Introduction to Systematic Reviews and Meta-analyses of Therapeutic Studies

Murtadha Al-Khabori[1] and Wasif Rasool[2]

[1]Department of Hematology, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Oman

[2]Department of Medicine, Sultan Qaboos University Hospital, Muscat, Oman

Received: 6 June 2021

Accepted: 8 August 2021

Corresponding author: mkkhabori@gmail.com

DOI 10.5001/omj.2022.42

Abstract

A systematic review is a specific and reproducible method to search, identify, select, appraise, and summarize all studies relevant to a particular health care question. In this paper, we will review the concept of level of evidence, define the terms systematic review and meta-analysis, and outline the steps in performing a systematic review and meta-analysis with an illustrative example. In addition, we will introduce some important concepts in systematic reviews and meta-analyses like heterogeneity, publication bias, forest plots, and quality assessment. Finally, this review will focus on systematic reviews addressing therapeutic research questions.

Main Text

Two important concepts support any recommendation in clinical practice guidelines: level of evidence and strength of recommendation. The level of evidence is based on the number and the quality of studies addressing a particular clinical question. It is a reflection of how we believe in the finding. The lowest level of evidence comes from expert opinions based only on *in vitro* studies and animal research. In contrast, the highest level of evidence comes from systematic reviews and meta-analyses of large randomized clinical trials.[1]

A systematic review is a specific and reproducible method to search, identify, select, appraise, and summarize all studies relevant to a particular health care question.[2] Several steps are usually undertaken to perform a systematic review. This process starts with identifying a question then formulating it in a PICO (Patients, Intervention, Comparison, Outcome) format.[3] The outcome does

not need to be limited, and a systematic review should include all relevant outcomes to a particular clinical question. After a question is formulated, previous systematic reviews are searched and reviewed to decide if a new systematic review addressing the same question adds information to what is already available. Once a decision is made to perform a systematic review on the question of interest, then a search strategy is formulated and executed. Then, relevant studies are then selected from the search results. Subsequently, relevant data fields are extracted and summarized using tables and pooled effect size when appropriate. The included studies are critically appraised, and the results are appropriately presented. Assessment of heterogeneity and publication bias is then performed. Finally, the manuscript is written illustrating these steps and results.

Identifying and formulating a question

Systematic reviews should have a clearly stated set of objectives. There are several ways a clinical research question can arise.[4] The two commonest sources are during patients' care and when reviewing the literature. The question should be formulated in a PICO format most suited for questions on therapies. Systematic reviews should have a comprehensive assessment of research questions and should include all relevant outcomes of the interventions involved. Selecting studies with specific outcomes only may limit the comprehensive analysis of outcomes. The authors of systematic reviews should be clear as to what outcomes are relevant to a specific intervention. In the example we will use to illustrate the concepts in this review, the question is on the therapeutic benefit of Autologous Stem Cell Transplantation (ASCT) in managing patients with previously untreated follicular lymphoma.[5] The question in PICO format is "In adults with previously untreated follicular lymphoma, does ASCT improve event-free survival when compared to chemotherapy alone?" Notice that the question includes all components of the PICO format, including the outcome event-free survival. With such an intervention, other outcomes besides event-free survival are important; overall survival, quality of life, adverse events, including secondary myelodysplasia, are important to be included in the systematic review. It is therefore important to include studies addressing other outcomes during the formation of a search strategy.

Before proceeding to the next step, a comprehensive search for previous systematic reviews addressing this question is performed, and the need for a new systematic review and meta-analysis is assessed. A prior systematic review and meta-analysis on a question of interest does not preclude performing a new one. For example, if there are important studies that are published additionally to what is included in the previous systematic review, a new review is justified. In addition, if the previous systematic review suffers from a major methodological limitation, again, a new one is warranted.

Formulation of a search strategy

A comprehensive search for published and unpublished studies addressing a particular question improves the validity of the results of a systematic review. This step includes the decision on what and where to search, followed by the search itself. The support of a librarian and good access to medical bibliographic databases are very important in this step. Based on the research question, inclusion criteria are formulated, including the elements in the PICO and the study design. The eligibility

criteria for inclusion should be predefined. The decision to limit the search period and language should be carefully thought of as this may affect the number of included studies and the conclusion of the systematic review. In our example,[5] the inclusion criteria are adults with follicular lymphoma (patients), ASCT or chemotherapy (intervention and the comparison) and randomized clinical trials (study design of interest).

In this example, there are no exclusion criteria. This means unpublished studies and studies published in languages other than English are not excluded. Moreover, "older studies" are not excluded. Excluding "older studies" may be reasonable if the practice in these studies was very different from the practice in the more recent studies, which is presumed to be contemporary. The next step is the selection of bibliographic databases. This depends on the question of the systematic review; for biomedical research, the search should include the U.S. National Library of Medicine (MEDLINE),[6] Excerpta Medica dataBASE (EMBASE),[7] and Cochrane Central Register of Controlled Trials (CENTRAL)[8] in addition to other databases. These databases can be searched directly or through vendors. Databases indexing conferences proceedings should also be searched. Clinical trial registries should be searched if the question of the review includes clinical trials. A common mistake, which potentially may bias the results, is to exclude studies not yet published (also called grey literature).

When conducting the search itself, all possible variations of the terms of interest are searched using a free-text format. Medical Subject Headings (MeSH) terms, when available in the database, e.g., MEDLINE, should also be used for the search. The search is conducted on each element in the inclusion criteria and then combined using Boolean operators offered by the database search tool, e.g., "AND". The search from different databases is combined in one document, which is later reviewed to select studies.

Selection of studies

The citations from different bibliographic databases are gathered in one document to make the review and study selection easy and organized. The citations in this document are searched, and studies not relevant to the review question are excluded. The first stage is to review the titles and abstracts of the imported citations and exclude studies based on information available in these if not relevant to the study question. It is important to record the reasons for the exclusion. Usually, at this stage, the number of excluded citations is large, especially if the search strategy was sensitive and extensive. For the remaining citations, full-text papers are retrieved to review and decide on the selection. If languages other than English were searched, these full-text papers are translated. Again, it is important to record the reasons for excluding any paper. Two reviewers usually do the process of selection and exclusion, and it is presented in a flow diagram as recommended by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[9] In our example,[5] the search returns 1661 citations; of which, 1625 citations are excluded at the title and abstract stage. Thirty-six papers are then retrieved and reviewed, and 29 are excluded after reviewing the full text resulting in seven papers to be included in the systematic review.

Extraction of information from the included studies

Early in the review, reviewers develop two data extraction forms: description and quality assessment of the included studies. These forms are developed based on the review question. They include important information relevant to that question. The forms are reviewed by experts in the field of the question, and then they are piloted by the reviewers before starting the data extraction from the included studies. The extracted data are usually presented in a table format similar to table 1 of baseline information in original studies. In our example, the extracted information (presented in tables) includes the first author's name, year of publication, study design, number of enrolled patients, and description of intervention and control. At the level of enrolled patients in each study, an additional description of key characteristics relevant to the description of this population should also be included. These are usually presented as means or medians for continuous variables and proportions for categorical variables. In our example, these includes proportions of males and females, median age, and proportions of high-risk patients, patients with poor performance status, and patients with bulky disease. A reviewer and a co-reviewer perform data extraction from the included studies. Discrepancies are resolved by discussion or a third reviewer.

Summarizing the results

Results of the include studies in a systematic review are usually summarized in a table format. In addition, many systematic reviews include a synthesis of an overall estimate of the combined results. This process of pooling the results of different studies into one effect size using weighted averaging is called a meta-analysis. This is usually produced, where appropriate, for each outcome of the included studies in systematic reviews. Not all studies contribute the same to the pooled (combined) effect size because of different weighting methods. There are two main ways of assigning weights to studies when pooling the results; fixed-effects and random-effects models.[10] The fixed-effects model assumes a single effect size shared by all studies, so variation is assumed to be within studies only. Therefore, this model gives more weight to studies with larger sample sizes.

On the other hand, the random-effects model assumes that the effect size varies among studies, and therefore, variation is assumed to be within and between studies. This variation leads to wider confidence intervals in the pooled effect size, making this a more conservative approach to combining results in meta-analyses than the fixed-effects model. This is an important concept when inconsistencies of results between studies are seen. This will be further detailed under the assessment of heterogeneity below. Reviewers should decide which model to use at the stage of study protocol development; however, the other model can also be used to supplement the argument of the systematic review in a sensitivity analysis.

The graphical representation of the effect size from different studies and the pooled effect size is called a forest plot. Figure 1 shows a hypothetical example of a forest plot. The columns in the forest plot represent the included studies, results from these studies, and the number of patients in the intervention and control (where appropriate) groups in each study (and the total of these studies). Sometimes, the weights of the studies are also included. In the figure itself, a square (or a circle) represents each study, and bars on each side represent a confidence interval. A vertical line represents the line of neutrality or the line of no effect. If the confidence interval crosses the line of no effect, this indicates that the study result is not statistically significant. The sides of the line of no effect are

usually labeled for clarifying with “favors intervention” and “favors control” as appropriate. These labels relate the effect sizes of each study and the pooled effect size. A different shape (e.g., diamond) from the individual studies usually represents the pooled effect, and the sides of the used shape usually represent the confidence interval. A random-effects model is used in the forest plot in figure 1. Figure 2 uses the same raw data as figure 1 except for using a fixed-effects model to combine the results.

Assessment of heterogeneity

Our confidence in the conclusion of systematic reviews and meta-analyses decreases if the results of the included studies are different. This inconsistency in the results is called heterogeneity.[11] We believe more in a pooled analysis when the results are consistent with less heterogeneity. Various methods can be used to assess heterogeneity. The first method is a simple inspection of the forest plot. Forest plots are inspected visually for the overlap between the 95% confidence intervals of the different studies. Unless the results are extremely different, this method is unlikely to be accurate, albeit simple, quick, and does not need interpretation from a statistical test. In our hypothetical example, the visual inspection of the forest plot in figure 1 suggests that the results are different, meaning that there is heterogeneity. The second method is a chi-squared test, which measures the variation in the results and assesses if this variation is expected by chance. The question it asks is, “Are the results between studies different?” Therefore, if the *p*-value is less than 0.05, we conclude that the results are different and not expected by chance. This test is widely available in many statistical software and packages to perform meta-analyses, and the interpretation is simple. However, as most systematic reviews and meta-analyses include a relatively small number of studies, the test is usually underpowered, which means that if the *p*-value is more than 0.05, we cannot conclude that there is no heterogeneity. Therefore, the test is helpful only when it indicates that there is heterogeneity. In the example in figure 1, the *p*-value is 0.07, and consequently, we cannot conclude that the results are inconsistent despite what the inspection of the forest plot suggests.

The last two tests, visual inspection and chi-squared, indicate only the presence or absence of heterogeneity, and they do not quantify it. A test that can quantify this has been developed and is called the Cochrane I^2 test. It estimates the magnitude of inconsistency beyond what is expected by chance.[11] Values of over 50% indicate moderate heterogeneity that needs to be explained.[11] Unless heterogeneity is explained, the confidence in the conclusion cannot be established, and many question the value of pooling the results in this setting. There are different ways of exploring heterogeneity to explain it. These need to be stated a priori at the stage of study protocol development. One option, when there is substantial heterogeneity, is not to perform the meta-analysis. The random-effects model incorporates heterogeneity into wider confidence intervals, and therefore it is a better option than the fixed-effects model when inconsistency in study results is seen. Subgroup analysis is another method to explore heterogeneity. In this method, studies are divided into subgroups based on a specific characteristic, and heterogeneity is estimated for each of these subgroups. If the substantial heterogeneity disappears, one can conclude that the effect size is different in different subgroups, which may explain the heterogeneity. This conclusion is usually supplemented by a statistical test called the test of interaction.[12] Finally, regression (called meta-regression here) can be used to explore heterogeneity in a meta-analysis. Detailing these methods is beyond the scope of this review.

Assessment of publication bias

It is known that studies with positive results are more likely to be published and cited.[13] This gives an advantage of better visibility to these studies and is easier to find. If the search strategy of a systematic review is limited to only the published results, publication bias is likely to influence the conclusion and the combined outcome. Although there are several statistical methods to assess publication bias, they are limited in their ability to find it, especially if the number of the included studies is small. In addition, finding an abnormal distribution using these methods may not be specific to publication bias.

Systematic reviews should aim to minimize bias. The funnel plot is one of the commonest methods used in systematic reviews to assess publication bias. In this method, the effect size (x-axis of the forest plot) is plotted against the sample size (or a measure reflecting the weight of the study). Each circle on this plot represents a study. The assumption in this method is that the distribution of these circles should be symmetrical, especially when it comes to small studies. If the visual inspection of the plot indicates an asymmetry in the distribution, especially when there are fewer circles on the side of the small negative studies, this suggests publication bias. The distribution can also be assessed using statistical tests like Egger's test. If the *p*-value is less than 0.05, the test indicates that the distribution is not symmetrical, suggesting publication bias. However, like the chi-squared test in heterogeneity assessment, this test is also underpowered, and when it fails to indicate asymmetry, publication bias cannot be ruled out. In another hypothetical example in figure 3, the visual inspection of the funnel plot is not indicative of asymmetry, and therefore we do not have evidence of publication bias. However, in figure 4, the plot is asymmetrical, suggesting that small studies favoring controls are not included likely due to publication bias.

Critical appraisal of the included studies

This is the most important step in a systematic review. The conclusion of any review depends on the quality of the included studies. The average of poor-quality studies is a poor quality estimate. The role of the authors of a review is to assess the quality of the included studies and present the results in an explicit way for the readers. Additionally, the quality should affect how strongly the reviewers recommend their conclusion.

There are several methods to assess and present the quality of the included studies in a systematic review. Different methods are used for different study designs. One of the well-described methods to assess randomized clinical trials is the Jadad score.[14] This score assesses several domains: randomization, blinding, dropout, inclusion and exclusion criteria, adverse events, and statistical analysis for each of the included studies and presents them with a total score in a table or a figure format. Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) profile is a more recently described method, which has been increasingly used to evaluate randomized clinical trials.[15] This method assesses outcomes rather than individual studies. The following domains are evaluated in this tool: risk of bias, imprecision, inconsistency, indirectness, and publication bias. Finally, the Cochrane risk of bias tool is increasingly used, especially in Cochrane reviews.[16] In contrast to the last three methods used for the assessment of clinical trials, the Newcastle-Ottawa Scale is used to assess the quality of observational studies.[17] This method uses star points to evaluate three

main areas: selection, comparability, and exposure. Again, the total score is representative of studies and not outcomes.

Conclusions

A systematic review is a specific and reproducible method to search, identify, select, appraise, and summarize all studies relevant to a particular health care question. The synthesis of a combined result of the included studies is called a meta-analysis. This is usually presented in a forest plot. Publication bias is not uncommon in systematic reviews, and it is a difficult area to assess appropriately. Critical appraisal of the included studies is the most important step in a systematic review as it affects the confidence in how accurate is the pooled result.

References

1. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg* 2011; **128**(1)**:** 305-310. doi: 10.1097/PRS.0b013e318219c171

2. Gopalakrishnan S, Ganeshkumar P. Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. *J Family Med Prim Care* 2013; **2**(1)**:** 9-14. doi: 10.4103/2249-4863.109934

3. Eriksen MB, Frandsen TF. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *J Med Libr Assoc* 2018; **106**(4)**:** 420-431. e-pub ahead of print 2018/10/01; doi: 10.5195/jmla.2018.345

4. Tully MP. Research: articulating questions, generating hypotheses, and choosing study designs. *Can J Hosp Pharm* 2014; **67**(1)**:** 31-34. doi: 10.4212/cjhp.v67i1.1320

5. Al Khabori M, de Almeida JR, Guyatt GH, Kuruvilla J, Crump M. Autologous stem cell transplantation in follicular lymphoma: a systematic review and meta-analysis. *J Natl Cancer Inst* 2012; **104**(1)**:** 18-28. e-pub ahead of print 2011/12/21; doi: 10.1093/jnci/djr450

6. MEDLINE®: Description of the Database. In: U.S. National Library of Medicine, 2020.

7. Elsevier. Embase Coverage and Content | Elsevier. In, 2020.

8. Cochrane Controlled Register of Trials (CENTRAL) | Cochrane Library. In, 2020.

9. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009; **6**(7)**:** e1000097. e-pub ahead of print 2009/07/21; doi: 10.1371/journal.pmed.1000097

10. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods* 2010; **1**(2)**:** 97-111. e-pub ahead of print 2010/11/21; doi: 10.1002/jrsm.12

11. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**(7414)**:** 557-560. doi: 10.1136/bmj.327.7414.557

12. Richardson M, Garnera P, Doneganb S. Interpretation of subgroup analyses in systematic reviews: A tutorial. *Clinical Epidemiology and Global Health* 2019; **7**(2)**:** 192-198. e-pub ahead of print 24 May 2018; doi: https://doi.org/10.1016/j.cegh.2018.05.005

13. Sedgwick P. What is publication bias in a meta-analysis? *BMJ* 2015; **351:** h4419. e-pub ahead of print 2015/08/14; doi: 10.1136/bmj.h4419

14. Berger VW, Alperson SY. A general framework for the evaluation of clinical trial quality. *Rev Recent Clin Trials* 2009; **4**(2)**:** 79-88. doi: 10.2174/157488709788186021

15. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; **336**(7650)**:** 924-926. doi: 10.1136/bmj.39489.470347.AD

16. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I *et al.* RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019; **366:** l4898. e-pub ahead of print 2019/08/28; doi: 10.1136/bmj.l4898

17. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M *et al.* The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. In, 2020.