The full texts of the references selected based on titles and abstracts are retrieved. Full texts can usually be found using the "find full text option" in EndNote, via searching local libraries or online search engines such as Google Scholar and Research Gate, or contacting the authors directly. If the full text of a reference is not directly available to the reviewers the reference should not be ignored, but instead the university library can assist in retrieving a copy of the article via interlibrary loan. Once all full texts are retrieved, two independent reviewers screen the articles using the

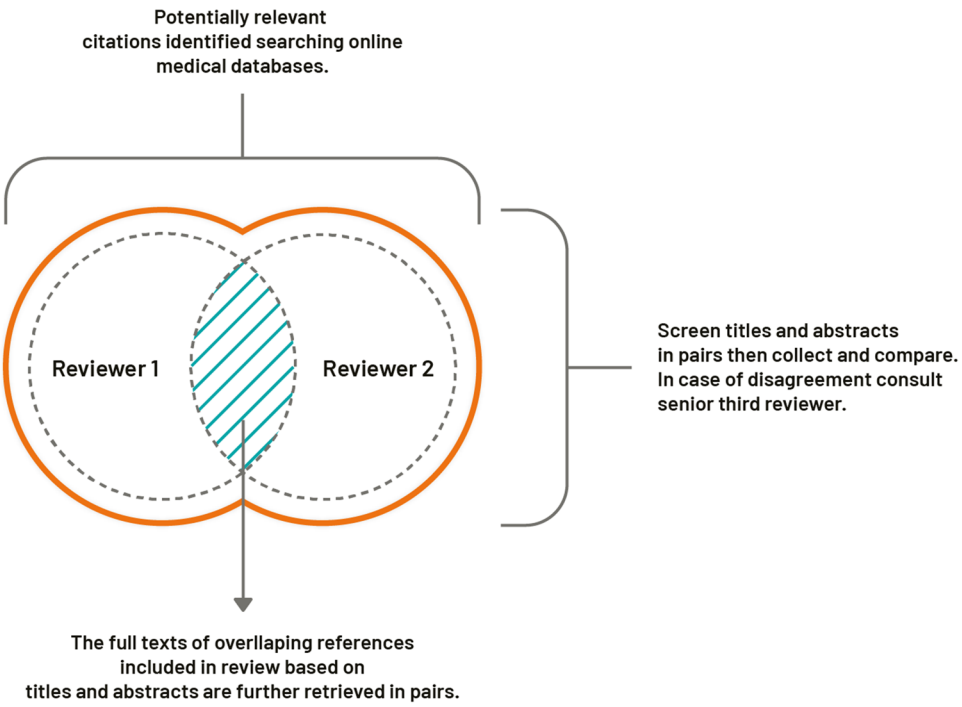

Fig. 2 Reference screening process

selection criteria to select those to be included in the systematic review. A third independent reviewer is available to solve disagreements. As in abstract screening, custom groups in Endnote can distinguish various reasons for exclusion, and articles may be assigned to specific groups for certain sub-questions.

Step 13: Contact experts

Contact authors who are experts in the field to identify any ongoing or missing study, find unpublished but relevant data (for example, estimates might be provided in an article for a certain outcome but not given for another), or assist recalculation of summary estimates from a published study with a standardized set of covariates for more meaningful combination of the estimates across all studies in meta-analysis. To identify experts, we recommend corresponding with authors of articles selected in step 12. Check whether the references suggested by these authors duplicate references already reviewed and, if they do not, repeat steps 9–13 to make a final decision on whether to include the suggested studies in the review.

Step 14: Search for additional references

In this step, the references assembled at the end of step 12 and through step 13 should themselves be reviewed for relevant studies cited in them (forward search) or by screening studies that have cited the articles (backward search). The abstract and citation database Elsevier Scopus, may facilitate this. Detailed instructions can be found at: https://www.elsevier.com/__data/assets/pdf_file/0005/79196/scopus-quick-reference-guide.pdf.

The reference lists of previous systematic reviews that are related to this systematic review should also be searched. For references selected from these sources, EndNote may be used in combination with the Scopus or Web of Science databases to semi-automatically download the references into an existing EndNote library [19]. This again requires another round of checking references and eliminating duplicates, and retrieving and screening full-texts in keeping with steps 8 through 12.

Step 15: Make the final selection list and draw the flow chart

The articles selected in steps 12, 13, and 14 will become the final articles included in the review. A well-designed flowchart will contain information on the number of relevant citations identified through database searches, experts, and reference lists; the number of studies excluded based on title and abstract search; the number of full texts screened; the number of studies excluded after full text assessment with the reasons for exclusion citing number of studies excluded for each reason; and the number of the studies included in the systematic review. An example of a flowchart can be found in online **w9**.

Step 16: Apply data collection form (in pairs)

The next step is to extract the data contained in the included studies using the predefined collection form of step 5. Using the items in the form, two independent reviewers extract detailed data from each article. Close attention to the data extraction process will yield an initial understanding and description of the shared characteristics of the body of evidence and will pave the way for the analytic and interpretive process of synthesis to follow. When extracting the data, make clear abbreviations, carefully convert all data using the same unit(s), maintain consistent definitions, and keep content brief.

Step 17: Evaluate study quality and risk of bias

Evidence and results should be interpreted in light of the quality of the included studies. The quality of the research encompasses how a study has been conducted (its methodological quality) and how it has been described (reporting quality and reproducibility). Poor methodological and reporting quality of primary studies included in the review may introduce bias and spurious conclusions. Thus, a valid assessment of study quality by two independent reviewers is essential to guarantee accuracy and generalisability. An important aspect of methodological quality is the risk of bias in the included studies. While its researchers might have done the best possible study, the study may still be at high risk of confounding, selection bias, and information bias. There are a number of checklists available to assist in assessing risk of bias in the included studies. The Cochrane Collaboration's tool RoB 2 [23] and the Newcastle–Ottawa Scale [24] are the most commonly used to evaluate the risk of bias in RCTs and in prospective observational studies, respectively. The RoB 2 tool (online Supplement I) evaluates five possible sources of bias [1]: bias arising from the randomization process [2], bias due to deviations from intended interventions [3], bias due to missing outcome data [4], bias in measurement of the outcome, and [5] bias in selection of the reported result. The Newcastle–Ottawa Scale (online Supplement II) uses a star system (with maximum of nine stars) to evaluate three domains: selection of participants, comparability of study groups, and the ascertainment of outcomes and exposures of interest. Studies that receive a score of nine stars are judged to be at low risk of bias, a score of seven or eight stars indicates a medium risk, while a score of six or less indicates a high risk of bias. Separate tools have been developed to assess risk of bias in nonrandomised

studies of interventions (ROBINS-I) [25] and in diagnostic accuracy studies, such as QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) [26]. Further, the QUIPS (Quality In Prognosis Studies) tool has been developed to assess risk of bias in predictor finding (prognostic factor) studies [27]. While recently, the Prediction model study Risk Of Bias Assessment Tool (PROBAST) has been developed to assess the quality of prediction model studies for development, validation, or updating of both diagnostic and prognostic models, regardless of the medical domain, type of outcome, predictors, or statistical technique used [28].

The quality of evidence of the included studies in the systematic review/meta-analysis should be clearly reported, discussed, and interpreted to provide readers an idea on how much confidence they can place in the conclusions provided. An example of how to evaluate the quality of RCTs can be found in **w10**.

Step 18: Prepare database for analysis

In this step, data from the newly formed database are collated, imported to a statistical evaluation program (SPSS or Excel), and prepared for analysis. Analyses may be descriptive (step 19), such as a structured summary and discussion of the studies' characteristics, findings, and quality, or they may be quantitative (step 20), which involves statistical analyses (step 21). The data quality depends on the studies that are included in the review. In online table **w11** we give an overview of effect measures by type of study and suggested random and fixed-effect models which shall be used to pool the data together. For example, in observational studies prevalence, mean difference values, beta regression coefficients, odds ratios, relative risks, or hazard ratios may be reported. While clinical trials, when the outcome is dichotomous, may report baseline and end of the trial values of outcomes, and the mean harmonized across included studies. For example, some studies may report risk estimates by comparing various extreme quantiles (top vs. bottom fifths, thirds, etc.), per unit or per standard deviation change in baseline exposure, etc. Also, outcomes may be measured on different scales: for blood glucose, mmol/L is the most common measurement used in the UK while mg/dL is predominantly used in the USA and continental Europe. Systematic reviews of health economics outcomes might be presented in different currencies from different years, and require currency conversions and adjustments for inflation [29]. There are online tools that can help with unit and currency conversion.

Step 19: Conduct descriptive synthesis

Descriptive synthesis relies primarily on words and text to summarize and explain findings. Whether including a meta-analysis or not, authors should describe the process flow of the systematic review by summarizing the number of references they found from the search strategy, the number of abstracts and full texts they screened, and the final number of primary studies they included in the review. This process is summarized in a flowchart flow-chart (**w9**). Authors should also describe the characteristics such as the populations studied, types of exposures, intervention details, and outcomes of the included studies in a table and in the main text of the manuscript. If meta-analysis is not feasible (see step 20), authors should describe the results of the included studies, including the direction and size of effect, effect consistency across studies, and the strength of evidence for the effect. Rodgers et al. offer further guidance [30]. An example of a table containing study characteristics is presented in **w12**.

Step 20: To meta-analyze or not

Prior to this step, systematic reviews and meta-analyses share the same steps. At step 20, the study team must decide whether the data gathered for each outcome is suitable for pooling using quantitative methods. By combining data from different studies, the sample size increases generating more statistical power and improving estimates of the size of the effect, and it has the potential of resolving uncertainty when primary studies disagree. Also, when possible, a meta-analysis makes it easier to describe the pooled effect of the findings (instead of describing the findings of each study separately). The decision to pool depends on the degree of heterogeneity. Heterogeneity is not a condition to ignore, but to report, and it can occur at multiple levels: study characteristics such as differences in design (interventional or observational); population characteristics including differences in age, gender, and geographical location; and methods and results encompassing differences in analyses, adjustments, and measures of association. While clinical, biological or methodological heterogeneity may be specific to certain topic, statistical heterogeneity can be examined using the same statistical methods across all meta-analyses [31]. The most commonly used methods to evaluate statistical heterogeneity include the Cochrane's Chi squared test (Cochran's Q), which examines the null hypothesis that all studies are evaluating the same effect but may not always accurately detect heterogeneity. Higgins's I^2 statistic is also widely used. Higgins's I^2 represents the percentage of variation between the sample estimates that is due to heterogeneity rather than to sampling error (tells us what proportion of the total variation across studies is beyond chance) [31]. It can take on values from 0 to 100%, with 100% being the maximum level of heterogeneity. Often I^2 values below 25% are considered low, 25 to 50% moderate, and above 75% high heterogeneity [32]. I^2 is routinely implemented in all Cochrane reviews and in meta-analyses published in medical journals. However, I^2 has some uncertainty, and Higgins and

Thompson provided methods to calculate this uncertainty, while, recently the other investigators indicated that I^2 has low statistical power with small numbers of studies and its 95% confidence intervals can be large [33] and given that I^2 is not precise, 95% confidence intervals should always be given [31]. For example, in STATA, it is possible to calculate 95% CI using either of two methods: a test based approach or a non-central χ^2 based approach (heterogi module). The performance of these two methods is comparable, although the test based approach often gives lower values for lower and upper confidence intervals, so that the non-central χ^2 based approach may be preferable [34]. The perception of statistical heterogeneity may influence researcher's decision on whether the data are similar enough to combine different studies. Therefore, when making a decision on whether or not to pool treatment estimates in a meta-analysis, Ruecker et al., suggest that the between-study variance ($\tau 2$), rather than I^2 may be appropriate measure for this purpose [35].

Stratification is a tool to explore sources of heterogeneity (see step 21). It is important that the studies in the meta-analysis are comparable in terms of definitions, coding, methods, comparisons, and categories of exposure between studies. Therefore, before synthesizing estimates it is crucial to use the same estimates and standardize the coding and definitions when possible. More details on which potential problems and how to deal with them while standardising the data for meta-analysis can be found in **w13**.

Another doubt a researcher may face is whether or not to synthase different types of studies which address the same research question (i.e. observational and RCTs or observational and experimental studies). The inclusion of more than one study design may improve the quality of systematic review significantly and contribute to better understanding, easier interpretation of findings and clarification of the contradictory results. Another important uncertainty when studding health interventions is whether or not to include RCTs only or also non-randomized. Although, RCTs are considered to be on the top of quality of evidence pyramid, Ioannidis et al. [36] reported that discrepancies between RCT and non-randomized studies were less common when only nonrandomized studies with a prospective design were considered. Also, the Cochrane Collaboration offers a guide for inclusion of nonrandomized studies [37] and has developed a tool for assessing the risk of bias in both RCT and controlled nonrandomized studies [38]. Therefore, it is of high interest not to neglect nonrandomized studies especially in cases where randomization may pose important ethical issues. Further, in clinical practice there are more than two interventions of interest for a single health condition and researchers often aim to determine the best available intervention in a single, coherent analysis of all the relevant RCTs [7, 39, 40]. A pairwise meta-analysis and its extension network meta-analysis (NMA) have been developed to facilitate indirect comparisons of multiple interventions that have not been studied in head-to-head studies. Network meta-analysis as compared to pairwise meta-analysis, allows the visualisation of a greater number of evidence, estimation of the relative effectiveness among all interventions, and ranks ordering of the intervention [39]. The underlying assumption of NMA is that there are no study or individual's characteristics that would modify the relative treatment effect of each treatment in comparison with other treatments included in the meta-analysis [40]. NMA can be performed for continuous and dichotomous RCTs outcomes but also for event rates and from survival models, using an appropriate scale (mean difference, odds ratio, hazard ratio, relative risk). Detailed instructions how to perform NMA can be found elsewhere [40].

Example We have studied the associations between phytoestrogen intake and type 2 diabetes (T2D) risk and glucose homeostasis. We included observational longitudinal studies and RCTs. Although, the estimates reported in observational studies (risk of developing T2D) and RCTs (mean serum change) could not be pooled together, they are complementary. In particular, the findings of beneficial effect of phytoestrogens on T2D risk were supported with findings from RCTs where we found that phytoestrogen supplementation improved glucose homeostasis [16]. Therefore, the conclusions of our review were stronger than if would have included only observational studies or solely RCTs. Therefore, the conclusions of our review were stronger than if would have included only observational studies or solely RCTs. Similarly, in an another systematic review and meta-analysis of alcohol intake and onset of menopause, the meta-analysis results of cross-sectional and longitudinal studies provided similar conclusions, strengthening the validity of the findings had we choose only cross-sectional studies [41].

When synthesising the evidence, authors often need to choose between two statistical methods: the fixed effect (FE) and the random effects (RE) model. These two methods may yield similar or discrepant results. However, even if results of the two models are similar summary estimates should be interpreted in a different way [42]. The basic assumption of the FE model is that the exposure or treatment effect under observation is fixed in all studies included in the meta-analysis, whereas the RE model allows the exposure effect to vary across the studies. In simple terms, RE model allows the true effects underlying the studies to differ and thus accounts for unexplained heterogeneity between studies. The main misconception is that the model should be chosen based on the test of heterogeneity. Indeed, often when heterogeneity variance is estimated to be 0% the results are identical under the two meta-analysis models. However, the choice of the model should not be made based on the test of heterogeneity since heterogeneity may exist even if it remains undetected by the test. In Fig. 3 we

compare the FE and RE models and give instructions on how to choose a model suitable for your analysis [42].

Finally, various software applications are able to perform a meta-analysis. However, one of the most commonly used and for the inexperienced researcher perhaps one of the simplest for meta-analysis is *metan* command in STATA [43]. Guidance on how to undertake meta-analysis using Stata is provided by Chaimani et al. [44] In case authors do not have a STATA subscription, meta-analysis packages are available in the open access statistical environment R (Metafor (R package)) [31]. For users who are not experiences with using R, we suggest JASP or Jamovi which are free, open-source programs used to perform statistical analysis tests by using R packages. Further, Review Manager (RevMan) developed by the Cochrane Collaboration may be a good choice for those who are new to the world of meta-analysis. Nevertheless, in order to perform a simple meta-analysis it is possible to use Excel add-on such as MetaEasy or MetaXL [34].

Step 21: Exploration of heterogeneity

Subgroup analyses, or stratification, should be taken into consideration from step 1 within the definition of primary and secondary aims. Factors by which results might differ—that is, effect modifiers—often include study characteristics such as study design, geographical location, date of publication, and type of intervention, and also population characteristics such as age, gender, ethnicity, and presence of disease [45]. Results should be presented and pooled by different categories of these factors to compare whether the pooled estimates differ within groups, and whether tau^2 changes. Heterogeneity should also be evaluated within specific strata. Meta-regression analysis can also be used to explore whether observed heterogeneity is a consequence of the specific study or population characteristics. Meta-regression is therefore similar to conventional statistical regression used to determine the effect of one factor upon an outcome variable. Meta-regression is often done when

References used to create the figure:

1.Meta-analyses of randomised controlled trials. Davey Smith G, Egger M Lancet. 1997 Oct 18; 350(9085):1182.
2.Random-effects model for meta-analysis of clinical trials: an update. DerSimonian R, Kacker R Contemp Clin Trials. 2007 Feb; 28(2):105-14.
3.Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychol Methods. 1998;3:486–504.
4.Nikolakopoulou A, Mavridis D, Furukawa TA, Cipriani A, Tricco AC, Straus SE, et al. Living network meta-analysis compared with pairwise meta-analysis in comparative effectiveness research: empirical study. BMJ (Clinical research ed). 2018;360:k585.

Fig. 3 Fixed versus random effects model

more than 10 studies are included in a meta-analysis [46]. An example of subgroup analyses is in **w14**. Recommendations for the interpretation of subgroup analyses in systematic reviews can be found elsewhere [47].

Step 22: Check reporting bias

Publication bias occurs whenever the published literature is systematically unrepresentative of all completed studies [48]. Publication bias originates in a decision to publish that is influenced by an experimental or research study's outcome. Most commonly, negative results or those judged not significant are less likely to be submitted and accepted for publication. Publication bias is usually evaluated through a funnel plot in which asymmetry may be assessed visually, and by using the Egger test. A funnel plot is a scatter plot of the exposure effect estimates from individual studies against a measure of study precision (typically the standard error) [49]. If the funnel is asymmetric, it can imply that there are studies missing from the literature. However, publication bias is not the only cause of funnel plot asymmetry; other causes of bias include heterogeneity, selective outcome reporting, and simply chance [49]. Particularly when representing a low number of studies a funnel plot may not detect publication bias [50, 51]. Harbord developed a modified version of the Egger test for small-study effects in meta-analysis of controlled trials with binary endpoints [52]. Yet, this test is not recommended in meta-analyses of cohort studies where there is large imbalance in the group sizes; however, in this situation the original Egger test will often perform well. Further, Begg proposed a bias indicator using Kendall's method (testing the interdependence of variance and effect size). This bias indicator makes fewer assumptions than that of Egger and in case of small number of studies, bias cannot be ruled out if the test is not significant. Yet, this test may be used as an exploratory tool for meta-analysis, as a formal procedure to complement the funnel-plot graph [53]. When the degree of between-trial heterogeneity is large, none of the three mentioned tests has uniformly good properties [52]. Finally, the presence of publication bias requires reporting and thorough discussion, but it need not prevent publishing the study. More information can be found in **w15**.

Step 23: Check the quality of the evidence: the confidence in the results presented

The strength of the results reported in a systematic review and meta-analysis relies, first, upon the quality of the review's evidence. Authors can apply the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to score the quality of evidence included in the systematic review. The GRADE approach bases judgment of the quality of evidence on the magnitude of effect, and consideration of the risk of bias, the study design, and consistency and directness of the findings. It grades evidence as high, moderate, low, or very low. RCTs start as high quality and observational studies start as low quality. Limitations in study quality, important inconsistency of results, or uncertainty about the directness of the evidence can lower the grade of evidence. Also, certain factors such as evidence of a dose–response gradient or strong evidence of association based on consistent evidence from two or more observational studies with no plausible confounders may increase the grade [54]. The evaluation should be performed independently by two reviewers, while any disagreement should be discussed with a third, independent reviewer. Detailed instructions how to use the GRADE approach are given in the online tutorial found at:

https://gdt.gradepro.org/app/handbook/handbook.html.

Step 24: Update, report, and submit for publication

When ready to submit the study for publication, if the interval since beginning the search of bibliographic databases is greater than 6–12 months the search should be updated to identify recently published articles.

Guidelines exist on how to report a systematic review and meta-analysis facilitating transparency, reproducibility, and comparability between studies. *PRISMA, QUOROM* (which evolved into PRISMA), and *MOOSE* are flowcharts that graphically describe the sequence of reporting a systematic review and meta-analysis. When submitting the study, it is essential to add as an attachment a detailed PRISMA or MOOSE report. PRISMA and MOOSE flowcharts are provided in online Supplements III and IV.

Finally, additional experts with content expertise may be invited to review and comment on the manuscript (and the published work should acknowledge their assistance). It is still possible to improve the quality of the publication further by appraising the interpretation of the results one last time.

Concluding remarks

Evidence syntheses constitute essential tools for evidence-based medicine and policy-making in a time of proliferating scientific publications and journals. Healthcare professionals and researchers must understand the principles of preparing such reviews and follow strict protocols to use them effectively. This 24-step guide can simplify the process of conducting a systematic review, provide healthcare professionals and researchers with the tools to conduct methodologically sound systematic reviews and meta-analyses, and enhance the quality of synthesis efforts already underway. The guide will increase readers' understanding of the complexity of

the process and the quality of published systematic reviews, and enhance the incorporation of knowledge synthesis into clinical decisions and policy-making.

Acknowledgements We would like to thank Georgia Salanti for the critical revision of the manuscript, Christopher Owen Ritter for English language editing, and 24-design.com for help with figures' design.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest related to this work.

References

1. Manchikanti L. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management, part I: introduction and general considerations. Pain Physician. 2008;11(2):161–86.
2. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Qual Rep Meta-Anal Lancet. 1999;354(9193):1896–900.
3. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264–9.
4. Dekkers OM, Vandenbroucke JP, Cevallos M, Renehan AG, Altman DG, Egger M. COSMOS-E: guidance on conducting systematic reviews and meta-analyses of observational studies of etiology. PLoS Med. 2019;16(2):e1002742. https://doi.org/10.1371/journal.pmed.1002742.
5. Conway A, Inglis SC, Chang AM, Horton-Breshears M, Cleland JG, Clark RA. Not all systematic reviews are systematic: a meta-review of the quality of systematic reviews for non-invasive remote monitoring in heart failure. J Telemed Telecare. 2013;19(6):326–37. https://doi.org/10.1177/1357633X13503427.
6. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018;555(7695):175–82. https://doi.org/10.1038/nature25753.
7. Nikolakopoulou A, Mavridis D, Furukawa TA, et al. Living network meta-analysis compared with pairwise meta-analysis in comparative effectiveness research: empirical study. BMJ (Clin Res). 2018;360:k585. https://doi.org/10.1136/bmj.k585.
8. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak. 2007;7:16. https://doi.org/10.1186/1472-6947-7-16.
9. Bettany-Saltikov J. How to do a systematic literature review in nursing: a step-by-step guide. Berkshire: McGraw-Hill Education; 2012.
10. Booth Andrew. Clear and present questions: formulating questions for evidence based practice. Library Hi Tech. 2006;24(3):355–68.
11. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res. 2012;22(10):1435–43. https://doi.org/10.1177/1049732312452938.
12. Haynes AS, Derrick GE, Redman S, et al. Identifying trustworthy experts: how do policymakers find and assess public health researchers worth consulting or collaborating with? PLoS ONE. 2012;7(3):e32665. https://doi.org/10.1371/journal.pone.0032665.
13. Oxman AD, Guyatt GH. The science of reviewing research. Ann NY Acad Sci. 1993;703:125–33 **discussion 33–4**.
14. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 2017;6(1):245. https://doi.org/10.1186/s13643-017-0644-y.
15. Rethlefsen ML, Farrell AM, Osterhaus Trzasko LC, Brigham TJ. Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. J Clin Epidemiol. 2015;68(6):617–26. https://doi.org/10.1016/j.jclinepi.2014.11.025.
16. Glisic M, Kastrati N, Gonzalez-Jaramillo V, et al. Associations between phytoestrogens, glucose homeostasis, and risk of diabetes in women: a systematic review and meta-analysis. Adv Nutr. 2018;9(6):726–40. https://doi.org/10.1093/advances/nmy048.
17. Franco OH, Chowdhury R, Troup J, et al. Use of plant-based therapies and menopausal symptoms: a systematic review and meta-analysis. JAMA. 2016;315(23):2554–63. https://doi.org/10.1001/jama.2016.8012.
18. Elamin MB, Flynn DN, Bassler D, et al. Choice of data extraction tools for systematic reviews depends on resources and review complexity. J Clin Epidemiol. 2009;62(5):506–10. https://doi.org/10.1016/j.jclinepi.2008.10.016.
19. Bramer WM, Milic J, Mast F. Reviewing retrieved references for inclusion in systematic reviews using EndNote. J Med Libr Assoc. 2017;105(1):84–7. https://doi.org/10.5195/jmla.2017.111.
20. Mourad Ouzzani HH, Fedorowicz Zbys, Elmagarmid Ahmed. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5:210. https://doi.org/10.1186/s13643-016-0384-4.
21. DistillerSR EP, Ottawa, Canada, Available at https://www.evidencepartners.com/.
22. Covidence. Cochrane Community. https://community.cochrane.org/help/tools-and-software/covidence. Accessed 3 Jul 2018.
23. Higgins JPT, Savovic J, Page MJ, Hróbjartsson A, Boutron I, et al. A revised tool for assessing risk of bias in randomized trials. Cochrane Database Syst Rev. 2016;10(Suppl 1):29–31.
24. Higgins JP, Altman DG, Gotzsche PC, et al. The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ (Clin Res Ed.). 2011;343:d5928. https://doi.org/10.1136/bmj.d5928.
25. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ (Clin Res Ed). 2016. https://doi.org/10.1136/bmj.i4919.
26. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
27. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158(4):280–6. https://doi.org/10.7326/0003-4819-158-4-201302190-00009.
28. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8. https://doi.org/10.7326/M18-1376.
29. Luhnen M, Prediger B, Neugebauer EAM, Mathes T. Systematic reviews of health economic evaluations: a protocol for a systematic review of characteristics and methods applied. Syst Rev. 2017;6(1):238. https://doi.org/10.1186/s13643-017-0639-8.
30. Rodgers M, Sowden A, Petticrew M, et al. Testing methodological guidance on the conduct of narrative synthesis in systematic reviews: effectiveness of interventions to promote smoke alarm ownership and function. Evaluation. 2009;15(1):49–73. https://doi.org/10.1177/1356389008097871.
31. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ (Clin Res ed.). 2007;335(7626):914–6. https://doi.org/10.1136/bmj.39343.408449.80.

32. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ (Clin Res Ed.). 2003;327(7414):557–60. https://doi.org/10.1136/bmj.327.7414.557.
33. Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods. 2006;11(2):193–206. https://doi.org/10.1037/1082-989X.11.2.193.
34. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58. https://doi.org/10.1002/sim.1186.
35. Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8:79. https://doi.org/10.1186/1471-2288-8-79.
36. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA. 2001;286(7):821–30. https://doi.org/10.1001/jama.286.7.821.
37. Reeves BCDJ, Higgins JPT. Wells GA Chapter 13: Including nonrandomized studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 510 [updated March 2011]. Oxford: The Cochrane Collaboration; 2011.
38. Abraham NS, Byrne CJ, Young JM, Solomon MJ. Meta-analysis of well-designed nonrandomized comparative studies of surgical procedures is as good as randomized controlled trials. J Clin Epidemiol. 2010;63(3):238–45. https://doi.org/10.1016/j.jclinepi.2009.04.005.
39. Tonin FS, Rotta I, Mendes AM, Pontarolo R. Network meta-analysis: a technique to gather evidence from direct and indirect comparisons. Pharm Pract (Granada). 2017;15(1):943. https://doi.org/10.18549/PharmPract.2017.01.943.
40. Dias S, Caldwell DM. Network meta-analysis explained. Arch Dis Child Fetal Neonatal Ed. 2019;104(1):F8–12. https://doi.org/10.1136/archdischild-2018-315224.
41. Taneri PE, Kiefte-de Jong JC, Bramer WM, Daan NM, Franco OH, Muka T. Association of alcohol consumption with the onset of natural menopause: a systematic review and meta-analysis. Hum Reprod Update. 2016;22(4):516–28. https://doi.org/10.1093/humupd/dmw013.
42. Nikolakopoulou A, Mavridis D, Salanti G. Demystifying fixed and random effects meta-analysis. Evid-Based Mental Health. 2014;17(2):53–7. https://doi.org/10.1136/eb-2014-101795.
43. Harris RBM, Deeks J, et al. Metan: fixed-and random-effects meta-analysis. Stata J. 2008. https://doi.org/10.1177/1536867X0800800102.
44. Chaimani A, Mavridis D, Salanti G. A hands-on practical tutorial on performing meta-analysis with Stata. Evid-Based Mental Health. 2014;17(4):111–6. https://doi.org/10.1136/eb-2014-101967.
45. Higgins JPT GSe. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.
46. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21(11):1559–73. https://doi.org/10.1002/sim.1187.
47. Marty Richardsona P, Doneganb S. Interpretation of subgroup analyses in systematic reviews: a tutorial. Clin Epidemiol Glob Health. 2019;7:192–8.
48. Hannah R, Rothstein AJS, Borenstein M. Publication bias in meta-analysis. New York: Wiley; 2005.
49. Mavridis D, Salanti G. Exploring and accounting for publication bias in mental health: a brief overview of methods. Evid Based Ment Health. 2014;17(1):11–5. https://doi.org/10.1136/eb-2013-101700.
50. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ (Clin Res Ed). 2006;333(7568):597–600. https://doi.org/10.1136/bmj.333.7568.597.
51. Sutton AJ, Higgins JP. Recent developments in meta-analysis. Stat Med. 2008;27(5):625–50. https://doi.org/10.1002/sim.2934.
52. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Stat Med. 2006;25(20):3443–57. https://doi.org/10.1002/sim.2380.
53. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50(4):1088–101.
54. Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group, 2013. Available from guidelinedevelopment.org/handbook. 2013.

Affiliations

Taulant Muka[1] · Marija Glisic[1,2] · Jelena Milic[3,4] · Sanne Verhoog[1] · Julia Bohlius[1] · Wichor Bramer[5] · Rajiv Chowdhury[6] · Oscar H. Franco[1]

1 Institute of Social and Preventive Medicine (ISPM), University of Bern, Mittelstrasse 43, 3012 Bern, Switzerland

2 Swiss Paraplegic Research, Nottwil, Switzerland

3 Department of Medical Informatics and Biostatistics, Institute of Public Health of Serbia, Belgrade, Serbia

4 Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands

5 Medical Library, Erasmus MC, Rotterdam, The Netherlands

6 Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge, UK