[ Guest Editorial ]

# Writing a Systematic Review and Meta-analysis: A Step-by-Step Guide

A systematic review aims to synthesize evidence on a specific topic through a structured, comprehensive, and reproducible analysis of the literature. This process is critical for developing an informed understanding of a given subject, allowing for evidence-based conclusions to guide further research, policy decisions, and clinical practice. Whereas scoping reviews follow a systematic approach to study selection similar to systematic reviews, the scope of the research question is broader, making them ideal when exploring emerging areas of research and more useful for providing an overview of the field and highlighting areas where further research is needed. On the other hand, systematic reviews focus on specific, well-defined research questions.

This paper will detail the step-by-step process to conducting a systematic review and meta-analysis, which involves the following: developing the research question and eligibility criteria, searching for and screening studies, extracting data, assessing study quality, synthesizing and analyzing data, assessing the certainty of the evidence, and interpreting results and making conclusions.

## SYSTEMATIC REVIEW VERSUS META-ANALYSIS

The terms "systematic review" and "meta-analysis" are often used erroneously interchangeably, but they serve distinct purposes. "Systematic review" refers to a comprehensive search and screening process that leads to the inclusion of all relevant studies on a specific topic. When data from a systematic review are pooled statistically, this is called a meta-analysis. This type of statistical analysis can be extremely useful when seeking to understand overarching trends or estimates of effect. When a systematic review is combined with a meta-analysis, this results in a quantitative synthesis of a comprehensive list of studies, allowing for a holistic understanding of the evidence through statistical evaluation.

## REPORTING SYSTEMATIC REVIEWS WITH PRISMA

Transparent and rigorous reporting is essential in systematic reviews, which is facilitated by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines.[5] These guidelines include a 27-item checklist encompassing the review's title, abstract, introduction, methods, results, discussion, and other relevant information as applicable. A checklist for journal and conference abstracts is also included in PRISMA 2020, along with a template flow diagram for tracking the study selection process, which can be modified depending on whether the systematic review is original or updated.[5] A well-reported systematic review should be reported according to the PRISMA guidelines. The PRISMA guidelines also recommend that any new systematic review be registered openly with PROSPERO - an international systematic review registry - before beginning the study, to encourage transparency and prevent duplicate work by multiple authors.

## DEVELOPING THE RESEARCH QUESTION AND ELIGIBILITY CRITERIA

The first step in the review process is to identify a topic of interest and formulate a focused research question. Typically, the PICO (Population, Intervention or Exposure, Comparator or Control, and Outcome) framework is used to define the scope of the review and formulate the research question. For example, a research question can take the following form: In [population of interest], does [intervention/exposure] compared with [comparator/control] lead to better or worse [outcome(s) of interest], although the specific form may vary. Once a research question is defined, eligibility (inclusion and exclusion) criteria must be established to guide reviewers in the study selection process. Defining a set of eligibility criteria will ensure that the included studies align with the objectives of the review.

*Address correspondence to Moin Khan, MD, MSc, McMaster University, Mary Grace Wing, Room G807, 50 Charlton Ave E., Hamilton, ON L8N4A6, Canada (email: khanmm2@mcmaster.ca).

The following author declared potential conflicts of interest: M.K. is the *Sports Health* Associate Editor for Systematic Reviews.

DOI: 10.1177/19417381251364686

© 2025 The Author(s)

It is important to note that the eligibility criteria for a systematic review and a primary research study differ in purpose and scope due to the nature of each type of study. For example, whereas the exclusion criteria for a primary research study will likely serve to exclude a certain subset of participants, the exclusion criteria for a systematic review typically serve to exclude certain types of publications such as review articles, conference abstracts, and technique articles, and study designs such as case reports and case series. The decision on the type of studies to include/exclude in a systematic review will ultimately depend on the research question and the availability of existing literature. While some topics may have numerous high-quality randomized controlled trials (RCTs) available, others may be limited to case series or other study designs of lower levels of evidence. Systematic reviews and meta-analyses of well-conducted, high-quality RCTs represent the pinnacle of evidence-based research. However, the strength of a systematic review is tied directly to the quality of the included studies. A review based on Level 1 evidence will be classified as Level 1, whereas a review based on Level 3 evidence remains Level 3. If a systematic review includes studies with varying levels of evidence, the overall classification is determined by the lowest level of evidence among the included studies. In other words, a systematic review is only as strong as its weakest component.

It is also important to note that the exclusion criteria should not be the opposite of the inclusion criteria. If the inclusion criteria specify that only RCTs will be included in the review, it would be redundant to mention the exclusion of other study types in the exclusion criteria.

## DEVELOPING A SEARCH STRATEGY

After defining the research question and eligibility criteria, the next step is to develop a comprehensive search strategy. A comprehensive search typically involves the use of ≥3 databases, with search strategies tailored to each individual database. Examples of commonly used databases include CENTRAL, MEDLINE, and Embase. Other specialized databases may also be utilized depending on the topic. It is important to note the difference between a platform and a database in this context. A platform serves as the interface that provides access to one or more database(s), whereas a database is where the actual literature is stored and retrieved. Examples of platforms include Ovid, PubMed, and Web of Science, which offer access to various databases such as MEDLINE and Embase. It is important to recognize this difference when creating a search strategy, as each platform has distinct search features, filters, and indexing systems that can influence search results. Although optional, collaborating with a professional librarian is strongly encouraged to help design and carry out a thorough and effective search.

Developing the search strategy involves combing key concepts from the research question, utilizing the Boolean operators "OR" and "AND," where similar concepts/words are grouped with "OR" and different concepts are tied together with "AND." To develop a focused search strategy, begin by identifying the most important PICO elements, typically focusing on P and I, with occasional emphasis on O or study design based on eligibility criteria. Start by compiling a list of relevant terms for both P and I and identify synonyms and related terms to expand the search using the Boolean operator "OR." Search for words used commonly by authors in titles, abstracts, and database indexing. Different databases will have different indexing terms (subject headings and keywords), which is why it is important to tailor the search strategy to each database. Subject headings are database-specific, controlled terms, such as MeSH terms in MEDLINE and Emtree terms in Embase, whereas keywords are uncontrolled terms that can take on different forms.

To use subject headings, identify the first key concept and enter it in the search bar. One or more options will appear. Select the preferred subject heading from the suggested options, choosing to "focus" or "explode" the term as appropriate. Multiple subject headings may be used depending on search results. Once all relevant subject headings have been added, keywords can be added to refine the search. When selecting keywords, consider synonyms, alternate spellings, acronyms, and truncation, as well as adding ".mp." at the end of the keyword to search across all fields in the database. Truncation involves adding an asterisk after the root of the keyword, which tells the database that the word can be completed in any way after the asterisk, e.g., dislocat* = dislocate, dislocating, dislocation, dislocations. Unlike subject headings, keywords are not standardized, which means all the variants of each term need to be considered and captured. Repeat this process for each keyword related to your first concept and combine all terms with "OR" to capture any variation within that concept. After completing this for each concept, combine the sets with "AND" to narrow the search scope effectively. Run the search and evaluate the number of results.

It is likely that this process will need to be repeated a few times before the search is finalized. It is best to start broadly and assess the number and relevance of initial hits by reviewing the first 30 results to get an idea of the types of articles that were retrieved. Depending on these results, the search may either need to be narrowed or broadened, or keywords may need to be adjusted. If exploding a subject heading produced too many irrelevant results, focusing it instead can help. Specific limits such as publication year or language may also be applied to narrow down results. It is important to note that the number of hits to expect will depend on the topic. Broad topics typically yield >2000 hits, whereas narrower topics have fewer. Once the search strategy is finalized, run the searches in their respective databases on the same day, export results, and import them into a reference management software to begin the screening process.

## SCREENING STUDIES

Once the databases searches are complete and the results have been imported to a reference management software (such as

Covidence, Rayyan, or similar tools), the next step is to screen the studies to identify those relevant for inclusion. Screening is performed in duplicate by independent reviewers to minimize bias and increase reproducibility. This process begins with the removal of duplicate entries, followed by title and abstract screening.

At the beginning of the title and abstract screening phase, reviewers may conduct a pilot screening of a small percentage of the total studies to identify and resolve any discrepancies in interpretation of the eligibility criteria. Once the piloting process is complete, both reviewers will complete the rest of the title and abstract screening independently. Any uncertainties or conflicts between reviewers at this stage should be included in the full-text screening for further assessment to avoid premature exclusion.

At the full-text screening stage, reviewers must document specific reasons for excluding each article. Any conflicts at this stage are typically addressed and resolved by discussion and consensus among the 2 reviewers, or by a third reviewer. The inter-rater reliability should be measured at both the title and abstract as well as full-text screening stages. This is typically calculated using Cohen's kappa (κ) coefficients and reported in the results section of the manuscript.[4] The reference lists of the included studies should also be searched manually to ensure that all relevant articles have been captured. This may also involve manually searching the reference lists of other similar systematic reviews.

The PRISMA flow diagram should be used to track and report the study selection process, including the databases searched and number of hits per database, total number of records screened at each stage, reasons for exclusion at the full-text stage, whether any articles were identified through manual searches, and final number of included studies.

## DATA EXTRACTION

Once the authors have finalized the list of included studies, the next step is to extract the data. This process should also be done in duplicate to ensure consistency and reliability and minimize transcription error. Data extracted from each study generally include author, year of publication, study design, sample size, population demographics, interventions, comparators, and outcomes. Other specific datapoints of interest may vary based on the research question. Of note, reviewers should avoid extracting individual study conclusions. Instead, they should draw their own conclusions based on their own analyses of the data.

A data extraction template should be created a priori and is often piloted on a small number of studies to verify that all relevant information is captured. When extracting outcome data, it is important to record specific details about the outcome measures, such as the name, direction of the scale (i.e., which direction represents a favorable outcome), version used (full version or abbreviated), total possible score, and other relevant specifications. It is also crucial to specify the follow-up timepoint(s) at which these outcomes are being reported. Since outcomes are often measured at multiple timepoints, reviewers should decide in advance whether they will extract data at all available timepoints or focus on a specific one of interest. This decision should align with the review's objectives and be applied consistently throughout the extraction process.

## ASSESSMENT OF STUDY QUALITY

The quality of all included studies will then be assessed. Similar to the process of study screening and data extraction, quality assessments should also be conducted in duplicate by 2 independent reviewers to increase the validity and reliability of the assessments. The selection of a quality assessment tool depends on the type of study design being evaluated. For RCTs, the revised Cochrane Risk of Bias (RoB-2) tool is widely regarded as the gold standard.[9] Nonrandomized or observational studies are typically assessed using the Cochrane Risk of Bias In Nonrandomized Studies - of Interventions/Exposures (ROBINS-I/E) tool,[8] or the Methodological Index for Nonrandomized Studies (MINORS) tool.[7] Tools like QUADAS-2 (Quality Assessment Tool for Diagnostic Accuracy Studies) and QUIPS (Quality in Prognostic Studies) are suited for diagnostic accuracy and prognostic studies, respectively.[2,12] Each tool evaluates specific risk areas, allowing reviewers to critically appraise the methodological rigor of included studies, identify potential sources of bias, and assess the reliability of the evidence base upon which conclusions are drawn.

Although the quality assessments may be discussed on their own, they also play a role in the Grading of Recommendations Assessment, Development and Evaluation (GRADE) assessment of the certainty of the evidence (discussed below).

## SYNTHESIZING THE DATA

Data synthesis in systematic reviews can be performed with or without meta-analysis, depending on the available data. Synthesis without meta-analysis is referred to as narrative synthesis. If there is significant heterogeneity in key characteristics, such as populations, interventions, or outcome measures, then combining the data may not be appropriate, and synthesis should be avoided altogether. To determine suitability for synthesis, it is essential to assess the heterogeneity among the included studies. Key considerations include whether the populations are comparable, whether the interventions and comparators are similar in terms of type, dosage, frequency, and delivery, and whether the same outcomes are being measured across studies. If these elements differ significantly, synthesis may not be methodologically sound. When studies are sufficiently homogeneous, the next step is to determine whether a meta-analysis is feasible based on the format and consistency of the data. If not, a narrative synthesis may be used instead to summarize the findings in a structured and systematic manner.

### Synthesis Without Meta-Analysis (Narrative Synthesis)

In cases where data cannot be combined quantitatively due to different reporting formats, but studies are otherwise

comparable, narrative synthesis is used. This method differs from "no synthesis," as it still involves analyzing how outcomes compare across studies, though without statistical synthesis as in a meta-analysis.

When synthesizing studies, both the direction and magnitude of the effects observed across the studies should be considered and discussed. When reporting the direction of effect, studies should be identified as showing either benefit or harm. For example, "Of the 6 studies that reported data on quality of life, 4 studies found improvements with respect to the intervention." Reporting the magnitude of effect involves evaluating the size of the reported effect measure, such as risk ratio (RR), odds ratio (OR), risk difference (RD), mean difference (MD), or other, depending on the data. This provides information about how substantial the observed effects are, beyond just whether they are statistically significant.

Regardless of the approach, the statistical significance of individual studies should not be summarized. The following should be avoided: "Of the 6 studies that reported data on quality of life, 4 studies reported statistically significant improvements with respect to the intervention." Statistical significance alone does not reflect the direction or magnitude of the effect and can be misleading, particularly in studies with small sample sizes.

### Synthesis With Meta-Analysis

For studies that report data using comparable formats and measurement scales, conducting a meta-analysis may be appropriate. Numerous statistical software, such as the Cochrane-endorsed Review Manager program (RevMan), MIX 2.0, MetaStat, and DataParty, are widely available to conduct meta-analyses. The results (effect sizes and confidence intervals [CIs]) should be reported both quantitatively and graphically with forest plots (discussed below).

For dichotomous outcomes, data requirements from each study include the total sample size and number of events per group. The choice of effect measure (RR, OR, or RD) will depend on the type of data, study design, and desired interpretation.

For continuous outcomes, the means and standard deviations, and total sample size from each study are needed. If a study reports medians and ranges instead of means and standard deviations, the appropriate values can be estimated using the data provided.[11] The choice of effect measure will also depend on the type of data. Mean difference (MD) should be used if the same scale is used across all studies (e.g., all studies assess pain on a visual analog scale (VAS) from 0 to 10, with 10 representing a worse outcome). Standardized mean difference (SMD) should be used if studies use different scales to assess the same outcome (e.g., quality of life assessed using the EQ-5D-5L, QOLS, WHOQOL, etc).

Because scales can differ in range (e.g., 0-10, 0-50, 0-100) and direction (higher scores may indicate better or worse outcomes), standardization is necessary. For example, to convert an outcome from a scale of 0 to 100 to a scale of 0 to 10, the result should be divided by 10. Similarly, if some scales are in different directions, the scores should be inverted so that all outcomes align in the same direction (e.g., lower number always represents worse outcome).

Once the data are transformed and ready to be pooled and analyzed, the statistical approach must be chosen, including selecting a fixed-effect or random-effects model. These models make different assumptions about heterogeneity and the decision is typically based on the research question. A fixed-effect model assumes a single true effect across all studies, where any observed variation is due solely to sampling error, making it more suitable for narrow PICO questions (less expected heterogeneity). In contrast, a random-effects model does not estimate one true effect but rather the mean of a distribution of effects, assuming treatment effects may vary between studies, which is often preferable for broader PICO questions (more expected heterogeneity). The model selection is made based on assumptions related to level of heterogeneity based on the research question before actual heterogeneity is known.

Once the model is selected and the data are inputted into the statistical software, the software will then create a forest plot displaying the results of the meta-analysis. In a forest plot, each individual study is represented by a square, with the position of the square indicating the effect size. The horizontal lines extending from the square show the CIs, illustrating the range of uncertainty around the effect estimate. A vertical line in the center typically represents the line of no effect (e.g., a RR/OR of 1 or a MD/SMD of 0). At the bottom of the plot, a diamond represents the overall pooled effect from all included studies. The center of the diamond shows the combined effect estimate, and the width reflects the overall CI.

## HETEROGENEITY AND SUBGROUP ANALYSES

When conducting a meta-analysis, heterogeneity between studies must be assessed before drawing conclusions. This is typically done through visual inspection of forest plots, which involves checking the alignment of point estimates and overlap of confidence intervals, as well as statistically using measures like the $I^2$ statistic.

If point estimates show similar direction and magnitude with overlapping confidence intervals, heterogeneity is likely minimal. The $I^2$ statistic quantifies the proportion of variability due to heterogeneity rather than chance, with values from 0% to 40% indicating minimal heterogeneity, 30% to 60% moderate, 50% to 90% substantial, and 75% to 100% considerable heterogeneity.[10] These thresholds overlap, as interpretation ultimately depends on the specific research question, the expectations established a priori, as well as the results of the visual inspection.

Potential sources of heterogeneity may be explored using subgroup analyses. Clinical parameters such as population, intervention, and outcome should be considered, as well as

methodological nuances such as study design and risk of bias assessments. After running the subgroup analysis, the results should be evaluated to identify whether the effects differ between subgroups and whether the subgroup variable adequately explains the observed heterogeneity. The ICEMAN criteria can be used to determine the reliability of subgroup findings, ensuring that only meaningful subgroup effects are considered.[6]

## GRADE FRAMEWORK FOR CERTAINTY OF EVIDENCE

The GRADE framework provides criteria for assessing the certainty of the evidence by evaluating 5 domains: risk of bias, inconsistency, indirectness, imprecision, and publication bias.[1] Levels of certainty are categorized into 4 levels: high (very confident in the effect estimate), moderate, low, and very low (very little confidence in the effect estimate). It is important to note that a high level of certainty does not necessarily indicate a strong or beneficial treatment effect, but rather confidence in the evidence that produced the observed result. For example, reviewers may conclude with high certainty that an intervention has minimal or no effect on an outcome. This distinction is essential, as the certainty rating informs how confidently reviewers can interpret and present findings without implying causality or overstating effect sizes.

## INTERPRETING RESULTS AND MAKING CONCLUSIONS

Once the results have been analyzed, the final step in the review process is to interpret the results based on effect sizes, CIs, and the certainty of the evidence. Discussing statistical significance of the findings should be avoided. This approach allows readers to understand the magnitude of the effect and the robustness of the findings. CIs, which capture potential variability in the results, provide a more nuanced view than dichotomizing findings as statistically significant or not. Making specific recommendations based on the findings should also be avoided. Reviewers should merely present objective results and discuss these findings with consideration to the certainty of the evidence. Although hypotheses may be generated, making definitive treatment recommendations is discouraged.

## MANUSCRIPT STRUCTURE

The manuscript should follow a structured format including the following sections: Introduction, Methods, Results, Discussion, and Conclusion.

The introduction should provide an overview of the topic, identify the evidence gap being addressed, and clearly state the purpose of the review. The structure should follow the problem/gap/hook approach to improve clarity and readability.[3]

The methods section then describes the review process. This section typically includes the following subsections: eligibility criteria, search strategy, screening, data extraction, quality assessment, heterogeneity assessment, assessment of the certainty of the evidence, and data analysis. The data analysis section should specify whether data will be synthesized, with or without meta-analysis if applicable, and outlines any planned subgroup analyses if applicable.

The results section will typically begin by detailing the results of the search and screening process (PRISMA flow diagram, inter-rater reliability calculation). This is then followed by a summary of the characteristics of the included studies and patient demographics, often displayed in a tabular format and presented briefly in text. After this, the results of the quality assessments should be presented, followed by the outcomes. For each outcome, results of the certainty and heterogeneity assessments should be included as applicable.

The discussion should provide a summary of the key findings, relate these findings to existing literature, and discuss implications for clinical practice and research. This section should also include a discussion of the review's strengths and limitations, focusing on the methodological aspects of the review rather than the strengths and limitations of the individual included studies.

Finally, the conclusion should be brief and focused, clearly stating the main findings and offer recommendations for future research if appropriate. It must remain concise, avoiding detailed analysis or discussion, and serve as a clear final takeaway.

## CONCLUSION

In sum, systematic reviews offer a meticulous, structured approach to synthesizing evidence, with a focus on transparency, consistency, and comprehensive assessment. Writing a high-quality systematic review and meta-analysis involves developing the research question and eligibility criteria, searching for and screening studies, extracting data, assessing study quality, synthesizing and analyzing data, assessing the certainty of the evidence, and interpreting results and making conclusions. By applying rigorous methodologies in each of these steps, systematic reviews and meta-analyses yield invaluable insights and guide subsequent research, policy development, and clinical practice.

—Danielle Dagher, MSc
—Moin Khan, MD, MSc*
*McMaster University, Hamilton, ON, Canada*

## REFERENCES

1. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction - GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol.* 2011;64(4):383-394. doi:10.1016/j.jclinepi.2010.04.026

2. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann Intern Med*. 2013;158(4):280-286. doi:10.7326/0003-4819-158-4-201302190-00009
3. Lingard L. Joining a conversation: the problem/gap/hook heuristic. *Perspect Med Educ*. 2015;4(5):252-253. doi:10.1007/s40037-015-0211-y
4. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-282.
5. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
6. Schandelmaier S, Briel M, Varadhan R, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ*. 2020;192(32):E901-E906. doi:10.1503/cmaj.200077
7. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (minors): development and validation of a new instrument. *ANZ J Surg*. 2003;73(9):712-716. doi:10.1046/j.1445-2197.2003.02748.x
8. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919. doi:10.1136/bmj.i4919
9. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. doi:10.1136/bmj.l4898
10. Thorlund K, Imberger G, Johnston BC, et al. Evolution of heterogeneity ($I^2$) estimates and their 95% confidence intervals in large meta-analyses. *PLoS One*. 2012;7(7):e39471. doi:10.1371/journal.pone.0039471
11. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol*. 2014;14:135. doi:10.1186/1471-2288-14-135
12. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009


For article reuse guidelines, please visit Sage's website at http://www.sagepub.com/journals-permissions.