Sanjay Patole *Editor*

Principles and Practice of Systematic Reviews and Meta-Analysis

Springer

Principles and Practice of Systematic Reviews and Meta-Analysis

Sanjay Patole
Editor

Principles and Practice of Systematic Reviews and Meta-Analysis

Springer

Editor
Sanjay Patole
School of Medicine
University of Western Australia
Perth, WA, Australia

ISBN 978-3-030-71920-3 ISBN 978-3-030-71921-0 (eBook)
https://doi.org/10.1007/978-3-030-71921-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Evidence-Based Medicine (EBM) is at the core of modern medicine [1, 2]. EBM is the integration of individual clinical expertise with the best available clinical evidence from systematic research and patient's values and expectations [1, 2]. EBM requires that decisions should be taken based on the body of evidence, and not just a single study [3]. Systematic reviews offer evidence that is as good as the best available evidence summarized by the review [3].

Systematic reviews are "the most reliable and comprehensive statement about what works," and involve identifying, synthesizing, and assessing all available evidence, quantitative and/or qualitative, to generate a robust, empirically derived answer to a focused research question [4]. Since their introduction in medical sciences in 1970s, systematic reviews have been adopted in a wide range of fields, from astronomy, international development, and global health, to zoology [5–7]. The importance of systematic reviews with meta-analyses as the best source of evidence cannot be overemphasized considering that health care staff, public health policy-makers, and researchers have limited time to catch up with and critically appraise the vast amount of literature that gets added every day [8].

Written by clinicians, the objective of this reader-friendly book is to introduce the readers from various faculties of science to the principles and practice of systematic reviews and meta-analysis. Our aim is to help them in developing skills to use this precious tool for guiding their clinical practice and research [8].

Perth, WA, Australia — Sanjay Patole

References

1. Sense about science because evidence matters: Evidence based medicine http://www.senseaboutscience.org/pages/evidence-based-medicine.html.
2. Guyatt GH, Haynes RB, Jaeschke RZ, et al. Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence Based Medicine Working Group. *JAMA* 2000; 284(10):1290–6.
3. Guyatt G, Rennie D, Meade MO, Cook DJ. User's guide to the medical literature: a manual for evidence-based clinical practice. 2nd ed. London: AMA; 2002.
4. van der Knaap, LM, Leeuv FL, Bogaerts S, Nijssen LTJ. Combining Campbell standard and the realist evaluation approach: the best of two worlds? *American Journal of Evaluation* 2008, 29(1), 48–57.
5. Petticrew M. Systematic reviews from astronomy to zoology: myths and misconceptions. *BMJ* 2001; 322(7278):98–101.
6. Malletta R, Hagen-Zankerb J, Slaterc R, Duvendackd M. The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness* 2012; 4(3): 445–455.
7. Schlosser RW. The role of systematic reviews in evidence-based practice, research, and development. Focus—A Publication of the National Center for the Dissemination of Disability Research (NCDDR) 2006; Technical Brief No 15: 1–4.
8. 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.

Acknowledgments

I would like to express my sincere gratitude to the Associate Editor of this book, A/Prof. Shripada Rao. This book would not have been completed without his significant contribution in every aspect, from concept to completion.

Sanjay Patole, MD, FRACP, DrPH

Contents

Contributors

Gayatri Athalye-Jape Neonatal Directorate, King Edward Memorial Hospital, Perth, WA, Australia;
School of Medicine, University of Western Australia, Perth, Australia

Sam Athikarisamy Neonatal Directorate, King Edward Memorial Hospital for Women, Perth, WA, Australia;
School of Medicine, University of Western Australia, Perth, WA, Australia

Mangesh Deshmukh Department of Neonatology, Fiona Stanley Hospital, School of Medicine, Curtin and University of Western Australia, Perth, WA, Australia

Kwi Moon Department of Pharmacy, Perth Children's Hospital, Perth, WA, Australia;
School of Medicine, University of Western Australia, Perth, WA, Australia

Gillian Northcott Graphic Design, Medical Illustration, Perth Children's Hospital, Perth, WA, Australia

Mohan Pammi Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA

Sanjay Patole Neonatal Directorate, King Edward Memorial Hospital for Women, Perth, WA, Australia;
School of Medicine, University of Western Australia, Perth, WA, Australia

Abhijeet Rakshasbhuvankar School of Medicine, Neonatal Directorate, King Edward Memorial Hospital for Women, University of Western Australia, Perth, WA, Australia

Shripada Rao School of Medicine, Neonatal Directorate, Perth Children's Hospital, University of Western Australia, Perth, WA, Australia

Sven Schulzke Neonatologist, Director of Research, University Children's Hospital Basel UKBB, Basel, Switzerland

Ravisha Srinivasjois School of Medicine, University of Western Australia, Perth, WA, Australia

Yemisi Takwoingi Institute of Applied Health Research, Public Health Building, University of Birmingham, Birmingham, UK

Systematic Reviews, Meta-Analysis, and Evidence-Based Medicine

Sanjay Patole

Abstract Evidence-based medicine (EBM) is at the core of current clinical practice. The philosophical origins of EBM date as far back as the mid-19th century earlier. David Sackett (1934-2015) considered as the father of EBM, described it as '*the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients'*. EBM requires that clinical decisions should be based on the evidence in totality, and not on just a single study. Systematic reviews offer the best available evidence for decision making in clinical practice. They are 'the most reliable and comprehensive statement about what works', and involve identifying, synthesising and assessing all available evidence by a systematic approach, to generate a robust, empirically derived answer to a focused research question. A systematic review may or may not contain a statistical analysis (Meta-analysis) depending on whether it is possible, and importantly, sensible to combine data from different studies on the same subject, or not. This chapter covers the history, principles and characteristics of systematic reviews and meta-analysis in the context of EBM.

Keywords Evidence-based medicine · Systematic reviews · Narrative reviews · Meta-analysis · History · Principles · Practice · Hierarchy

Introduction

Evidence-based medicine (EBM) is at the core of current clinical practice (http://www.senseaboutscience.org/pages/evidence-based-medicine.html; Guyatt et al. 2000). The philosophical origins of EBM date as far back as the mid-19th century

S. Patole (✉)
School of Medicine, University of Western Australia, Perth, WA 6009, Australia
e-mail: sanjay.patole@health.wa.gov.au

S. Patole
Neonatal Directorate, King Edward Memorial Hospital for Women,
Perth, WA 6008, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_1

Paris and earlier (Anderson 2015). David Sackett (1934-2015) considered as the father of EBM, described it as'the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients' (Anderson 2015; Sackett 1997). To him, the practice of EBM meant 'integration of individual clinical expertise with the best available external clinical evidence from systematic research'. EBM requires that clinical decisions should be based on the evidence in totality, and not on just a single study (Guyatt et al. 2002). The role of a systematic review is to offer the best available evidence for decision making in clinical practice (Guyatt et al. 2002).

Narrative Reviews

Narrative reviews often reflect 'Eminence-based medicine' as they are mostly written by invited experts (Isaacs and Fitzgerald 1999). Needless to say, they are influenced by the author's intuition, experience, and inevitably, their bias. Critics point out that narrative reviews are a quick, easy, and an inexpensive way to reach desired conclusions! (Isaacs and Fitzgerald 1999).

The definition of what constitutes 'evidence' is subjective. Narrative reviews could be evidence-based, but still not truly useful as scientific evidence. In the absence of a clear section titled 'methods', it is difficult to understand how the evidence was derived and interpreted in narrative reviews. The lack of clarity and transparency and the element of subjectivity make it challenging to derive reliable, unbiased interpretation and conclusions on a specific topic when appraising narrative reviews (Isaacs and Fitzgerald 1999). For example, a comparison of seven narrative reviews, including the same studies showed that different reviewers reached different conclusions!!(Cipriani and Geddes 2003) The case of vitamin C as an intervention for cold illustrates the issues with narrative reviews quite well. The narrative review of vitamin C ('How to live longer and feel better, Linus Pauling 1986') had concluded that "We should be getting 200 times the amount of vitamin C that the Food and Nutrition Board recommends" (Linus Planning 2006). The author, Linus Pauling, was one of the founders of quantum chemistry and molecular biology, and one of the 20 greatest scientists of all time who went on to win the Nobel Prize in Chemistry in 1954 (Global Firsts and Facts 2017). Furthermore in the year 2000 he was acknowledged as the 16th most influential scientist in history. A subsequent systematic review of vitamin C for the cold by investigators from Oxford involved an exhaustive search of databases, journals and special collections (Knipschild 1995). It identified 61 trials, of which 15 were methodologically sound. The results of this systematic review suggested that even in megadoses Vitamin C cannot prevent a cold, though it might shorten its duration if already infected. The reviewers pointed out that the narrative review had missed five of the 15 methodologically sound trials, and had referred to other two only in passing (Knipschild 1995). Considering their limitations, narrative reviews are becoming less and less prevalent in the era of EBM.

Systematic Reviews

Systematic reviews are 'the most reliable and comprehensive statement about what works', and involve identifying, synthesising and assessing all available evidence by a systematic approach, to generate a robust, empirically derived answer to a focused research question (Isaacs and Fitzgerald 1999). Systematic reviews have been used in a wide range of fields, 'from astronomy and zoology' to international development, and global health, and were introduced to the medical field only in the 1970s (Petticrew 2001; Malletta et al. 2012; Schlosser 2006; O'Rourke 2007).

A Brief History of Systematic Reviews

The phrase 'systematic review' was mentioned in the early and mid-19th century in few publications on the classification of species in biology and zoology (Mees 1957; Alm 1916). In the late 1970s and early 1980s a group of health researchers in Oxford prepared the ground for EBM by beginning a programme of systematic reviews on the effectiveness of health care interventions.

Archie Cochrane called for developing medicine based on randomised controlled trials (RCTs) in his seminal book in 1972 titled 'Effectiveness and Efficiency: Random Reflections on Health Services' (Cochrane 1972). Later, his call for the critical summary of all RCT's (1979) led to the establishment of a collaborative database of perinatal trials (The Cochrane Collaboration 2017). Systematic reviews of RCTs started to get published in the 1980s, and in 1987 he encouraged others to adopt the methodologies used in these reviews. Archie Cochrane's untiring efforts and the increasing acceptance of EBM subsequently led to the opening of the Cochrane Collaboration Centre in Oxford, the UK in 1992, shortly after his death (Cochrane 1972; EPPI 2017; Brent Thoma 2013). The push for systematic reviews in the medical world started with the meeting organised by the British Medical Journal and the Cochrane Centre in London in 1993 (Chalmers and Altman 1995). The group at this meeting aimed to improve the scientific rigour of reviews in clinical medicine for a reliable and evidence-based approach in advising treatments. They believed that in the absence of scientific methods, advice on some lifesaving therapies had been delayed for over a decade, while others shown to be harmful in controlled trials continued to be offered (Oxman and Guyatt 1988). Systematic reviews, as we understand them today, represent the structured approach to undertaking literature reviews on earlier research studies and they are tied closely to meta-analyses, i.e. a statistical method for combining the data from the previous studies. We will learn more about meta-analysis later in this book.

The importance of systematic reviews as the best source of evidence for practising EBM cannot be overemphasised considering that health care providers, public health policymakers, and researchers often have limited time to catch up with and critically appraise the vast amount of literature that gets added every day

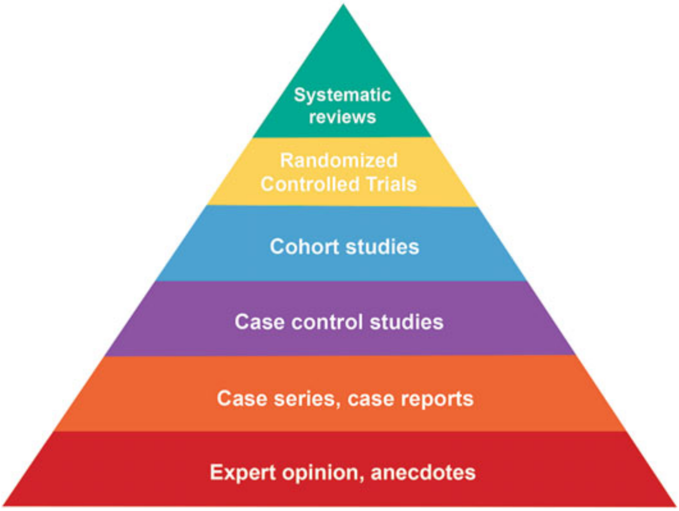

Systematic reviews
Randomized Controlled Trials
Cohort studies
Case control studies
Case series, case reports
Expert opinion, anecdotes

Fig. 1 Hierarchy of evidence pyramid

(Malletta et al. 2012; Glasziou et al. 2004). RCTs are considered as the gold standard in clinical research as they address the issue of not only the known but also the unknown confounders, something that other study designs ('Non-RCTs': cohort studies, case-control studies) cannot do. Systematic reviews of RCTs are, therefore, at the top of the pyramid of the hierarchy of evidence in EBM. However, assessing the risk of bias in various domains (e.g. randomisation, allocation concealment) of the included trials is important before accepting systematic reviews of RCTs as the gold standard in EBM (Fig. 1).

What Does Systematic Review Involve?

A systematic review involves systematic identification and evaluation of all the available relevant evidence to guide clinical practice, research, and policy. A systematic review focuses on a specific question; uses clearly stated, prespecified scientific methods to identify, select, assess, and summarise the findings of similar but separate studies. As Gene Glass said 'a systematic review is an analysis of analyses'. It is important to know that a systematic review may or may not contain a statistical analysis (Meta-analysis) depending on whether it is possible, and importantly, sensible to combine data from different studies on the same subject, or not (Douglas Altman 2013; Chinchilli 2007).

Why Do We Need Systematic Reviews?

Limited time to catch up with and critically appraise the vast amount of literature is not the only reason why we need systematic reviews. A comprehensive search and unbiased interpretation of the best available evidence—a critical component of EBM, is difficult without being systematic. Systematic reviews are useful in interpreting conflicting results of primary studies, synthesising results of a large number of primary studies, and judging external applicability of the evidence, especially when there are only a few primary studies. Reproducibility of results is another important benefit of systematic reviews given the transparency and clarity of their methodology. Systematic reviews help us know existing research (and its quality) in our area of interest, prevent duplication of efforts by letting us know what has already been done, and provide insights through the comparison and/or combination of different studies (Oakley et al. 2005).

What Are the Principles of Systematic Reviews?

A systematic review needs to have a focused, well defined, useful, and importantly, an answerable question. It requires a clear title and objectives with explicit and justified predefined inclusion and exclusion criteria. The question needs to convey, with clarity, the patients (P), intervention (I), control/comparison (C), the outcome of interest (O), and the study design (S). This is the PICOS format of the question that the systematic review is addressing. Some prefer to add the study time frame (T) to the phrased question, resulting in the abbreviation PICOT.

Considering the aim is to provide comprehensive and best available evidence, it should have a clearly documented *and* comprehensive search strategy for tracing all relevant studies-published as well as unpublished. Providing details of the search strategy makes it possible to reproduce the search results, increasing the validity of search methodology. To assure minimisation of bias, it should have a pre-stated method for critical appraisal of included studies using pre-stated methods.

The type of synthesis of the results (Quantitative, i.e. meta-analysis or Qualitative) depends on whether it is possible, and sensible to combine the data from ‘more or less similar; but different individual studies together. This is perhaps the most important step in systematic reviews.

Unbiased interpretation and conclusions and putting research into context are important. Finally, systematic reviews are required to have a structured report for the dissemination of results with clarity to the broader community.

As discussed above, assuring transparency, clarity, and objectivity at each step of the systematic review is important (Table 1). The practical approach to a systematic review is summarised in Table 2. The approach can be summarised in a sentence: *Ask a focussed question; tell the readers what exactly you did in an attempt to answer it, how and why?* Baumeister et al. have emphasised the

Table 1 Characteristics of a systematic review*

• **T**: Transparency at each step
• **R**: Reproducible and robust methodology
• **U**: Unbiased (Best precautions at each step to minimise bias)
• **E**: Explicit objective criteria for each step (e.g. inclusion)
*Systematic reviews have to be 'truly' systematic

Table 2 Practical approach to a systematic review

• Ask a useful and answerable question
• Before you start, check if it has been answered already!
• Be specific in deciding the type of studies (PICOS) you wish to search for
• Be comprehensive in the literature search
• Get help (Subject experts, Methodologist)
• Avoid the temptation to conduct a 'Meta-analysis' just to impress!
• Don't combine apples with oranges unless assessing 'fruits in general'
• Keep the mindset of a judge and jury (Fair judgement), rather than a lawyer (Make the best case for one side (Baumeister 2013))

importance of an additional aspect—the mindset of a systematic reviewer (Baumeister 2013). The responsibility of systematic reviewers is to provide the comprehensive and best available evidence in the context of current clinical practice and let the reader judge the applicability (safety and efficacy) of the evidence to their patient. Considering human behaviour, it is not uncommon for reviewers to take sides, consciously or subconsciously!

It is important to know what systematic reviews tell us and what they don't. If conducted and reported using a robust methodology, systematic reviews tell us in a scientific, structured, and transparent way as to Who did what, why, and for whom? How? What did they find? What does it mean in the current context? What needs to be done? Systematic reviews do NOT tell what one should do for an individual patient. That process is left to the health care provider and the patient as a shared responsibility.

What Is a Meta-Analysis?

Systematic reviews represent the structured scientific approach for undertaking literature reviews on earlier research studies addressing the desired focussed and properly framed question (PICOS/T). They are tied closely to meta-analyses, i.e. a statistical method for combining the data from the previous studies. A systematic review may or may not contain a meta-analysis depending on whether the data from previous studies addressing the desired question can or cannot be combined. When meta-analysis is possible, it's a systematic review *with* meta-analysis' (i.e. *quantitative* systematic review); otherwise, it is only a systematic review. When meta-analysis is not done for various reasons, the reviewers take a structured descriptive/narrative approach to discuss various aspects of the included studies. Such a systematic review *without* meta-analysis is called as a *qualitative* systematic review.

There no reason why data from different' *more or less similar*' studies answering the desired question cannot be combined using the technique of meta-analysis, however, the evidence from such a meta-analysis will not be reliable if the included studies are not derived by a systematic review. Appreciating the importance of a systematic review for identifying the studies included in a meta-analysis is critical.

It is important to know that meta-analysis can be used to synthesise results from RCTs, as well as non-RCTs ('Observational studies') and epidemiological studies.

A Brief History of Meta-Analyses

The 17th-century French mathematician Blaise Pascal developed methods to determine the value of possible gambles and to compare and combine observations by different astronomers (https://www.biography.com/people/blaise-pascal-9434176) (Table 3). Later, the 18th and 19th-century astronomers and mathematicians such as Gauss and Laplace dealt with the concept of summarising the results from different studies (https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss; https://en.wikipedia.org/wiki/Pierre-Simon_Laplace). These were presented in a book published by the British Royal Astronomer George Biddell Airy (Wright 1988). The British statistician Karl Pearson (1904) is considered to be the first person to combine observations from different studies using special methods (http://adsabs.harvard.edu/full/; Shannon 2008; Pearson 1900). Pearson compared infection and mortality among soldiers who had volunteered for vaccination against typhoid fever with those who had not volunteered. It is remarkable that he commented not only on the 'significance' of results, irregularity of correlation (i.e. heterogeneity) between vaccination and mortality, and the 'lowness' of the values (poor efficacy) reflecting the need for a better vaccine but also on the need for a better method (direction for further research) to get unbiased results (http://adsabs.harvard.edu/full/; Shannon 2008; Pearson 1900). Sir Ronald Aylmer Fisher (1890–1962), the famous English statistician and biologist who used mathematics to combine Mendelian genetics and natural selection, developed the combined probability test for combining data the, i.e. conducting "meta-analysis" (analysis of analyses) (Pearson 1904; Ronald Fisher 2017; Fisher 1925). The test is used to combine the results from several independent tests bearing upon the same overall hypothesis (H0). The credit for coining the term 'meta-analysis' is given to Gene Glass, an American statistician and researcher in educational psychology and social sciences (Mosteller and Fisher 1948; Gene 2017; Glass 1976). It is said that he used the term for the first time in his presidential address to the American Educational Research Association in San Francisco in April, 1976 (Mosteller and Fisher 1948).

Table 3 Meta-analysis–the historical milestones

- K. Pearson 1904: The effects of a vaccine against typhoid. (11 studies)
- Fisher RA 1932: Statistical Methods for Research Workers. London: Oliver & Boyd
- Gene Glass 1976: Coined the phrase "Meta-analysis"
- First book 1981: Meta-Analysis in Social Research. Beverly Hills, CA: Sage. GV Glass, B McGraw and ML Smith
- Cochrane Collaboration: 1993: Iain Chalmers

One of the earliest books on meta-analysis is said to have been published in 1981. Subsequent statisticians have contributed to further development of the methods for meta-analysis. The most recent milestone in the journey of EBM is the development of the Cochrane Collaboration devoted to systematic reviews and meta-analyses of clinical studies to guide clinical practice and research.

The next chapters in this book are devoted to the various steps in systematic reviews and meta-analysis of RCTs. Except for a few differences, the principles for systematic reviews and meta-analysis of non-RCTs, diagnostic studies, and animal studies are similar to those for RCTs. We have covered the essentials of the methodology for systematic reviews of these three different types of studies as a detailed discussion on them is beyond the scope of this book.

References

Alm G. Monography of Swedish fresh water ostracoda along with the systematic review of Tribus Podocopa. Zoologiska Bidrag Från Uppsala. 1916; 4. 1–248.

Anderson JD. David Sackett 1934–2015: the father of evidence-based medicine. Int J Prosthodont. 2015;28:343–4.

Baumeister RF (2013) Writing a literature review. In: Prinstein MJ, Patterson MD (editors) The portable mentor: Expert guide to a successful career in psychology, pp. 119–132, 2nd ed.). New York: Springer Science + Business Media.

Blaise Pascal Biography.com: Philosopher, Mathematician, Theologian, Physicist, Scientist (1623–1662). https://www.biography.com/people/blaise-pascal-9434176. Accessed 11 June 2017.

Brent Thoma. A review of systematic reviews in knowledge translation, January 27, 2013. Caandiem. https://canadiem.org/a-review-of-systematic-reviews/. Accessed 7 July 2017.

Carl Friedrich Gauss: https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss. Accessed 11 June 2017.

Chalmers I, Altman DG, editors. Systematic reviews. London: BMJ Publishing Group; 1995. ISBN 0-7279-0904-5.

Chinchilli VM. General principles for systematic reviews and meta-analyses and a critique of a recent systematic review of long-acting beta-agonists. J Allergy Clin Immunol. 2007;119:303–6.

Cipriani A, Geddes J. Comparison of systematic and narrative reviews: the example of the atypical antipsychotics. Epidemiol Psichiatr Soc. 2003;12:146–53.

Cochrane AL. Effectiveness and efficiency: random reflections on health services. London: Royal Society of Medicine Press; 1972.

Douglas Altman, EQUATOR Workshop 2013 (http://www.equator-network.org/wp-content/uploads/2013/10/Reporting-workshop-SR-principles-Oct2013-v4.pdf). Accessed 25 April 2017.

Fisher RA (1925). Statistical methods for research workers. Oliver and Boyd (Edinburgh). ISBN 0-05-002170-2.

Gene V Glass: https://en.wikipedia.org/wiki/Gene_V._Glass Accessed on 6/11/2017.

Glass GV. Primary, secondary, and meta-analysis of research. Educ Res. 1976; 5 (10):3–8. http://nutrigen.ph.ucla.edu/files/view/epi-m258-spring-2012/Glass.pdf Accessed 11 June 2017.

Glass GV, McGraw B, Smith ML. Meta-analysis in social research. Beverly Hills, CA: Sage; 1981.

Glasziou P, Vanderbroucke J, Chalmers I. Assessing the quality of research. BMJ. 2004;328:39–41.

Global Firsts and Facts. Inventions Discoveries and More. http://globalfirstsandfacts.com/2017/09/16/linus-carl-pauling-first-person-to-receive-nobel-prize-in-two-different-categories/. Accessed 11 July 2017.

Guyatt GH, Haynes RB, Jaeschke RZ, et al. Users' Guides to the medical literature: XXV evidence-based medicine: principles for applying the Users' Guides to patient care, Evidence Based Medicine Working Group. JAMA. 2000; 284:1290–1296.

Guyatt G, Rennie D, Meade MO, Cook DJ. User's guide to the medical literature: a manual for evidence-based clinical practice. 2nd ed. London: AMA; 2002.

History of Systematic Reviews. EPPI: Evidence-informed policy and practice, https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=68. Accessed 7 July 2017.

Linus Planning. How to live longer and feel better, 30th Anniversary Edition, May 1st 2006. Amazon Books. https://www.amazon.com/How-Live-Longer-Feel-Better/dp/0716717816/ref=tmm_hrd_swatch_0?_encoding=UTF8&qid=&sr=. Accessed 11 June 2017.

Isaacs D, Fitzgerald D. Seven alternatives to evidence based medicine. BMJ. 1999;319:1618.

Knipschild P. Some examples of systematic reviews. In: Chalmers I, Altman D, editors. Systematic reviews. London: BMJ Publishing Group; 1995.

Malletta R, Hagen-Zankerb J, Slaterc R, Duvendackd M. The benefits and challenges of using systematic reviews in international development research. J Dev Eff. 2012;4:445–55.

Mees GF. A systematic review of the Indo-Australian zosteropidae. Leiden: Brill; 1957.

Mosteller F, Fisher RA. Questions and answers. Am Stat. 1948; 2: 30–31. http://www.jstor.org/stable/2681650 https://doi.org/10.2307/2681650.

Oakley A, Gough D, Oliver S, Thomas J. The politics of evidence and methodology. Evid Policy. 2005;1:5–31.

O'Rourke K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. J R Soc Med. 2007;100:579–82.

Oxman AD, Guyatt GH. Guidelines for reading literature reviews. CMAJ. 1988;138:697–703.

Pae CU. Why systematic review rather than narrative review? Psychiatry Invest. 2015;12:417–9.

Pearson K. Report on certain enteric fever inoculation statistics. BMJ. 1904;3:1243–6.

Pearson K. Mathematical contributions to the theory of evolution: VII On the correlation of characters not quantitatively measurable. Philos Trans R Soc Lond (Series A, containing Papers of a Mathematical or Physical Character). 1900; 195:1–47.

Petticrew M. Systematic reviews from astronomy to zoology: myths and misconceptions. BMJ. 2001;322:98–101.

Pierre-Simon Laplace: https://en.wikipedia.org/wiki/Pierre-Simon_Laplace. Accessed 11 June 2017.

Ronald Fisher: https://en.wikipedia.org/wiki/Ronald_Fisher. Accessed 11 June 2017.

Sackett DL. Evidence-based medicine. Semin Perinatol. 1997;21:3–5.

Schlosser RW. The role of systematic reviews in evidence-based practice, research, and development. Focus-A Publication of the National Center for the Dissemination of Disability Research (NCDDR) 2006; Technical Brief No 15: 1–4.

Sense about science because evidence matters: Evidence based medicine. http://www.senseaboutscience.org/pages/evidence-based-medicine.html.

Shannon H (2008). A statistical note on Karl Pearson's 1904 meta-analysis. JLL Bull. Commentaries on the history of treatment evaluation (http://www.jameslindlibrary.org/articles/a-statistical-note-on-karl-pearsons-1904-meta-analysis/).

The Cochrane Collaboration: About us; our name. http://www.cochrane.org/about-us/our-name. Accessed 11 June 2017.

Wright DC. The published works of sir George Biddell Airy. J Br Astro Assoc. 1988;98:355–61. http://adsabs.harvard.edu/full/1988JBAA…98..355W. Accessed 11 June 2017.

Literature Search for Systematic Reviews

Shripada Rao and Kwi Moon

Abstract A thorough literature search is an essential step in the conduct of systematic reviews. Inadequate literature search can adversely influence the results and conclusions of a systematic review. Important databases to be searched are Medline, EMBASE, Cumulative Index to Nursing and Allied Health Literature (CINAHL), EmCare (Nursing and Allied Health), Cochrane Library, Clinical Trial Registries, conference proceedings and Grey Literature. Majority of the databases can be accessed freely via the internet except for EMBASE, CINAHL and Emcare, which are subscription-based. Boolean operators AND, OR and NOT are used to identify relevant articles for systematic reviews. Searching of at least two major databases is the minimum prerequisite. However, it is better to search as many databases as possible. At least two reviewers should independently perform the literature search. Citation managers such as Endnote, Mendeley, Zotero and Reference Manager are useful in searching, organising, and sharing the literature. This chapter covers the strategy for a comprehensive literature search with optimal transparency and reproducibility.

Keywords Boolean operator · Cochrane · CINAHL · Databases · EMBASE · EMCARE · Medline · Search · Strategy · Systematic review

S. Rao (✉)
School of Medicine, Neonatal Directorate, Perth Children’s Hospital,
University of Western Australia, 15 Hospital Ave, Nedlands, Perth,
WA 6008, Australia
e-mail: shripada.rao@health.wa.gov.au

K. Moon
Department of Pharmacy, Perth Children’s Hospital, 15 Hospital Ave, Nedlands,
Perth, WA 6008, Australia
e-mail: kwi.moon@health.wa.gov.au

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_2

Introduction

A thorough literature search is an essential step in systematic reviews. Inadequate literature search can adversely influence the results and conclusions of a systematic review. Important databases that need to be searched are Medline, EMBASE, Cumulative Index to Nursing and Allied Health Literature (CINAHL), EmCare (Nursing and Allied Health), Cochrane Library, Clinical Trial Registries and Grey Literature.

MEDLINE

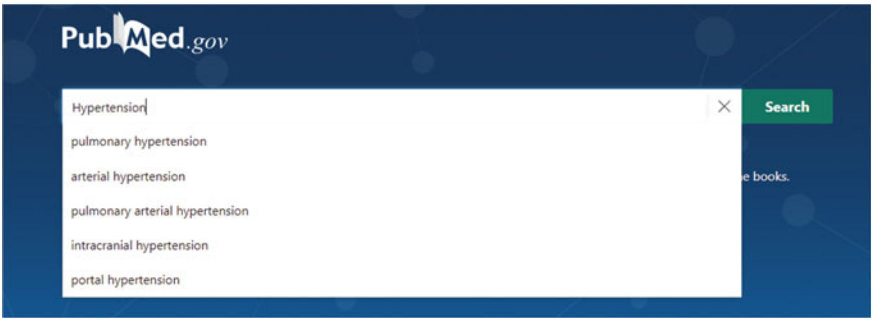
MEDLINE is the most widely used database and contains over 24 million citations from more than 5600 biomedical journals dating back to 1946. Since 1996, MEDLINE is made freely available via PubMed from the National Library of Medicine, USA. While many people use the terms Medline and PubMed interchangeably, it is essential to know that they are not the same. PubMed is the gateway to Medline, i.e. Medline can be searched via PubMed. MEDLINE can also be accessed via commercial platforms such as Ovid.

MEDLINE uses a controlled vocabulary called The Medical Subject Headings (MeSH®) for indexing journal articles. The MeSH terms are arranged in a hierarchical categorised manner called MeSH Tree Structures and updated annually. Familiarity with this vocabulary will enable optimal searching of PubMed.

Searching MEDLINE Through PubMed

PubMed can be searched using basic search strategy by typing in simple terms in the search box and clicking the ‘search’ icon (Fatehi et al. 2013). While typing the search term in the search box, an autocomplete feature will suggest relevant terms, which appear as a list from which one could select, instead of typing the complete term (Fig. 1). By default, the results are sorted by the date added to PubMed and displayed with 20 citations on each page (Fatehi et al. 2013). It could be changed to up to 200 citations per page (Fig. 2). In addition to showing the author/s, institution, journal, date and year of publication and the abstract, the left-hand side of the results page will show filters that can be applied to narrow the search based on the type of study (i.e. study design), year of publication, human vs animal studies and many other options (Fig. 3).

On the left-hand side of the results page also shows a histogram showing the trend of articles during the past years (Figs. 2 and 3). Once the “Advanced” link is clicked, it takes you to the “History and Search details”. If you click on “details”, it



Hypertension

pulmonary hypertension

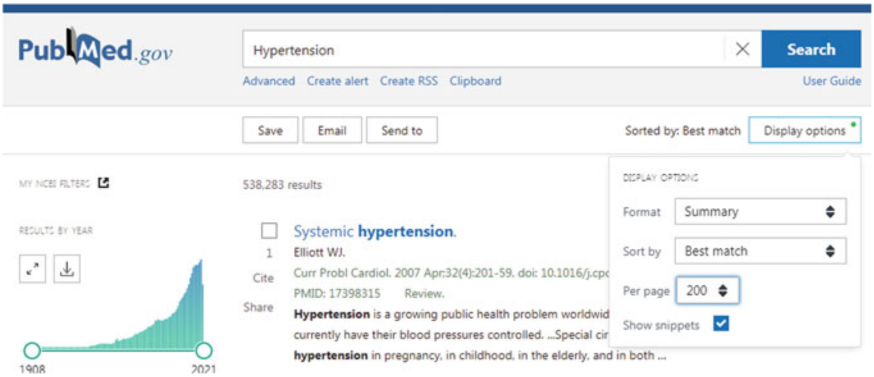
arterial hypertension

pulmonary arterial hypertension

intracranial hypertension

portal hypertension

Fig. 1 While searching PubMed with simple terms, the auto-search feature provides options to select relevant terms/phrases



Hypertension

Sorted by: Best match

538,283 results

□ 1 Systemic **hypertension**.
Elliott WJ.
Curr Probl Cardiol. 2007 Apr;32(4):201-59. doi: 10.1016/j.cpc
PMID: 17398315 Review.
Hypertension is a growing public health problem worldwid
currently have their blood pressures controlled. ...Special cir
hypertension in pregnancy, in childhood, in the elderly, and in both ...

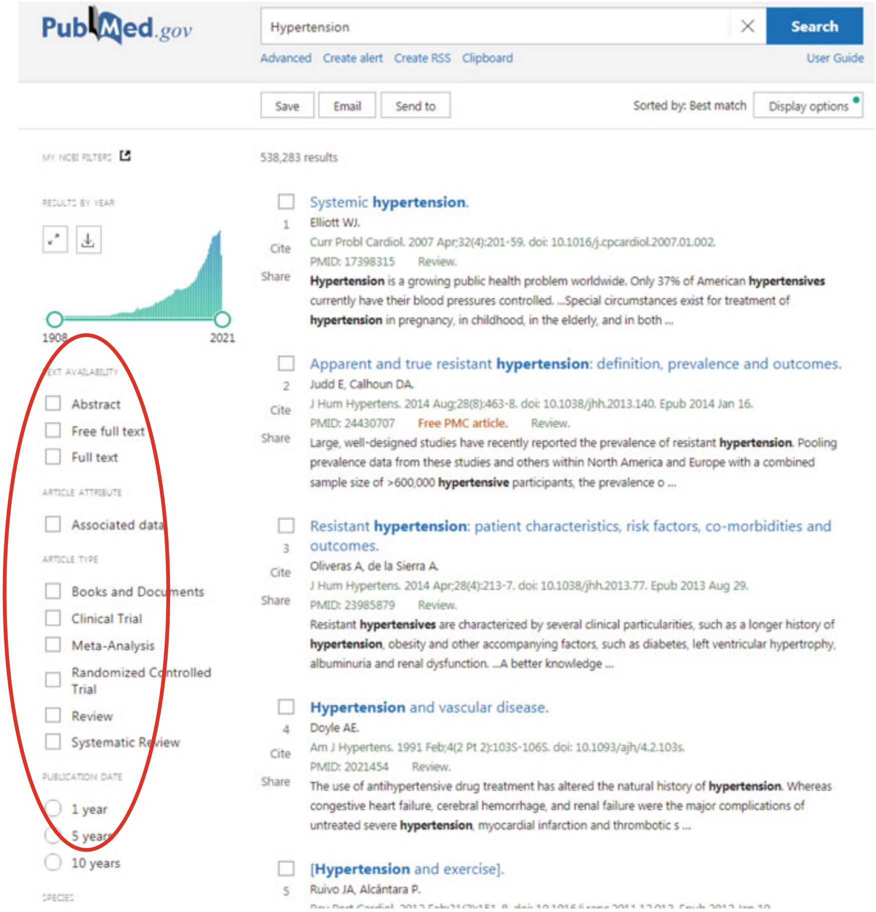
Format: Summary
Sort by: Best match
Per page: 200
Show snippets ☑

Fig. 2 Up to 200 citations can be displayed per page on PubMed

shows how PubMed has translated the search query, and reviewing this information will enable the searcher to modify a query to include or exclude specific terms, if required (Fig. 4).

Searching PubMed using Boolean operators: PubMed accepts Boolean operators AND, OR and NOT. They could be typed directly in the search box on the main screen or in the advanced search builder. For example, if you want to search articles on probiotics in preterm infants, you need to type Probiotics AND Preterm Infant, so that articles, where both terms are present, will be retrieved (Figs. 5and 6). If you want to search for articles about Vitamin D, you need to type Vitamin D OR Cholecalciferol OR Ergocalciferol to ensure all articles on this topic are retrieved (Fig. 7). If you want to search for articles on Hypertension, but not interested in Pulmonary Hypertension, you could type Hypertension NOT Pulmonary Hypertension.

Incorrect use of the Boolean operator results in the retrieval of irrelevant citations. For example, while searching for articles on probiotics in preterm infants, if



538,283 results

□ 1 Systemic **hypertension**.
Elliott WJ.
Curr Probl Cardiol. 2007 Apr;32(4):201-59. doi: 10.1016/j.cpcardiol.2007.01.002.
PMID: 17398315 Review.
Hypertension is a growing public health problem worldwide. Only 37% of American **hypertensives** currently have their blood pressures controlled. ...Special circumstances exist for treatment of **hypertension** in pregnancy, in childhood, in the elderly, and in both ...

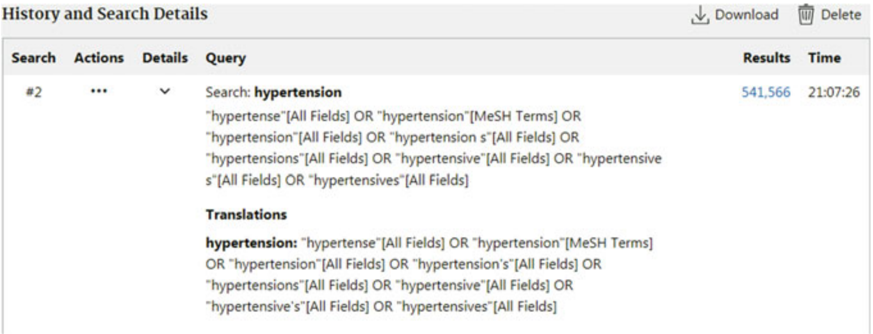
□ 2 Apparent and true resistant **hypertension**: definition, prevalence and outcomes.
Judd E, Calhoun DA.
J Hum Hypertens. 2014 Aug;28(8):463-8. doi: 10.1038/jhh.2013.140. Epub 2014 Jan 16.
PMID: 24430707 Free PMC article. Review.
Large, well-designed studies have recently reported the prevalence of resistant **hypertension**. Pooling prevalence data from these studies and others within North America and Europe with a combined sample size of >600,000 **hypertensive** participants, the prevalence o ...

□ 3 Resistant **hypertension**: patient characteristics, risk factors, co-morbidities and outcomes.
Oliveras A, de la Sierra A.
J Hum Hypertens. 2014 Apr;28(4):213-7. doi: 10.1038/jhh.2013.77. Epub 2013 Aug 29.
PMID: 23985879 Review.
Resistant **hypertensives** are characterized by several clinical particularities, such as a longer history of **hypertension**, obesity and other accompanying factors, such as diabetes, left ventricular hypertrophy, albuminuria and renal dysfunction. ...A better knowledge ...

□ 4 **Hypertension** and vascular disease.
Doyle AE.
Am J Hypertens. 1991 Feb;4(2 Pt 2):103S-106S. doi: 10.1093/ajh/4.2.103s.
PMID: 2021454 Review.
The use of antihypertensive drug treatment has altered the natural history of **hypertension**. Whereas congestive heart failure, cerebral hemorrhage, and renal failure were the major complications of untreated severe **hypertension**, myocardial infarction and thrombotic s ...

□ 5 [**Hypertension** and exercise].
Ruivo JA, Alcântara P.
Rev Port Cardiol. 2012 Feb;31(2):151-8. doi: 10.1016/j.repc.2011.12.012. Epub 2012 Jan 10.

Fig. 3 Various filters can be applied while searching PubMed



History and Search Details

Search	Actions	Details	Query	Results	Time
#2	•••	⌄	Search: **hypertension** "hypertense"[All Fields] OR "hypertension"[MeSH Terms] OR "hypertension"[All Fields] OR "hypertension s"[All Fields] OR "hypertensions"[All Fields] OR "hypertensive"[All Fields] OR "hypertensive s"[All Fields] OR "hypertensives"[All Fields] **Translations** **hypertension:** "hypertense"[All Fields] OR "hypertension"[MeSH Terms] OR "hypertension"[All Fields] OR "hypertension's"[All Fields] OR "hypertensions"[All Fields] OR "hypertensive"[All Fields] OR "hypertensive's"[All Fields] OR "hypertensives"[All Fields]	541,566	21:07:26

Fig. 4 History and search details: Shows how PubMed has translated the search query

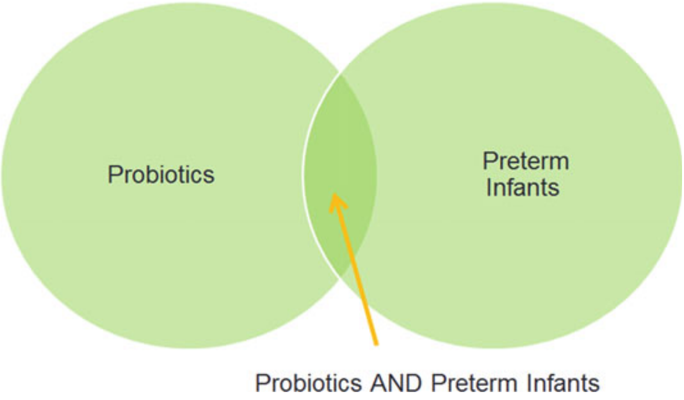
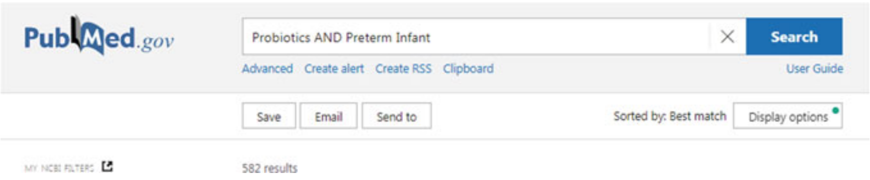


Fig. 5 Search using the Boolean operator AND



Probiotics AND Preterm Infant

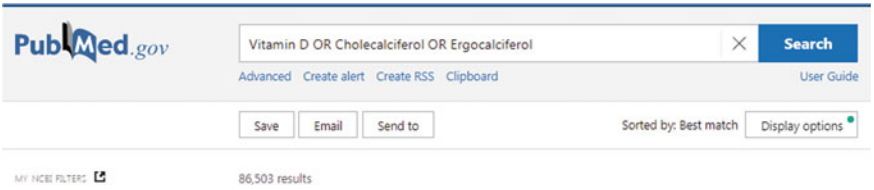
582 results

Fig. 6 Search using the Boolean operator OR

you type in Probiotics OR Preterm infant instead of Probiotics AND Preterm infant, you will end up with more than 100000 non-relevant citations instead of the correct 411 citations.

Searching PubMed using singular vs plural words: Some publications use singular words, some plural, and hence it is essential to search with both terms. For example, the phrase Drug retrieved 5,760,287 citations (Fig. 8a), whereas the plural phrase Drugs retrieved only 1,543,528 citations (Fig. 8b). A search using both terms with an OR in between yielded 6254560 citations.

Searching using "PubMed Advanced Search Builder": The advanced search builder is used for highly focussed searches. For example, you may want to search



Vitamin D OR Cholecalciferol OR Ergocalciferol

86,503 results

Fig. 7 Use of the Boolean Operator OR

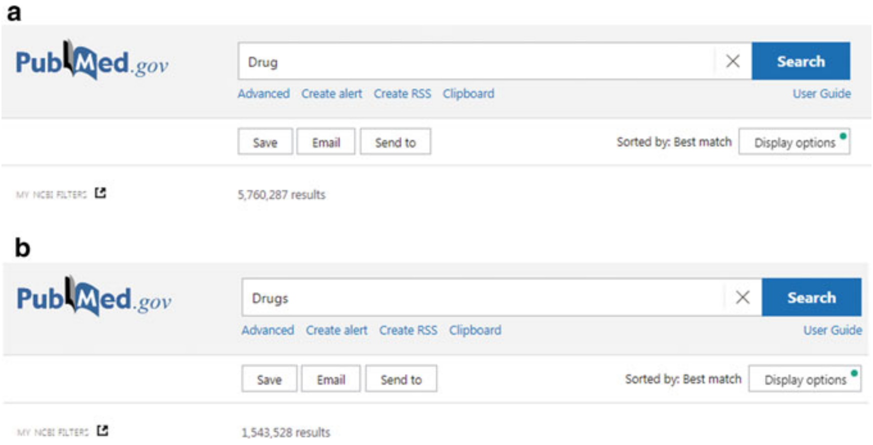


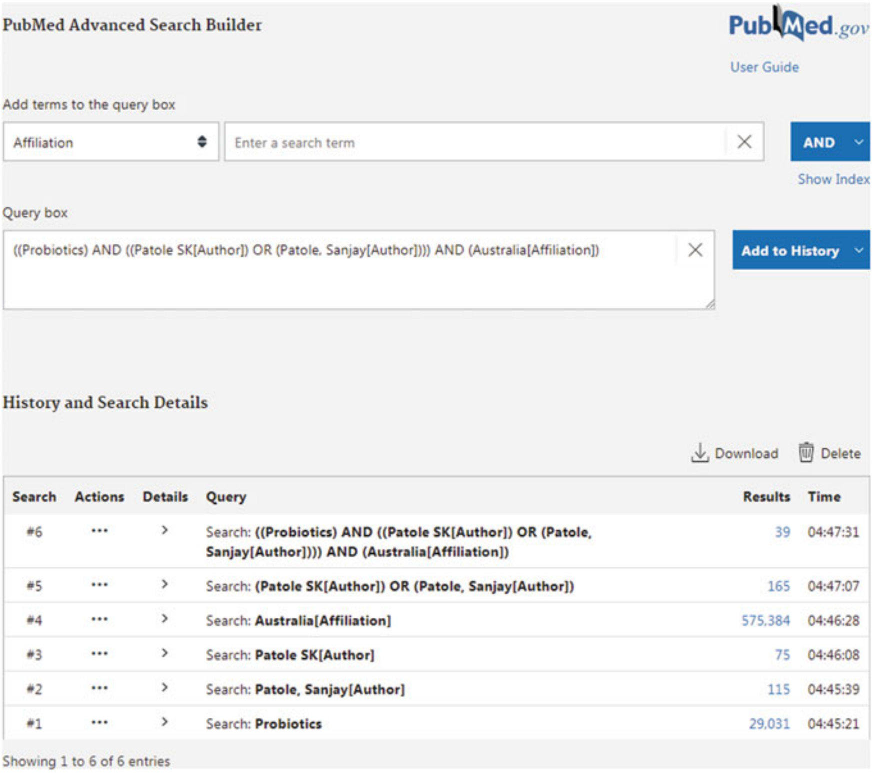
Fig. 8 **a** Singular versus plural word while searching PubMed. **b** Singular versus plural word while searching PubMed

for articles on Omeprazole, where the name Omeprazole is present in the title of the article. There are at least 41 fields that can be utilised to build a search strategy. For example, one could search for articles on probiotics Sanjay Patole from Australia by using the advanced search builder, which will yield 39 relevant citations (Fig. 9).

Searching PubMed for articles with keywords that are present in title only, title/abstract or all fields: Through advanced search builder one could search for articles that have the keyword in the title or title/abstract or any area in the manuscript (Fig. 10). As one would expect, searching for articles where the keyword is present in the title itself will yield less number of citations and miss relevant citations, but the ones identified are expected to be highly relevant. On the other hand, searching for the keyword in any field would result in the maximum number of retrievals, many of which may not be relevant. Hence the systematic reviewers need to be aware of the trade-off between high sensitivity vs high specificity between these strategies.

Searching using MeSH terms: Since different authors use different terminologies for the same concept (e.g. “preterm infant” and “premature infant” are often used interchangeably), a standard vocabulary system is needed to enable retrieval of articles that have used either of these terminologies. The MeSH database provides a controlled vocabulary and index terms (Lowe and Barnett 1994). MeSH database can be searched from PubMed by clicking the MeSH database at the bottom right-hand corner of the PubMed screen (Fig. 11). For example, you will notice that the MeSH term for a preterm infant is “Infant, Premature” OR “Infant, Extremely Premature”. It means, by using these MeSH terms, you will be able to retrieve articles that have used the term “Preterm Infant”.

Searching PubMed using MeSH terms may not reveal the latest articles because they have not yet been indexed using those terms. This is because skilled subject analysts at the National Library of Medicine (NLM) regularly examine journal


PubMed Advanced Search Builder

Add terms to the query box

Affiliation | Enter a search term

Query box

((Probiotics) AND ((Patole SK[Author]) OR (Patole, Sanjay[Author]))) AND (Australia[Affiliation])

History and Search Details

Search	Actions	Details	Query	Results	Time
#6	•••	>	Search: **((Probiotics) AND ((Patole SK[Author]) OR (Patole, Sanjay[Author]))) AND (Australia[Affiliation])**	39	04:47:31
#5	•••	>	Search: **(Patole SK[Author]) OR (Patole, Sanjay[Author])**	165	04:47:07
#4	•••	>	Search: **Australia[Affiliation]**	575,384	04:46:28
#3	•••	>	Search: **Patole SK[Author]**	75	04:46:08
#2	•••	>	Search: **Patole, Sanjay[Author]**	115	04:45:39
#1	•••	>	Search: **Probiotics**	29,031	04:45:21

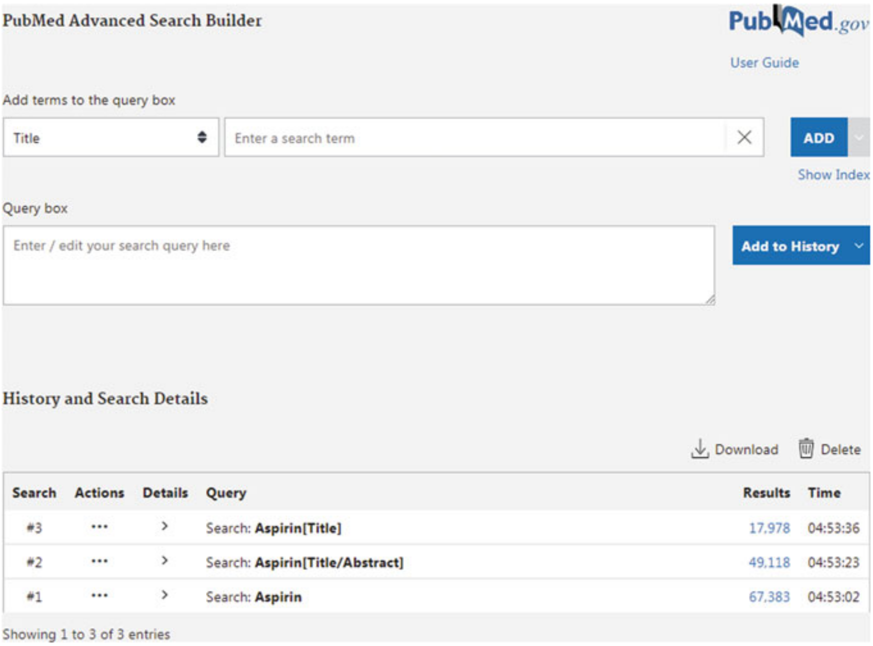
Showing 1 to 6 of 6 entries

Fig. 9 Searching PubMed using Advanced Search builder

articles and assign to each the most specific MeSH terms applicable - typically ten to twelve. Until such allocation, searching using MeSH term will not pick up those articles. It is better to search PubMed using both keywords and MeSH terms to ensure optimal results.

Use of double quotes on PubMed search: If you enclose a search term using double quotes, e.g. "Heart Attack" while searching PubMed, only 3895 citations are retrieved, whereas a search without the double quotes yields 233902 citations (Fig. 12). This is because enclosing a search term or phrase in double-quotes turns off automatic term mapping (ATM) capability of PubMed to search for other relevant terms (e.g. Myocardial Infarction). Hence PubMed does not recommend the use of double-quotes.

Finding related citations: Once a highly relevant article is found, one could click on the "similar articles" (Fig. 13) link to identify them. PubMed uses a robust word-weighted algorithm to compare words from the Title and Abstract of each citation, as well as the MeSH headings assigned. The best matches for each citation are pre-calculated and stored as a set (https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_190.html).



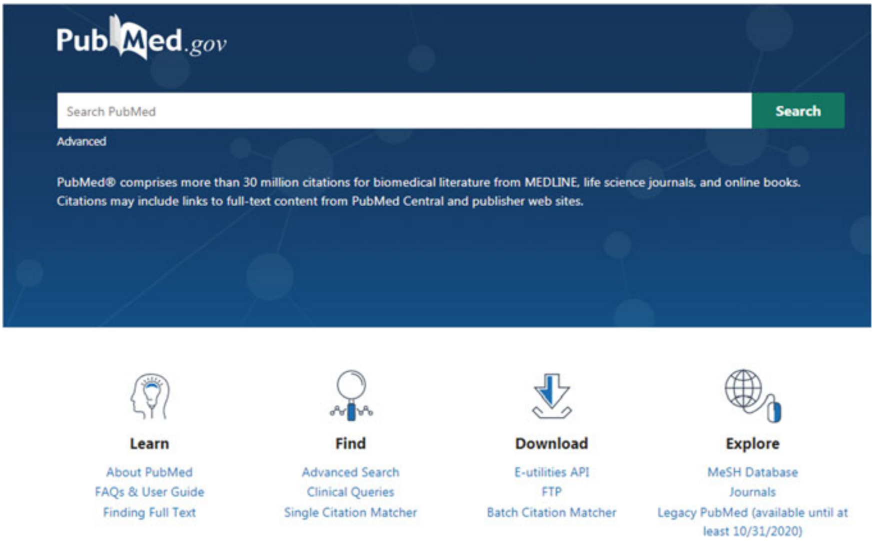
PubMed Advanced Search Builder

History and Search Details

Search	Actions	Details	Query	Results	Time
#3	•••	>	Search: **Aspirin[Title]**	17,978	04:53:36
#2	•••	>	Search: **Aspirin[Title/Abstract]**	49,118	04:53:23
#1	•••	>	Search: **Aspirin**	67,383	04:53:02

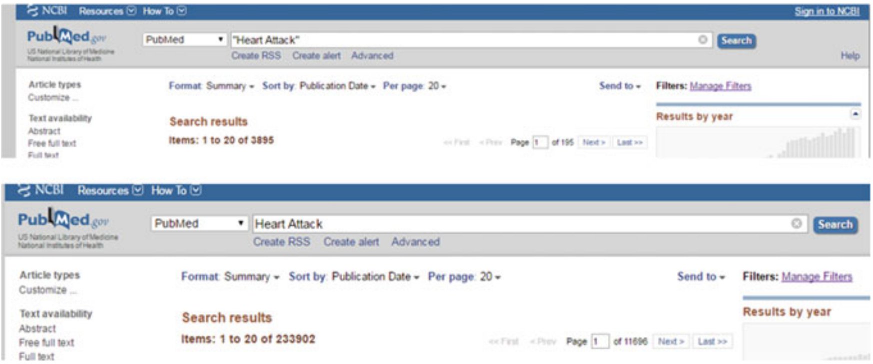
Showing 1 to 3 of 3 entries

Fig. 10 Searching all fields vs Title/Abstract vs Title



PubMed® comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Fig. 11 Searching PubMed through MeSH database



"Heart Attack"

Search results

Items: 1 to 20 of 3895

Page 1 of 195

Heart Attack

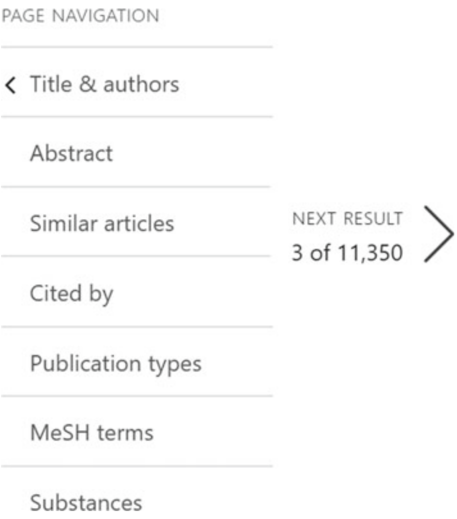
Search results

Items: 1 to 20 of 233902

Page 1 of 11696

Fig. 12 Searching PubMed using terms with and without double quotes

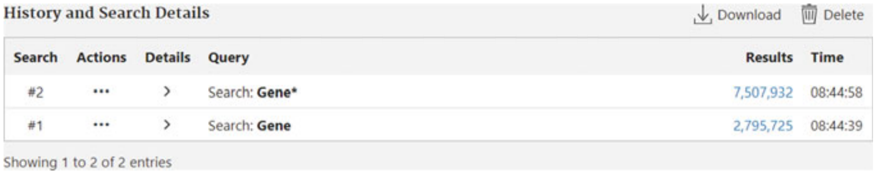
Fig. 13 Finding similar articles



Searching using truncations/wildcards: Asterisk (*) is the wildcard symbol for PubMed.

To search for all terms that begin with a word, enter the word followed by an asterisk (*): the wildcard character. For example, if the truncated search term gene* is used, PubMed will search for articles containing any of the terms such as gene, genetics, gene therapy and many more. The use of truncation and wild card inhibits the efficient automatic mapping capability of PubMed for that particular search. Moreover, truncations result in the search strategy becoming unmanageable; for example, a search using Gene* results nearly 7.5 million articles instead of 2.8 million (Fig. 14).

Applying filters: While searching PubMed, you could apply filters to retrieve highly relevant citations. The available filters are article types (e.g. clinical trials,



History and Search Details

Search	Actions	Details	Query	Results	Time
#2	•••	>	Search: **Gene***	7,507,932	08:44:58
#1	•••	>	Search: **Gene**	2,795,725	08:44:39

Showing 1 to 2 of 2 entries

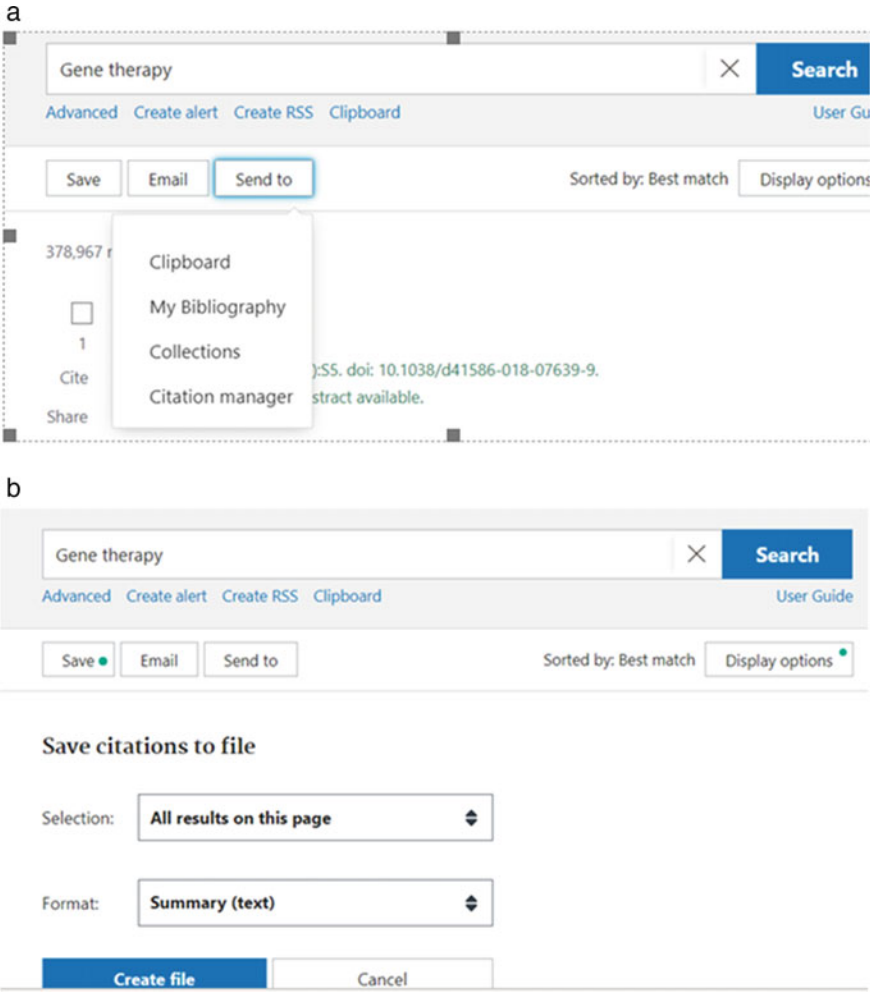
Fig. 14 Searching PubMed with truncations and wildcards

systematic reviews, letters to the editor, meta-analyses), availability of free full texts, age of the study population in the paper, studies in humans, animal studies, publication dates and many other options (Fig. 3).

Obtaining full-text articles: PubMed does not store full-text articles but provides a link to the publisher's website through which one could purchase the article for a fee. Articles are often provided free of cost by the publisher or have already been deposited in the PMC. Institutional libraries may have subscriptions to those journals, and hence full texts may be available to download free of cost to the searcher.

Saving PubMed searches: There are various options for saving the searches: (a). While searching, if articles of interest are identified, they could be added to the clipboard (Fig. 15a), where they will be available for eight hours. A maximum of 500 citations can be temporarily saved here; (b). The citations could be saved as text-files or to endnote by clicking on the "save" icon or send to "citation manager" button respectively (Fig. 15b). The citations could be e-mailed to self or anyone, by clicking on the "E-mail" icon (Fig. 15a and b). A handy approach is to send to "collections" and create a cost-free NCBI account and save them as "my collections" or my bibliography on the NCBI account. Follow the simple steps on this website to create your own NCBI account (https://www.ncbi.nlm.nih.gov/account/?back_url=https%3A%2F%2Fwww.ncbi.nlm.nih.gov%2Fpubmed) (Fig. 16). Once created, the search terms and results will be stored for the long term and can be re-run in the future (Fig. 17). The advantages of having an NCBI account are described in Table 1.

PubMed Central® (PMC): It is a free archive of biomedical and life sciences PubMed Central journal literature at the US (Fig. 18) National Institutes of Health's National Library of Medicine (NIH-NLM). In keeping with NLM's legislative mandate to collect and preserve the biomedical literature, PMC serves as a digital counterpart to NLM's extensive print journal collection. Launched in February 2000, PMC was developed and is managed by NLM's National Center for Biotechnology Information (NCBI). PMC contains 6.3 million free full-text articles, most of which have a corresponding entry in PubMed. PMC is a repository for journal literature deposited by participating publishers, as well as for author manuscripts that have been submitted in compliance with the NIH Public Access Policy and similar policies of other research-funding agencies. Some PMC journals are also MEDLINE journals (Shashikiran 2016).



Gene therapy

Save citations to file

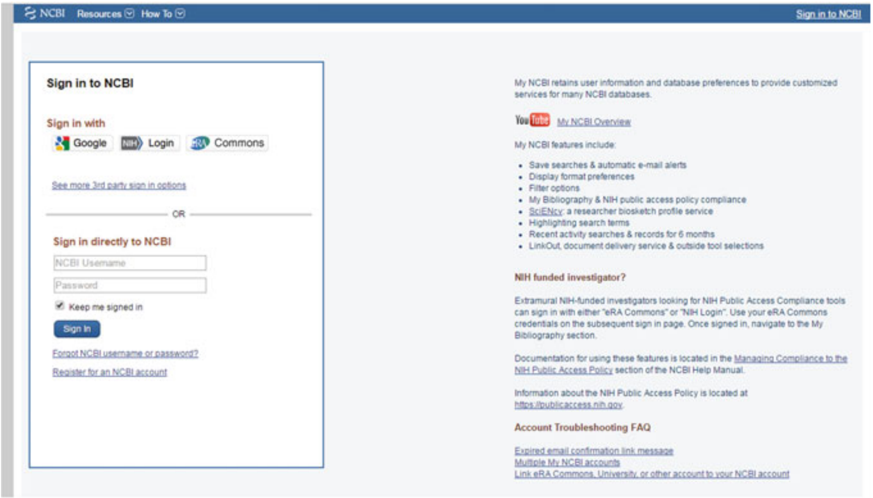
Selection: All results on this page

Format: Summary (text)

Fig. 15 **a** Saving PubMed search results. **b** Saving PubMed search results

Embase (Excerpta Medica Database)

Excerpta Medica Database is a biomedical database published by Elsevier (https://www.elsevier.com/solutions/embase-biomedical-research). It covers the most important international biomedical literature from 1947 to the present day. All articles are indexed in-depth using Elsevier's Life Science thesaurus Embase Indexing and Emtree®. The entire database is also available on platforms such as Ovid. Through Embase, one could search over 32 million records, including MEDLINE titles, over 8,500 journals from over 95 countries, including MEDLINE titles, over 2,900 indexed journals unique to Embase, over 1.5 million records



Sign in to NCBI

Sign in with

See more 3rd party sign in options

OR

Sign in directly to NCBI

NCBI Username

Password

☑ Keep me signed in

Sign In

Forgot NCBI username or password?

Register for an NCBI account

My NCBI retains user information and database preferences to provide customized services for many NCBI databases.

YouTube My NCBI Overview

My NCBI features include:

- Save searches & automatic e-mail alerts
- Display format preferences
- Filter options
- My Bibliography & NIH public access policy compliance
- SciENcv: a researcher biosketch profile service
- Highlighting search terms
- Recent activity searches & records for 6 months
- LinkOut, document delivery service & outside tool selections

NIH funded investigator?

Extramural NIH-funded investigators looking for NIH Public Access Compliance tools can sign in with either "eRA Commons" or "NIH Login". Use your eRA Commons credentials on the subsequent sign in page. Once signed in, navigate to the My Bibliography section.

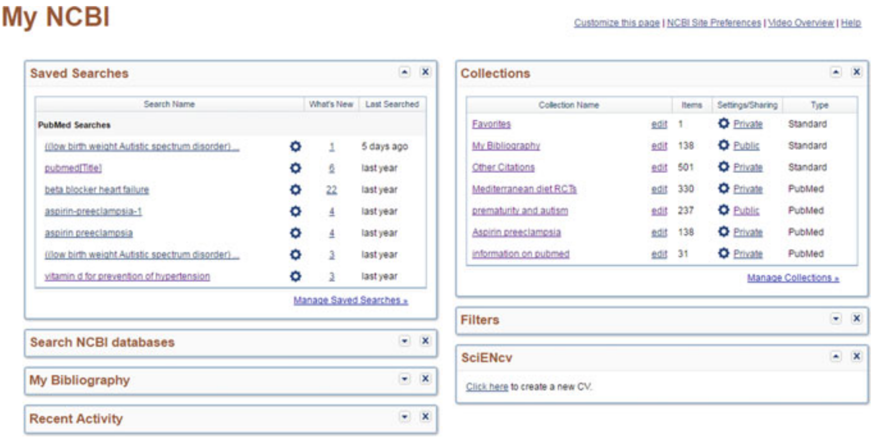
Documentation for using these features is located in the Managing Compliance to the NIH Public Access Policy section of the NCBI Help Manual.

Information about the NIH Public Access Policy is located at https://publicaccess.nih.gov.

Account Troubleshooting FAQ

Expired email confirmation link message
Multiple My NCBI accounts
Link eRA Commons, University, or other account to your NCBI account

Fig. 16 Creating a free NCBI account


My NCBI

Customize this page | NCBI Site Preferences | Video Overview | Help

Saved Searches

Search Name		What's New	Last Searched
PubMed Searches			
((low birth weight Autistic spectrum disorder) ...	⚙	1	5 days ago
pubmed[Title]	⚙	6	last year
beta blocker heart failure	⚙	22	last year
aspirin-preeclampsia-1	⚙	4	last year
aspirin preeclampsia	⚙	4	last year
((low birth weight Autistic spectrum disorder) ...	⚙	3	last year
vitamin d for prevention of hypertension	⚙	3	last year

Manage Saved Searches »

Search NCBI databases

My Bibliography

Recent Activity

Collections

Collection Name		Items	Settings/Sharing	Type
Favorites	edit	1	⚙ Private	Standard
My Bibliography	edit	138	⚙ Public	Standard
Other Citations	edit	501	⚙ Private	Standard
Mediterranean diet RCTs	edit	330	⚙ Private	PubMed
prematurity and autism	edit	237	⚙ Public	PubMed
Aspirin preeclampsia	edit	138	⚙ Private	PubMed
information on pubmed	edit	31	⚙ Private	PubMed

Manage Collections »

Filters

SciENcv

Click here to create a new CV.

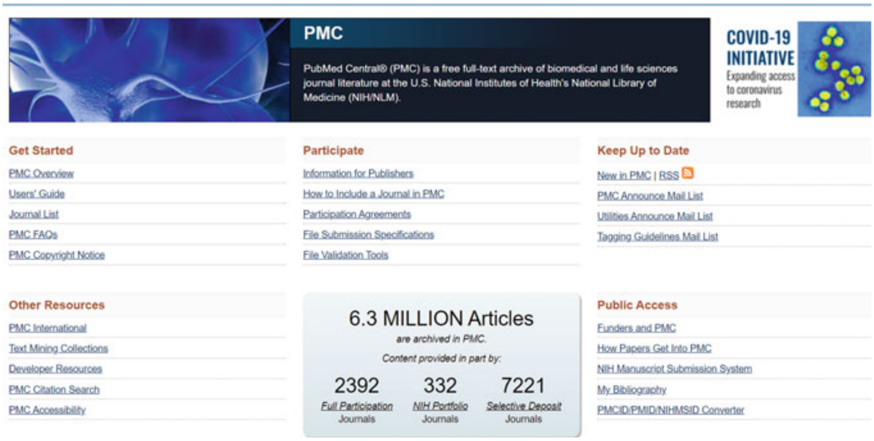
Fig. 17 Saving searches in My NCBI account

added yearly, with an average of over 6,000 each day, and over 2.3 million conference abstracts indexed from more than 7,000 conferences dating from 2009. Embase has full-text indexing of drug, disease and medical device data. The comparison between Medline and Embase is described in Table 2.

Many Universities and hospitals have a subscription to Ovid, through which Embase can be searched (Fig. 19). Once a keyword or concept is written in the search box, the database will provide a list of potentially relevant search terms, of which the relevant ones can be selected and combined with the Boolean operator OR (Fig. 20).

Table 1 Advantages of PubMed search with NCBI account

PubMed search without NCBI account	PubMed search with NCBI account
Search history is stored only for 8 h	Search history is stored for six months
Saving search strategy for the long-term duration not possible	Search strategy can be saved for many years
Creating alerts not possible	Alerts can be created to receive regular e-mails when new relevant studies are published
Since the search strategy is not saved re-running of the search is not possible after 8 h	Since the search strategy is saved re-running of the search can be done even after many years



PMC

PubMed Central® (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM).

6.3 MILLION Articles

are archived in PMC.

Content provided in part by:

2392 Full Participation Journals

332 NIH Portfolio Journals

7221 Selective Deposit Journals

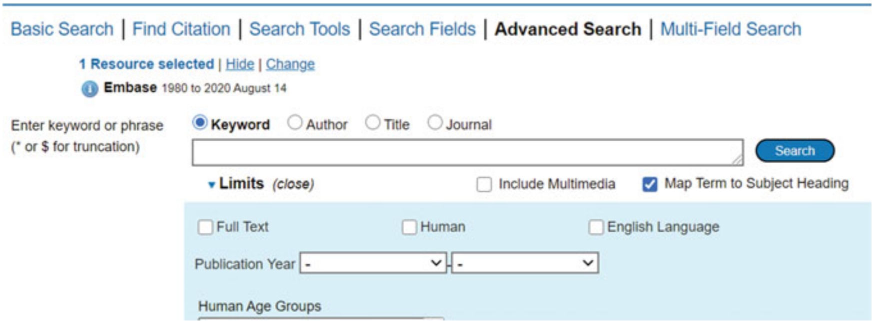
Fig. 18 PubMed Central

While searching Embase, select the “Explode” box if you wish to retrieve results using the selected term and all of its more specific terms. Select the “Focus” box if you wish to limit your search to those documents in which your subject heading is considered the major point of the article (Fig. 20).

Search terms could be combined using the Boolean operator AND, OR (Fig. 21). Similar to Medline, various filters (limits) can be applied based on the design of the study (e.g. clinical trials), age group (children, infants, or adults), animal studies, human studies, year of publication, and many others (Fig. 22). A free Ovid account could be created to enable storage of the search terms and strategy for the long run (Fig. 23). Once the Embase search for the systematic review is completed, it is important to save the search date, terms used, limits applied and the yield by selecting the “export” option on the screen and ticking the “search history” box and pressing “export”.

Table 2 Comparison between Medline and Embase

Database features	Medline	Embase
Focus	Biomedicine and health	Biomedicine and health; drugs and pharmacology
Produced by	US National Library of Medicine	Elsevier
Content	Journal articles	Journal articles plus conference abstracts
# of records	27 million dating back to 1946	32 million records, dating back to 1947
# of journals	5600	8500 journals
Journal origins (2012)	41% North America 49% Europe	34% North America 50% Europe
Conference abstracts	Does not contain conference abstracts	Gives access to some conference abstracts dating back from 2009
Price	Free via PubMed	Needs paid subscription


Basic Search | Find Citation | Search Tools | Search Fields | Advanced Search | Multi-Field Search
Hide | Change
Search
Limits (close)
Human Age Groups

1 Resource selected

Embase 1980 to 2020 August 14

Enter keyword or phrase (* or $ for truncation)

● Keyword ○ Author ○ Title ○ Journal

□ Include Multimedia ☑ Map Term to Subject Heading

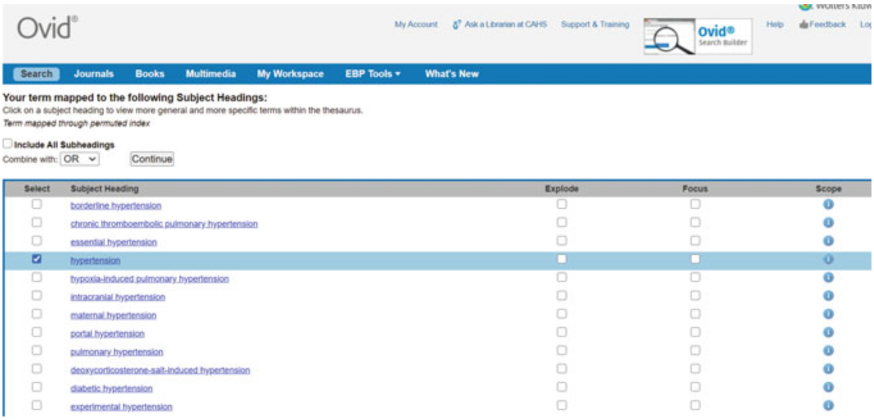
□ Full Text □ Human □ English Language

Publication Year - -

Fig. 19 Searching Embase through Ovid platform

Emcare (Nursing and Allied Health)

Emcare was launched in 2005 by Elsevier to improve the search related to nursing and allied health literature (http://www.ovid.com/site/catalog/databases/14007.jsp; Ulincy 2006). It contains over 3,700 international journals and nearly 5 million records dating back to 1995. 50% of journals from North America, 40% from Europe; 10% from other regions; 9% of all records reference non-English articles, though most have English-language abstracts—70% of records contain online abstracts. It can be accessed via the Ovid platform. The search methodology is similar to searching Embase via Ovid.



Your term mapped to the following Subject Headings:

Click on a subject heading to view more general and more specific terms within the thesaurus.

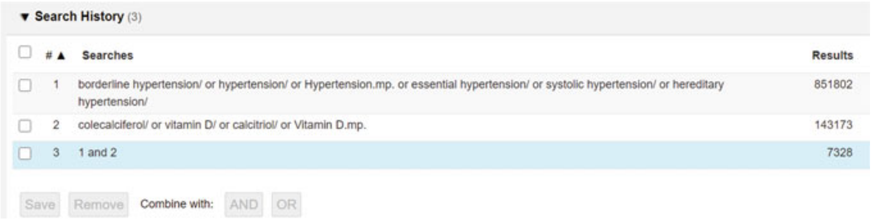
Term mapped through permuted index

☐ **Include All Subheadings**

Combine with: OR ⌄ Continue

Select	Subject Heading	Explode	Focus	Scope
☐	borderline hypertension	☐	☐	●
☐	chronic thromboembolic pulmonary hypertension	☐	☐	●
☐	essential hypertension	☐	☐	●
☑	hypertension	☐	☐	●
☐	hypoxia-induced pulmonary hypertension	☐	☐	●
☐	intracranial hypertension	☐	☐	●
☐	maternal hypertension	☐	☐	●
☐	portal hypertension	☐	☐	●
☐	pulmonary hypertension	☐	☐	●
☐	deoxycorticosterone-salt-induced hypertension	☐	☐	●
☐	diabetic hypertension	☐	☐	●
☐	experimental hypertension	☐	☐	●

Fig. 20 Suggested words for the term “Hypertension” in Embase


▼ **Search History** (3)

☐	# ▲	Searches	Results
☐	1	borderline hypertension/ or hypertension/ or Hypertension.mp. or essential hypertension/ or systolic hypertension/ or hereditary hypertension/	851802
☐	2	colecalciferol/ or vitamin D/ or calcitriol/ or Vitamin D.mp.	143173
☐	3	1 and 2	7328

Save Remove Combine with: AND OR

Fig. 21 Searching Embase using Boolean operators

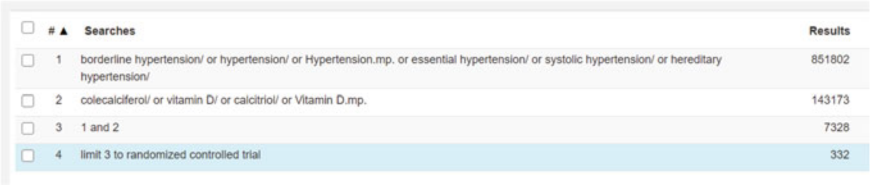
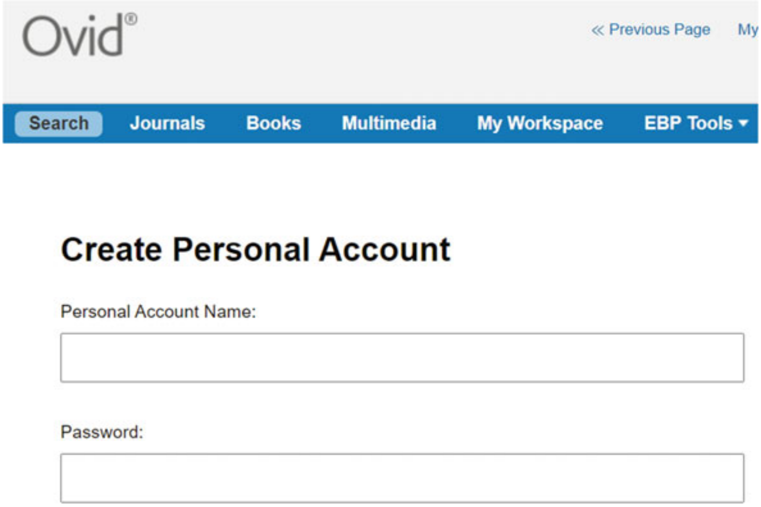

☐	# ▲	Searches	Results
☐	1	borderline hypertension/ or hypertension/ or Hypertension.mp. or essential hypertension/ or systolic hypertension/ or hereditary hypertension/	851802
☐	2	colecalciferol/ or vitamin D/ or calcitriol/ or Vitamin D.mp.	143173
☐	3	1 and 2	7328
☐	4	limit 3 to randomized controlled trial	332

Fig. 22 Applying filters in Embase

Subject coverage	Emcare (%)	CINAHL
Allied Health	35	20
Biomedicine	49	47
Nursing	13	27
Other	3	3

Source Table 1 from Ref. (Ulincy 2006)



Create Personal Account

Personal Account Name:

Password:

Fig. 23 Creating an Ovid account

Cumulative Index to Nursing and Allied Health Literature (CINAHL)

CINAHL is an important database for searching nursing and allied health literature. It contains 3115 journals dating back to 1981 and has various subscription options such as CINAHL, CINAHL Plus, and CINAHL Plus with full text and CINAHL complete. It is published by Ebsco Health (https://www.ebscohost.com/nursing/products/cinahl-databases/cinahl-complete).

Cochrane Library

The Cochrane Library (ISSN 1465–1858) is a collection of six databases that contain different types of high-quality, independent evidence to inform healthcare decision-making and a seventh database that provides information about Cochrane groups. The contents of the Cochrane library are given in the table below (Source: Cochrane Library, Accessed 15 August 2020).

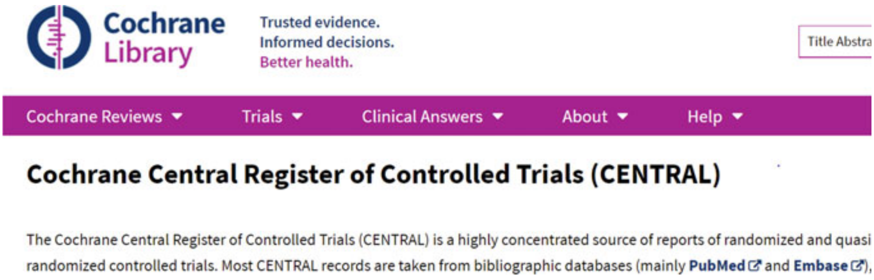
Contents of the Cochrane Library

Cochrane Reviews	8365
Cochrane Protocols	2416
Trials	1663083
Editorials	133
Special collections	37
Clinical answers	2544

Systematic reviewers who are interested in only randomised controlled trials can directly search the Cochrane Central Register of Controlled Trials (CENTRAL) to identify the relevant trials (Fig. 24).

Grey Literature

Despite advances in the dissemination of study information, nearly half of health-related studies go unpublished (Song et al. 2010). Searching such grey literature is important to enhance the reliability of results of systematic reviews. Grey literature is defined as a literature that is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers (Canadian Agency for Drugs Technologies in Health 2015). They are inaccessible via conventional bibliographic databases. Published articles, when compared with grey literature, yield significantly larger estimates of the intervention effect by nearly 15% (McAuley et al. 2000). A Cochrane review on this subject showed similar results and concluded that published trials tend to show an overall greater treatment effect than grey trials (Hopewell et al. 2007). Hence, the exclusion of grey literature can lead to spuriously exaggerated benefits/harms of an intervention. The Canadian Agency for



Cochrane Central Register of Controlled Trials (CENTRAL)

The Cochrane Central Register of Controlled Trials (CENTRAL) is a highly concentrated source of reports of randomized and quasi randomized controlled trials. Most CENTRAL records are taken from bibliographic databases (mainly **PubMed** and **Embase**),

Fig. 24 Searching through the Cochrane Central Register of Controlled Trials (CENTRAL)

Drugs and Technologies in Health has published an excellent guide and resources for searching medical grey literature (Canadian Agency for Drugs Technologies in Health 2015) (https://www.cadth.ca/resources/finding-evidence/grey-matters; accessed on 15 August 2020). It provides details of such databases from each country and also checklists that could be used by systematic reviewers.

Some of the important portals for searching the grey literature are Mednar (https://mednar.com/mednar/desktop/en/search.html), Trove (http://trove.nla.gov.au/; accessed on 15 August 2020) and OAIster (http://oaister.worldcat.org/; accessed on 15 August 2020). The AACODS checklist developed by the researchers from Flinders University is a precious tool while assessing the quality of the identified grey literature (https://dspace.flinders.edu.au/xmlui/bitstream/handle/2328/3326/AACODS_Checklist.pdf;jsessionid=49D97F7AACC861E2BB731E2227025CC5?sequence=4; accessed on 23 September 2020).

Non-english Literature

The Chinese biomedical literature has been rapidly growing over the recent years. China's share in the world's total published scientific papers was less than 1% in 1980 whereas it was about 12% in 2011 and is currently ranking second behind the US. Hence systematic reviewers need to search Chinese literature diligently (Cohen et al. 2015). The important Chinese biomedical databases are Chinese Biomedicine Literature Database (CBM), Chinese Medical Current Content CMCC, China National Knowledge Infrastructure (CNKI, http://www.cnki.net), VIP information and WANFANG (Xia et al. 2008). Unfortunately, it is difficult to gain access to these databases.

LILACS is the Latin American and Caribbean health sciences database. It has interfaces in Portuguese, Spanish and English language and hence relatively easy to navigate. Most of the journals are not indexed in other databases can be accessed at www.bireme.br.free of charge. A free tutorial on how to search LILACS is available on http://bvsalud.org/en/howtosearch/.

It is important to enlist the help of professional language translators to translate Non-English articles to English; however, financial limitations and the limited availability of expert translators are significant barriers.

Google Translate, a free Web-based resource for translation, has the potential to assist systematic reviewers in translating information from non-English literature. A recent study evaluated the utility of Google translate in extracting information from 10 randomised controlled trials in five languages (Chinese, French, German, Japanese, and Spanish) (Balk et al. 2013). The average length of time required to translate articles was 30 min (range 5 min to 1 h). Data extraction from translated articles was less accurate than from English language articles. Extraction was most accurate from translated Spanish articles and least accurate from translated Chinese articles. They concluded that the use of Google Translate has the potential of being an approach to reduce language bias; however, reviewers need to be cautious about using data from such translated articles.

Google Scholar

Recently, Google scholar has gained popularity among clinicians and researchers. While it cannot be considered as a principal search system by the systematic reviewers (Gusenbauer and Haddaway 2020; Bohannon 2014, Younger 1987), it could be an addition to other traditional search methods (Haddaway et al. 2015).

Clinical Trial Registries

Researchers (and journal editors) are generally interested in the publication of trials that show either a large effect of a new treatment (positive trials) or equivalence of two approaches to treatment (non-inferiority trials) (Gusenbauer and Haddaway 2020). They may be less enthusiastic about trials that show that a new treatment is inferior to the conventional treatment (negative trials) and even less interested in trials that are neither clearly positive nor negative, since inconclusive trials will not influence change in clinical practice. Trial results that place financial interests at risk are especially likely to remain unpublished and hidden from public view (De Angelis et al. 2004). Such selective reporting of trials can lead to publication bias and erroneous conclusions of a systematic review (de Vries et al. 2016; Hart et al. 2012; Turner et al. 2008). If all trials are registered in a public repository, all stakeholders can explore the full range of clinical evidence.

In 2004, the International Committee of Medical Journal Editors (ICMJE) proposed that the ICMJE member journals will require, as a condition of consideration for publication, registration in a public trials registry, and that trials must register at or before the onset of patient enrolment (De Angelis et al. 2004). Since then, many countries have established their country-specific trial registries.

ClinicalTrials.gov provides access to summary information on clinical studies on a wide range of diseases and conditions. It was established in 1999 and is maintained by the National Library of Medicine (USA). It consists of a clinical study registry and results' database (Tse et al. 2018).

Researchers are responsible for submitting information about their studies to ClinicalTrials.gov at the start of the study and update it regularly during the project. The summary results are reported on the same website once the study is completed.

As of October 2016, ClinicalTrials.gov contained information on over 227,000 studies (Zarin et al. 2017). By 2018, it had information for nearly 270 000 studies from over 200 countries and summary results of over 30 000 studies (Tse et al. 2018). Currently, it has 348,891 research studies from 216 countries (accessed on 15 August 2020).

The World Health Organization International Clinical Trials Registry Platform (WHO-ICTRP) was established in 2006 (Gülmezoglu et al. 2005), and it continues to coordinate a global network of trial registries. It is important to note that the WHO-ICTRP is not a trial registry, but a platform which accesses other registries from various countries including Australia and New Zealand (ANZCTR), China (ChiCTR), South Korea (CRiS), India (CTRI), European Union (EU-CTR), Iran (ISRCTN), Japan (JPRN), Thailand (TCTR), Africa PACTR and Srilanka (SLCTR) (de Vries et al. 2016). While there are limitations to it, The ICTRP represents a concentrated effort towards data accessibility from various trials. Systematic reviewers should include the trial registries while searching the literature.

In summary, for increasing the reliability of the results, systematic reviewers should search as many databases as possible to ensure all relevant studies are identified. The importance of a pre-planned, explicit, and robust strategy for comprehensive literature search with optimal transparency and reproducibility cannot be overemphasised.

References

Balk EM, Chung M, Chen ML, Trikalinos TA, Kong Win Chang L. Assessing the accuracy of Google translate to allow data extraction from trials published in non-english languages. Rockville (MD): Agency for Healthcare Research and Quality (US); January 2013.

Bohannon J. Scientific publishing: Google Scholar wins raves–but can it be trusted? Science (New York, NY). 2014;343(6166):14.

Canadian Agency for Drugs Technologies in Health. Grey Matters: a practical tool for searching health-related grey literature [internet]. Ottawa (ON): CADTH; 2015 (cited 2016 Oct).

Cohen JF, Korevaar DA, Wang J, Spijker R, Bossuyt PM. Should we search Chinese biomedical databases when performing systematic reviews? Syst Rev. 2015;4:23.

De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. N Engl J Med. 2004;351(12):1250–1.

de Vries YA, Roest AM, Beijers L, Turner EH, de Jonge P. Bias in the reporting of harms in clinical trials of second-generation antidepressants for depression and anxiety: A meta-analysis. Eur Neuropsychopharmacol. 2016;26(11):1752–9.

Fatehi F, Gray LC, Wootton R. How to improve your PubMed/MEDLINE searches: 1 background and basic searching. J Telemed Telecare. 2013;19(8):479–86.

Gülmezoglu AM, Pang T, Horton R, Dickersin K. WHO facilitates international collaboration in setting standards for clinical trial registration. Lancet. 2005;365(9474):1829–31.

Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods. 2020;11(2):181–217.

Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of google scholar in evidence reviews and its applicability to grey literature searching. PLoS ONE. 2015;10(9):e0138237.

Hart B, Lundh A, Bero L. Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses. BMJ (Clinical research ed). 2012;344:d7202.

Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomised trials of health care interventions. Cochrane Database of Syst Rev. 2007(2):Mr000010.

Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA. 1994;271(14):1103–8.

McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet. 2000;356 (9237):1228–31.

Shashikiran ND. MEDLINE, pubmed, and pubmed central (®): Analogous or dissimilar. J Indian Soc Pedodontics Prev Dent. 2016;34(3):197–8.

Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. dissemination and publication of research findings: an updated review of related biases. Health Technol Assess (Winchester, England). 2010;14(8):iii, ix–xi, 1–193.

Tse T, Fain KM, Zarin DA. How to avoid common problems when using ClinicalTrials.gov in research: 10 issues to consider. BMJ (Clinical research ed). 2018;361:k1452.

Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008;358(3):252–60.

Ulincy L. EMCare. J Med Libr Assoc. 2006;94(3):357–60.

Xia J, Wright J, Adams CE. Five large Chinese biomedical bibliographic databases: accessibility and coverage. Health Inform Libr J. 2008;25(1):55–61.

Younger P. Using google scholar to conduct a literature search. Nursing Standard (Royal College of Nursing (Great Britain): 1987). 2010;24(45):40–6; quiz 8.

Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on trial registration 11 years after the ICMJE policy was established. N Engl J Med. 2017;376(4):383–91.

Assessing and Exploring Heterogeneity

Sven Schulzke

Abstract Meta-analysis is a statistical method for combining the results of studies included in the systematic review. It is justified only when the potentially eligible studies are similar enough. However, some differences in their clinical or methodological characteristics are inevitable as no two studies are expected to be identical in the true sense. Clinical heterogeneity is caused by diversity in important characteristics such as participants, interventions, comparators, or outcomes (in extreme cases, 'apples vs. oranges'). Methodological heterogeneity involves differences in the design (e.g., randomised vs. quasi-randomised) and methodological quality of studies (e.g., masked vs. non-masked allocation) included in a systematic review. A fair amount of clinical judgement is thus necessary to decide whether or not studies are similar enough to be combined in a meta-analysis. Statistical heterogeneity in a meta-analysis means that the between-study variation in the effect of intervention varies beyond the extent expected by chance alone. This chapter is focussed on understanding, assessing and handling heterogeneity from various sources in meta-analysis.

Keywords Clinical heterogeneity · Chi-squared test · I-squared statistic · Methodological heterogeneity · Statistical heterogeneity · Sensitivity analysis · Subgroup analysis

Introduction

Meta-analysis involves a statistical method for combining the results of studies that are included in the systematic review. It is only reasonable to conduct meta-analysis when patient populations, interventions, outcomes, and follow-up of the considered studies are similar enough ('only compare apples to apples'). On the other hand,

S. Schulzke (✉)
Neonatologist, Director of Research, University Children's Hospital Basel UKBB,
Spitalstrasse 33, 4056 Basel, Switzerland
e-mail: sven.schulzke@ukbb.ch

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_3

there is always some difference in the clinical or methodological characteristics of studies given that no two studies are entirely identical (‘all apples are different’). Thus, a fair amount of clinical judgement is necessary to decide whether or not studies are similar enough to be combined in a meta-analysis.

Clinical Heterogeneity

Clinical heterogeneity (Table 1) is caused by clinical diversity in important study characteristics such as participants, interventions, or outcomes (in extreme cases, ‘apples vs. oranges’) (Cochrane 2017). However, even comparing apples with oranges may be adequate if we are indeed interested in evaluating the ‘effects of fruits’ in general.

Methodological Heterogeneity

In addition to clinical heterogeneity, study design (e.g., randomised vs. quasi-randomised) and methodological quality of studies (e.g., masking vs. non-masking of allocation) included in a systematic review may vary creating another important level of heterogeneity between studies. Such variability in the design and quality of the study is typically termed *methodological heterogeneity* (Table 2) (Cochrane 2017).

Table 1 Examples of clinical heterogeneity

Clinical characteristics	Examples of diversity
P: Patients	Age, sex, type of disease, the severity of the disease, stage of the disease
I: Intervention	Dose, duration, timing, frequency of treatment; different personnel administering the intervention
C: Control intervention	Placebo, standard care, no control treatment
O: Outcome	Type and definition of the event, duration of follow-up, different instruments to measure the outcome
T: Timing	Study setting, e.g., time of year, geographic setting, local setting (where were data collected)

Table 2 Examples of methodological heterogeneity

Methodological characteristic	Examples of diversity
Design	Randomised/non-randomised, parallel/crossover
Allocation	Concealed/non-concealed
Masking	Intervention masked/non-masked, outcome assessment masked/ non-masked
Analysis and reporting	Intention-to-treat vs. per protocol

Conceptual Heterogeneity

Clinical and methodological heterogeneity is summarised with the term *conceptual heterogeneity* (Fletcher 2007). Some degree of conceptual heterogeneity is inevitable (Higgins 2008), however, it can be reduced pre-emptively by asking a focused question (PICO) and by considering only specific trial/study designs at protocol stage of the systematic review (Gagnier et al. 2012).

Statistical Heterogeneity

'*Statistical heterogeneity'* in a meta-analysis means that the between-study variation in the effect of intervention varies beyond the extent expected by chance alone (Cochrane 2017). Statistical heterogeneity is a result of clinical or methodological heterogeneity or the combination of two heterogeneities among the studies (Fig. 1). Methods of exploring statistical heterogeneity are outlined below.

Exploring the presence and sources of statistical heterogeneity are important to understand differences in treatment effects between studies (Fletcher 2007). Moreover, creatively exploring the sources of statistical heterogeneity may enable us to identify subgroups of patients who benefit most from interventions or those who are particularly vulnerable to adverse events. Additionally, recognising heterogeneity is important when applying evidence from a systematic review to individual patients. For example, if a meta-analysis demonstrates the beneficial effects of an intervention across several studies in the presence of heterogeneity, our confidence in the applicability of the study results to a given patient may be increased. In other words, if a treatment has beneficial effects across different patient populations, at different dosages, and different treatment durations, the observed treatment effect is likely to occur in a given patient under the circumstances similar but not identical to those considered in the studies. In this case, the presence of heterogeneity may be advantageous when applying evidence to individual patients. Finally, assessing inconsistency of effect estimates is important when grading the quality of evidence and the strength of recommendations. Depending on the severity of inconsistencies, our confidence in the effect estimates

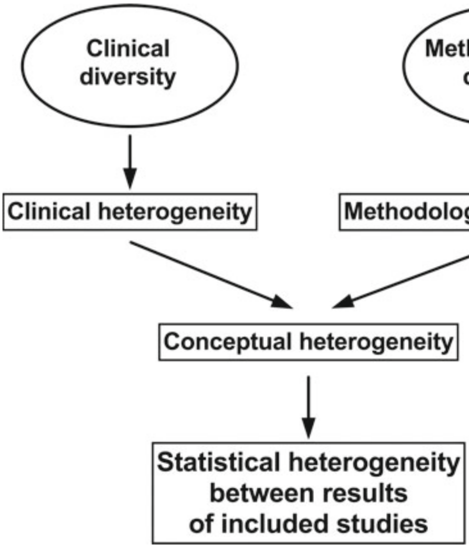

Clinical diversity
Methodological diversity
Clinical heterogeneity
Methodological heterogeneity
Conceptual heterogeneity
Statistical heterogeneity between results of included studies

Fig. 1 Associations between clinical, methodological, conceptual and statistical heterogeneity

of a meta-analysis might decrease considerably. In severe cases, this leads to the downgrading of the quality of evidence and strength of recommendation (Andrews et al. 2013). The approach to the assessment of the quality of evidence and strength of recommendations has been covered elsewhere in this book.

Measuring Statistical Heterogeneity

Investigating the sources of heterogeneity in meta-analysis is by nature exploratory, and therefore should always be interpreted with caution. However, careful assessment of heterogeneity may provide critical insight into the results and interpretation of a meta-analysis and lead to evidence that can be useful in suggesting the direction of future research (Song et al. 2001). Heterogeneity can be assessed using the ‘eyeball’ test or more formally with statistical tests such as the Cochran chi-squared (Cochran Q) test or the I-squared statistic (Centre for Evidence-Based Medicine 2014; Higgins and Thompson 2002; Patsopoulos et al. 2008a; b).

The ‘eyeball’ test involves visually assessing the confidence intervals of studies in a forest plot (Fig. 2). Heterogeneity is likely when confidence intervals of included studies in a forest plot do not overlap with the confidence interval of the summary effect estimate (Fig. 2a). Exploring the reasons for non-overlapping confidence intervals in a forest plot might reveal important differences in clinical or methodological characteristics of included trials. For example, the three studies favouring treatment over control in Fig. 2a might assess patients who are

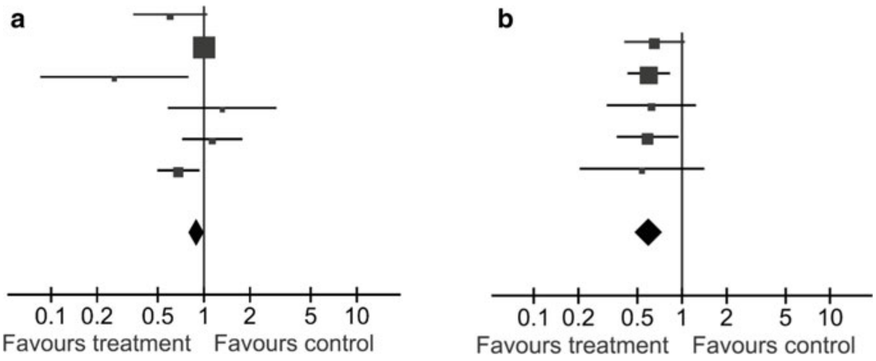


Fig. 2 The forest plots display the relative risk of an outcome in treatment vs. control group among studies included in the meta-analysis. The squares represent the point estimate of the treatment effect of each study with a horizontal line extending on either side of the square representing the 95% confidence interval. The size of the squares reflects the sample size and weight of each study in the meta-analysis, given that larger studies result in more precise point estimates and narrower confidence intervals. The diamonds represent the overall relative risk estimate of the studies presented in the meta-analysis. The widths of the diamonds represent the 95% confidence interval of the relative risk. The vertical midline of the forest plot corresponding to a relative risk of 1 represents a 'no effect' line

considerably younger (or less diseased) than those from the other studies, resulting in more beneficial effect estimates in the former and heterogeneity among all studies. The 'eyeball' test indicates the presence of homogeneity, i.e., no heterogeneity when confidence intervals of all included trials in a forest plot overlap with the summary effect estimate (Fig. 2b).

Panel a: Visual inspection ('eyeball' test) indicates that there is heterogeneity among studies as the confidence intervals of some of the six included trials are not overlapping with the confidence interval of the summary relative risk estimate. Further, there are three studies favouring treatment, one very large study (as indicated by the large square) at the 'no effect' line, and two other studies favouring the control intervention. In this example, heterogeneity does influence the summary relative risk estimate; the latter is tending towards a benefit of the intervention, although the studies are unlikely to assess the same true effect.

Panel b: Confidence intervals of all trials overlap indicating that there is no heterogeneity between the five included studies. In other words, study results are homogeneous.

The Cochran chi-squared (Cochran Q) test: This is a non-parametric statistical test assuming the null hypothesis of homogeneity among all studies in a meta-analysis. The test considers the differences between observed effects in the individual studies and the pooled effect estimate. It squares those differences, then divides by variance, and sums up. This gives the chi-squared test statistic T with the degrees of freedom (df) equal to the number of studies -1. The expected chi-squared test statistic T if the null hypothesis is true equals the degrees of freedom. A very low p-value in the Cochran chi-squared test indicates heterogeneity among studies. Unfortunately, the reliability of the Cochran chi-squared test is poor. The sensitivity of the Cochran chi-squared test to detect heterogeneity is low if there are few studies included in the meta-analysis (which often is the case), i.e., the p-value of the Cochran chi-squared test may not be very low although there is considerable heterogeneity among trials. In order to alleviate this issue, a p-value of 0.1 (rather than 0.05) is typically used as a cut-off to indicate statistically significant heterogeneity in the Cochran chi-squared test. For example, the Cochran chi-squared test of the meta-analysis in Fig. 2a results in a p-value of 0.02, indicating significant heterogeneity. If Cochran chi-squared is not statistically significant, but the ratio of T and the degrees of freedom (T/df) is > 1, there is potential heterogeneity. If the test is not statistically significant, but T/df is < 1 then heterogeneity is unlikely (Centre for Evidence-Based Medicine 2014). For example, the p-value of the Cochrane chi-squared test of the meta-analysis in Fig. 2b is 1.00. With df = 4, T/df is 0.25; thus, heterogeneity is unlikely based on the Cochran chi-squared test. However, It is important to know that the absence of heterogeneity does not equate with evidence of homogeneity given the low sensitivity of the Cochran chi-squared test (see above). On the other hand, a single outlying study with confidence interval not overlapping with those from all other studies in a meta-analysis may cause a significant test result in the Cochran chi-squared test but may not be important in the overall interpretation of the results. Therefore, even a positive test result may not be beneficial. Lastly, it is probably too simplistic to answer the question about the presence of heterogeneity with a simple yes/no answer based on a single statistical test.

The I-squared statistic: This statistic is an index of the degree of heterogeneity among the included studies. Rather than providing a simple yes/no answer as in the Cochrane chi-squared test, the I-squared statistic expresses the degree of inconsistency between study results as a percentage. I-squared can be anywhere between 0 and 99% with low values indicating no or little heterogeneity and high values indicating a high probability of heterogeneity (Patsopoulos et al. 2008a). There is no definitive cut-off to prove heterogeneity, and like any test statistic, I-squared has a level of uncertainty. I-squared thresholds can be misleading as the importance of inconsistency among trials depends on several factors such as the magnitude of the effect, the direction of effect, and strength of evidence for heterogeneity (e.g., a p-value of chi-squared test or confidence interval of I-squared statistic) (Higgins and Green 2011). In order to address this, overlapping I-squared thresholds for interpretation of the importance of heterogeneity may be given. As a rough guide, I-squared values below 30–40% may represent low heterogeneity, 30–60% might reflect moderate heterogeneity, 50–90% might represent considerable heterogeneity, and 75–100% might represent high heterogeneity (Higgins and Green 2011). For example, the I-squared statistic of the meta-analysis in Fig. 2A is 63%, indicating moderate to considerable heterogeneity. The I-squared statistic of the meta-analysis shown in Fig. 2b is 0%, indicating no statistical heterogeneity.

Dealing with Statistical Heterogeneity

Several options and strategies for addressing statistical heterogeneity are available (Higgins and Green 2011).

(1) *Check accuracy of data*

As a first step, one should always re-check the accuracy of the data entered into the software used for generating the forest plots in order to rule out simple coding errors causing invalid effect estimates.

(2) *Do not combine studies in a meta-analysis*

In the presence of substantial heterogeneity, especially if the direction of effect varies between studies, it may be misleading to conduct a meta-analysis and to provide an average effect estimate. In these cases, a narrative description of the results of the included studies might be preferable. Reviewers repeatedly face the question of whether to describe study results only narratively or to combine them mathematically in a meta-analysis. *This question needs to be answered for each comparison and each outcome of each study in the systematic review.* E.g., some outcomes of some studies assessing a specific comparison may be judged to be amenable to meta-analysis while others maybe not. In other cases, all outcomes of a comparison of intervention vs. control may be reported narratively because conceptual heterogeneity between studies may be deemed too severe to combine results in a meta-analysis. It is often possible to answer those questions by using clinical judgement and assessing the degree of conceptual heterogeneity (comparability of patients, interventions, control interventions, outcomes and follow-up, and study settings/methods). This solution, however, implies that the reviewers have the necessary clinical and methodological knowledge to provide a sensible judgement. Therefore it is recommended to assemble reviewer teams with members who complement each other in these qualifications.

(3) *Explore the reasons for heterogeneity*

Exploring heterogeneity can be formally accomplished by subgroup analysis, sensitivity analysis, and meta-regression.

Subgroup analysis: Heterogeneity frequently can be anticipated at protocol stage of a systematic review, allowing for pre-emptive declaration of subgroups in order to assess the influence of heterogeneity once the studies have been selected for inclusion in the review (Higgins and Green 2011). For example, cooling of asphyxiated neonates with hypoxic-ischaemic encephalopathy can be carried out using selective head-cooling or whole-body cooling devices. Studies considered for a meta-analysis of cooling for neonates with this condition could thus be analysed both in prestated subgroups based on the type of cooling device and by pooling of results from all studies.

Sensitivity analysis: Given that the methodology of included studies might influence the results of a systematic review, sensitivity analysis may be useful

(Higgins and Green 2011; Thabane et al. 2013). Sensitivity analysis assesses the robustness of effects of interventions towards changes in methodology, models, or assumptions of included trials. For example, one might carry out a sensitivity analysis by comparing the effect of interventions in truly randomised vs. quasi-randomised trials included in a systematic review. Ideally, sensitivity analysis should be declared at the protocol stage.

Meta-regression: Statistical heterogeneity can further be explored by quantitative techniques such as meta-regression (Baker et al. 2009; Thompson and Higgins 2002).This involves mathematically assessing the impact of interactions between covariates and treatment effects using regression techniques. Meta-regression typically has low statistical power and requires at least 5–10 studies in a meta-analysis to detect relationships between covariates and treatment effects reliably. Caution is warranted in conduction and interpretation of results, as relationships are purely observational, potentially affected by bias originating from the aggregation of data, and can be confounded by other variables not considered in the analysis (Baker e al. 2009; Thompson and Higgins 2002). Preferably, a limited set of scientifically well-founded covariates is chosen at the protocol stage, and a permutation test be carried out in order to reduce the risk of false-positive findings of a meta-regression (Higgins and Thompson 2004).

(4) *Mathematically allow for heterogeneity*

Effects of studies can be mathematically combined in a meta-analysis using either *fixed-effects or random-effects models* (Higgins and Green 2011). A fixed-effects model ignores heterogeneity by assuming that all included studies measure the same effect. This assumption implies that the observed differences among study results are entirely due to chance, i.e. that there is no statistical heterogeneity. If there is no heterogeneity, the summary effect estimate is interpreted as being the best estimation of the treatment effect. Fixed-effects models are powerful and provide narrow confidence intervals. In the presence of statistical heterogeneity that cannot be readily explained or explored, using a fixed-effects model might be inadequate because studies might assess different effects (e.g., differences in the population). In such circumstances, using a random-effects model allows for heterogeneity among results of included studies. A random-effects model assumes that the effects in the different studies are not identical but follow a distribution (Higgins and Green 2011). The typical choice of distribution is a normal distribution; however, it is difficult to show that this choice is adequate (which is a common criticism of random-effects meta-analyses). A random-effects model does not adjust for statistical heterogeneity in the sense that it is no longer an issue. It instead incorporates heterogeneity between studies as the confidence interval. A random-effects meta-analysis reflects both between and within study variations of effect estimates. Consequently, random-effects meta-analysis has less statistical power and wider confidence intervals compared to fixed-effects meta-analysis. It is essential to realise that in the presence of heterogeneity, a random-effects meta-analysis weights the studies more equally than a fixed-effect analysis, i.e., small studies receive a higher weight in random-effects meta-analysis. Therefore, if an intervention is more beneficial in the smaller studies of a meta-analysis, the overall random-effects estimate of the intervention effect will be more beneficial than the fixed-effect estimate (Higgins and Green 2011).

References

Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, Rind D, Montori VM, Brito JP, Norris S, Elbarbary M, Post P, Nasser M, Shukla V, Jaeschke R, Brozek J, Djulbegovic B, Guyatt G. GRADE guidelines: Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013;66(7):726–35.

Baker WL, White CM, Cappelleri JC, Kluger J, Coleman CI; Health Outcomes, Policy, and Economics (HOPE) Collaborative Group. Understanding heterogeneity in meta-analysis: the role of meta-regression. Int J Clin Pract. 2009;63(10):1426–34.

Centre for evidence-based medicine. Exploring heterogeneity. http://www.cebm.net/wp-content/uploads/2014/06/SYSTEMATIC-REVIEW.docx. Accessed 28 June 2017.

Cochrane. Exploring heterogeneity—slidecast. http://training.cochrane.org/resource/exploring-heterogeneity. Accessed 27 June 2017.

Fletcher J. What is heterogeneity and is it important? BMJ. 2007;334(7584):94–6.

Gagnier JJ, Moher D, Boon H, Beyene J, Bombardier C. Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. BMC Med Res Methodol. 2012;12(111):2.

Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. Int J Epidemiol. 2008;37(5):1158–60.

Higgins JPT, Green S (eds) Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011. www.handbook.cochrane.org.

Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58.

Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med. 2004;23(11):1663–82.

Patsopoulos NA, Evangelou E, Ioannidis JP. Heterogeneous views on heterogeneity. Int J Epidemiol. 2008a;38(6):1740–2.

Patsopoulos NA, Evangelou E, Ioannidis JP. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. Int J Epidemiol. 2008b;37(5):1148–57.

Song F, Sheldon TA, Sutton AJ, Abrams KR, Jones DR. Methods for exploring heterogeneity in meta-analysis. Eval Health Prof. 2001;24(2):126–51.

Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, Thabane M, Giangregorio L, Dennis B, Kosa D, Borg Debono V, Dillenburg R, Fruci V, Bawor V, Lee J, Wells G, Goldsmith CH. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol. 2013;13:92.

Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21(11):1559–73.

Assessment of the Risk of Bias

Kwi Moon and Shripada Rao

Abstract Biases (systematic errors or deviations from the truth) can result in under-estimation or over-estimation of results. If the studies included in the systematic review have a low risk of bias (ROB), the evidence is more likely to be reliable, and vice versa. Hence, systematic reviewers should carefully assess and report the ROB in the included studies and explore its impact on meta-analyses. Biases in randomised controlled trials (RCTs) include inadequate randomisation process, deviations from intended interventions, missing outcome data, the bias in measuring outcomes, and selective reporting of results. ROB-2 is the currently recommended tool for assessing ROB in RCTs. At least two reviewers should assess the ROB independently and resolve differences of opinion by discussion or by consulting a third reviewer. ROB assessment should be performed separately for each outcome. Robust processes for generating random sequence numbers, adequate allocation concealment, blinding of participants, clinicians, researchers and outcome assessors, and maximising follow up rates will minimise the ROB in RCTs.

Keywords Risk · Bias · Assessment · Systematic review · Randomised controlled trials · ROB-2

K. Moon (✉)
Department of Pharmacy, Perth Children's Hospital, 15 Hospital Ave, Nedlands, Perth,
WA 6009, Australia
e-mail: kwi.moon@health.wa.gov.au

S. Rao
Neonatal Directorate, Perth Children's Hospital, 15 Hospital Ave, Nedlands, Perth,
WA 6009, Australia
e-mail: shripada.rao@health.wa.gov.au

K. Moon · S. Rao
School of Medicine, University of Western Australia, Perth, WA 6009, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_4

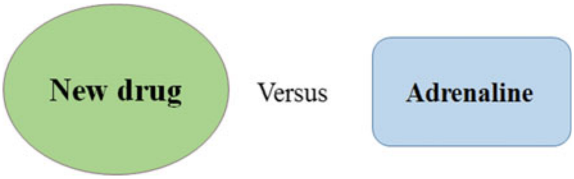
Introduction

The Cochrane Handbook defines bias as a systematic error, or deviation from the truth, in results (Higgins et al. 2019). Biases can lead to under-estimation or over-estimation of the true intervention effect and can vary in magnitude (Higgins et al. 2019). Risk of bias (ROB) assessment is an essential step while conducting systematic reviews. If the randomised controlled trials (RCTs) included in the systematic review have low ROB, the evidence is more likely to be reliable. Conversely, if the majority of the included RCTs have a high ROB, the evidence will become less reliable (da Costa et al. 2017). The major types of biases in RCTs that need to be assessed by systematic reviewers include the following: (1) Bias arising from the randomisation 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 due to selective reporting of results (Higgins et al. 2019; Sterne et al. 2019).

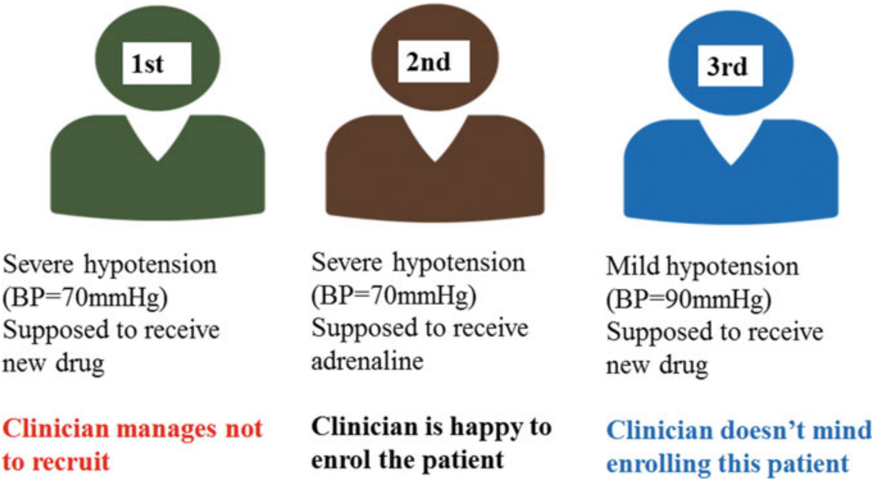
1. **Bias arising due to inadequacy in the randomisation process (Previously known as "Selection bias")**

This type of bias occurs when recruiters selectively enrol participants into the trial based on the knowledge about what the next treatment allocation is likely to be (Kahan et al. 2015). It can result in systematic differences in baseline characteristics of the groups that are compared, which then impact on the results of the study.

a. **Bias may occur if appropriate methods are not used for random sequence generation**. Take the example of a hypothetical RCT in which intensive care unit (ICU) patients with low systolic blood pressure (<95mmHg) are randomised to receive adrenaline infusion or the new drug.



The hypothesis is that the new drug can reduce the duration of ICU stay when compared to adrenaline, and mortality will be similar. As per the study protocol, allocation of the study drug is to be done alternatively, i.e. the first patient receives adrenaline, and the next patient receives the new drug and so on. If the treating clinician doesn't believe in the new drug, he/she might be reluctant to recruit a patient with severe hypotension (e.g. systolic BP=70 mmHg), being aware that the patient will be allocated to the new drug. On the other hand, the clinician may be happy to recruit the next patient who has mild hypotension (e.g. systolic BP= 90 mmHg), to receive the new drug knowing that the patient is not too ill. If the next patient has severe hypotension (e.g. systolic BP=70 mmHg), the clinician will be happy to recruit knowing that the patient will be allocated to receive adrenaline infusion.


1st
2nd
3rd
Severe hypotension (BP=70mmHg) Supposed to receive new drug
Severe hypotension (BP=70mmHg) Supposed to receive adrenaline
Mild hypotension (BP=90mmHg) Supposed to receive new drug
Clinician manages not to recruit
Clinician is happy to enrol the patient
Clinician doesn't mind enrolling this patient

If this process of selection is repeated throughout the trial, the majority of patients with milder hypotension will end up receiving the new drug, and those with severe hypotension will receive adrenaline. When data is analysed after full recruitment of 100 patients, the results might favour the new drug as shown below.

	New drug	Adrenaline	P-value
Duration of ICU stay	3 days (SD 1.2)	6 days (SD 1.9)	0.003

However, upon closer inspection of the baseline characters, it is evident that the real reason for the difference in results is that the patients in the new drug group had milder hypotension compared to those who received adrenaline.

	New drug	Adrenaline	P-value
Mean BP on entry into the trial	93 mm Hg (SD 9)	75 mm Hg (SD 8)	0.001

This is an example of selection bias, which occurred because the sequence of allocation was alternate and hence predictable. Therefore, allocations based on alternate patients, or other similar methods such as odd/ even day of admission, the month of the year, and date of birth are not ideal. This type of bias could have been minimised if the *sequence generation* was random. The robust methods include the use of computer-generated random sequence numbers or random number tables where the generated sequences are unpredictable. Manual methods of achieving random allocation such as coin tossing or throwing dice may become non-random, are challenging to implement and do not leave an audit trail and hence not recommended (Dettori 2010).

Simple randomisation is the simplest and most effective method to prevent selection bias (Kahan et al. 2015). It works by assigning each patient to one of the treatment groups with a certain probability (usually 50%). This probability is the same for every patient, regardless of previous allocations. For example, consider a trial where 35 of the first 50 patients are assigned to the intervention and only 15 to the control. When the 36th patient presents for randomisation, he/she would still have an equal chance of being assigned to either treatment group, regardless of the imbalance in numbers. Because the probability is always the same, recruiters will not be able to guess with any accuracy which treatment the patient will be assigned to. Thus selection bias cannot occur in such a scenario. The main disadvantage of simple randomisation is that it could lead to an unequal number of participants in the two groups. However, if the overall sample size is large, this imbalance has only a small impact on power and should not be used as a reason to avoid simple randomisation. Experts recommend that simple randomisation should be used more frequently in practice (Kahan et al. 2015). Methods such as block randomisation and stratified randomisation may be used to ensure balance in size and important prognostic factors (Setia 2016; Randomization 2011), but have the potential to increase the ability of the recruiter to guess the sequence.

b. **The second type of bias at the stage of the randomisation process occurs if *allocation concealment* is not adequate**.

In an RCT, even though robust methods have been used to generate random numbers, selection bias would still occur if the recruiter has access to the randomisation list. That means even before approaching for consent; the recruiter will be aware of the group, the participant will be randomised to if consent is obtained. It can introduce selection bias similar to the previous example.

Allocation concealment is a technique used to prevent selection bias by concealing the allocation sequence from those assigning participants to intervention groups, until the moment of assignment. It prevents researchers from influencing (unconsciously or otherwise) which participants are assigned to a given intervention group.

Inadequate allocation concealment may produce an exaggerated estimate of treatment effects (Schulz and Grimes 2002; Kjaergard et al. 2001; Odgaard-Jensen et al. 2011) by up to 30–40% (Moher et al. 1998; Nunan et al. 2018). Pindal et al. examined the impact of allocation concealment in RCTs on the conclusion of meta-analyses by reviewing 38 Cochrane reviews. They found that 2/3rd of conclusions that favoured interventions no longer did so when studies with inadequate or unclear allocation concealment were excluded (Pildal et al. 2007). For this reason, allocation concealment has been recommended as an essential tool (Schulz and Grimes 2002; Nuesch et al. 2009).

The following methods of concealing the allocation sequence are considered appropriate:

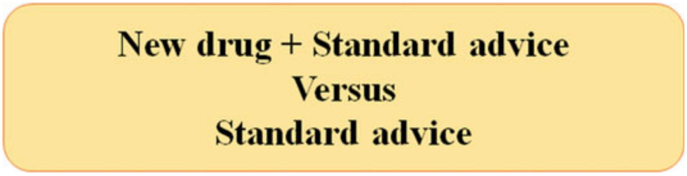
a. **Centralised telephone system administered by the trial co-ordination centre**: In this method, the recruiter telephones the trial administration centre after obtaining the participant's consent. As per the protocol, basic clinical details of the participant are then entered by the receiver into a customised database to generate the allocation, which is then given to the recruiter (Kennedy et al. 2017).
b. **Web-based allocation**: In this method, all random sequence numbers that have been generated are stored on a secure trial-specific website. Once the trial participant gives consent, the investigator logs into that website and enters participant details. Inbuilt algorithms in the software take into consideration the baseline characters of the trial participant and select the intervention to which the participant/patient should be allocated. Various online tools are available for this purpose; with many having an open access (Morice 2012; Cai et al. 2010).
c. **Sequentially numbered, opaque, sealed envelopes (SNOSE)**: The generated random sequence numbers stored in SNOSE are usually considered adequate. However, they may not be completely free of bias (Kennedy et al. 2017). Unsealed or transparent envelops are not ideal for concealing the allocation sequence.
d. **Pharmacy controlled**: Another frequently used method of allocation concealment, common in drug trials, is to get the allocation done by a pharmacy (Altman and Schulz 2001). The trial drugs are provided in containers of identical appearance and weight according to the allocation sequence.

In summary, randomising trial participants into a treatment group is a two-step process. The first step is to generate an **unpredictable** randomisation sequence, and the next step is to **conceal** this sequence from everyone involved in the recruitment process to prevent selection bias (Clark et al. 2013). It is important to note that allocation concealment refers to the technique used to implement the sequence, not to generate it (Schulz and Grimes 2002).

2. **Bias due to deviations from intended interventions (Previously known as "Performance bias")**

This type of bias occurs in an RCT when there are systematic differences in the care provided to the participants between the groups other than the intervention(s) being evaluated (McCambridge et al. 2014). It occurs especially if investigators and participants are not blinded to the interventions.

Take the example of a hypothetical RCT in which a new drug is being evaluated to treat obesity.



New drug + Standard advice
Versus
Standard advice

As per the trial design, the new drug is being compared with the standard care (i.e. advice on diet and exercise). The outcome of interest is weight loss at the end of 6 months. Patients randomised to the new drug might become complacent and stop regular exercises and good dietary habits, comfortable in the knowledge that the new drug would help them lose weight. On the other hand, patients randomised to standard care might spur into action and take better control of their weight issues. The clinician might also inadvertently provide additional support to those not receiving the new drug (e.g. referral to a dietician and an exercise therapist). In effect, systematic differences between the two groups have been introduced, other than the new drug being tested. The final results of the study may look as below:

Weight loss: New drug group	Weight loss: Standard group	P-value
5%	10%	0.03

Hence one could incorrectly conclude that the new drug is worse than standard therapy in obesity. The erroneous conclusion was reached because bias was introduced that affected the **performance** of the new drug because the recruiter and the participant were **not blinded** to the intervention. This bias could have been avoided by giving placebo tablets to the control group. The placebo tablet should be identical (or very similar) in colour, taste and texture to the new drug. This would have ensured that patients in both groups did not realise what they were receiving, and even the investigators were unaware of what group any patient was in. In controlled trials the term blinding, and in particular "double-blind," usually refers to keeping the study participants, their caretakers, and those collecting and analysing data unaware of the assigned treatment, to avoid getting influenced/biased by that knowledge (Day and Altman 2000). Overall, randomised trials that do not use the appropriate level of blinding tend to show larger treatment effects than blinded studies (Hrobjartsson et al. 2014, 2012; Savovic et al. 2012). For subjective outcomes, ROB with non-blinding is even greater such that intervention effects may be exaggerated by as much as 36% (Hrobjartsson et al. 2012).

Blinded participants are less likely to have biased responses to intervention, more likely to comply with trial regimens, less likely to seek additional adjunct interventions and less likely to leave trial (Schulz and Grimes 2002). Blinded investigators are less likely to transfer their inclinations to participants, and less likely to differentially adjust the dose, withdraw participants and encourage or discourage participants from continuing in the trial (Schulz and Grimes 2002).

Overall, blinding of the caregivers and patients/participants makes it difficult to bias results intentionally or unintentionally and increases the internal validity of the trial and credibility of conclusions (Day and Altman 2000). Hence every effort should be made to achieve adequate blinding in RCTs. However, in many RCTs, it is not possible to blind the clinicians and the participants. For example, in a trial comparing conventional ventilation versus high-frequency oscillator, it is impossible to achieve blinding. Similarly, it is difficult to achieve blinding in surgical RCTs (McCulloch et al. 2002). In the past, such open trials were always considered as carrying a high ROB. However, open trials can be at low ROB if there are no deviations from the intended intervention that arose because of the trial context. When patients/participants or clinicians cannot be blinded, trialists should ensure that, apart from the intervention, the allocation groups are treated as equally as possible. This may involve standardising the care of participants, including co-interventions, frequency of follow-up and management of complications (Karanicolas et al. 2010). Blinding of other team members and outcome assessors helps in minimising bias in such situations (McCulloch et al. 2002; Karanicolas et al. 2010).

3. **Bias due to missing outcome data**

This type of bias occurs due to systematic differences between the study groups in the number of participants lost from a study and the reasons (Nunan et al. 2018). In many trials, participant data are missing because of loss to follow-up or incomplete data collection (Hewitt et al. 2010). Data could be missing because the subjects were unable to provide it (e.g., dead, severely impaired), withdrew from the study voluntarily or because of an adverse event, or were lost to follow-up (e.g., moved away from the study area). The situation most likely to lead to bias is when reasons for the missing outcome data differ between the two groups. For example, if the study subjects who became seriously unwell withdrew from the comparator group while those who recovered withdrew from the experimental intervention group (https://training.cochrane.org/handbook/current/chapter-08#section-8-5-2*)*. Results of a trial become less reliable, especially if outcome data is missing from greater than 20% of participants (Schulz and Grimes 2002). Unfortunately, there is no threshold for defining *'small enough'* in relation to the proportion of missing outcome data.

In addition to 'complete-case analyses' wherein the statistical analysis is done for all patients where full information is available, trial authors may need to conduct analyses to address bias caused by missing outcome data. Approaches include single imputation (e.g. assuming the participant had no event; last observation carried forward), multiple imputation (Sullivan et al. 2016) and likelihood-based methods. However, such an approach may lead to spurious conclusions. Trialists should hence consider other strategies to minimise attrition bias (e.g. attempt to follow up all participants and sensitivity analyses) (White et al. 2011).

4. **Bias in measurement of outcomes**

Some circumstances when this type of bias occurs include the following:

a. **When an inappropriate method is used for measuring the outcome**. For example, in a randomised trial comparing conventional antihypertensive medication versus new medication, low-quality home blood pressure monitors may not correctly measure blood pressure levels above 200 mm Hg. The true incidence of severe hypertension will hence be unreliable.
b. **When ascertainment of the outcome differs between intervention groups**. For example, in a randomised trial comparing aspirin versus new drug for migraine, patients in the new drug group may have more headaches and undergo

more MRI scans leading to the detection of more number of benign asymptomatic brain tumours. The conclusion that patients in the new drug have a higher incidence of benign brain tumours is biased and hence incorrect.

c. **If the blinding of outcome assessors is inadequate**. Take the example of a hypothetical open-label trial comparing the effect of a new drug versus the conventional drug morphine for postoperative pain relief in newborn infants. The outcome of interest is pain-score as assessed by the nurses caring for the infants. Higher scores are considered to indicate severe pain. These scores are subjective and hence dependent on the opinion of the nurse looking after the infant. Since it is an open-label trial, the nurse will be aware of whether the infant is receiving morphine or the new drug. If the infant was randomised to the new drug, she might consciously or subconsciously give higher pain scores because of personal belief that the new drug is not good. On the other hand, if she is caring for an infant who has been randomised to morphine, she may allocate lower pain scores because she believes that morphine is a better analgesic. Hence a bias has been introduced due to the personal opinion of the nurse. Blinding of outcome assessors (nurses, in this case) could have prevented this type of bias.

In summary, blinded assessors are less likely to have a bias that can affect their assessments of the outcomes (Schulz and Grimes 2002). Hence every effort should be made to blind the outcome assessors in an RCT.

Trials Where Blinding Is Difficult

In some trials, it may be difficult to blind the participants and investigators (e.g. trials involving surgical procedure). An extra effort should be placed on blinding the outcome assessors in such studies. Additionally, novel strategies to maintain blinding for participants and investigators should be considered. (e.g. simulation of interventional procedures, imitation of incision/surgical access point, standardisation of interventions and care) (Wartolowska et al. 2017).

Difference Between Blinding and Allocation Concealment

One should not confuse “blinding” with “allocation concealment”. Allocation concealment is aimed to protect allocation sequence before and until assignment whereas blinding is a process undertaken to conceal group identity after assignment (Schulz et al. 2002). Blinding seeks to prevent performance and detection bias.

5. **Bias in selecting the results for reporting**

Reporting bias occurs if trial investigators selectively report the results of their trial (Kirkham et al. 2010, 2018).

There are various types of selective reporting in RCTs (Higgins et al. 2019):

a. **Not analysing as per the pre-specified plan**.

Systematic reviewers should attempt to retrieve the pre-specified analysis plans for each trial. It allows identification of outcome measures or analyses that have been omitted from or added to the reported results, post hoc.

Take the example of a hypothetical RCT that tested a new drug for the treatment of influenza in a community setting. The intervention group is to receive the new drug and control group is to receive the current standard (symptomatic therapy). Pre-specified outcomes of interest are "duration of illness" and "side effects of the drug". The trial finds that the new drug decreases the duration of illness by two days compared to symptomatic therapy, and patients receiving the new drug develop transient hepatic dysfunction. The trial investigators publish information on the duration of illness but withhold the information regarding the abnormal liver functions.

b. **Selective reporting of particular outcome measurement**.

For example, in a trial comparing new drug versus morphine for postoperative analgesia in neonates, the pain scores are measured using two scales: Scale A and Scale B. The results show a new drug is superior to morphine on Scale A but no different on Scale B. The results for scale A scale are reported, whereas Scale B results are not.

c. **Selective reporting of particular analysis (based on results) from multiple analyses**.

For example, in the above scenario, the investigators analysed the difference in pain scores between the new drug and morphine at six hours after commencing the medication but found no difference between the groups. They also compared the change in pain scores from baseline and found that infants in the new drug group had a significant drop in pain scores from baseline. Hence, they decided to report only the latter result.

In summary, reporting bias is a common phenomenon (Page et al. 2014) of concern (Ioannidis et al. 2017; Jones et al. 2015). To avoid the potentially harmful effects due to reporting bias, all trials should be registered, and provide the explicit list of pre-specified outcomes before starting recruitment. The investigators should provide a transparent description of all changes that occur afterwards from during the trial until publication (Ioannidis et al. 2017). In addition, editors should implement better quality control measures to prevent selective reporting of outcome (Ioannidis et al. 2017).

Areas of Debate

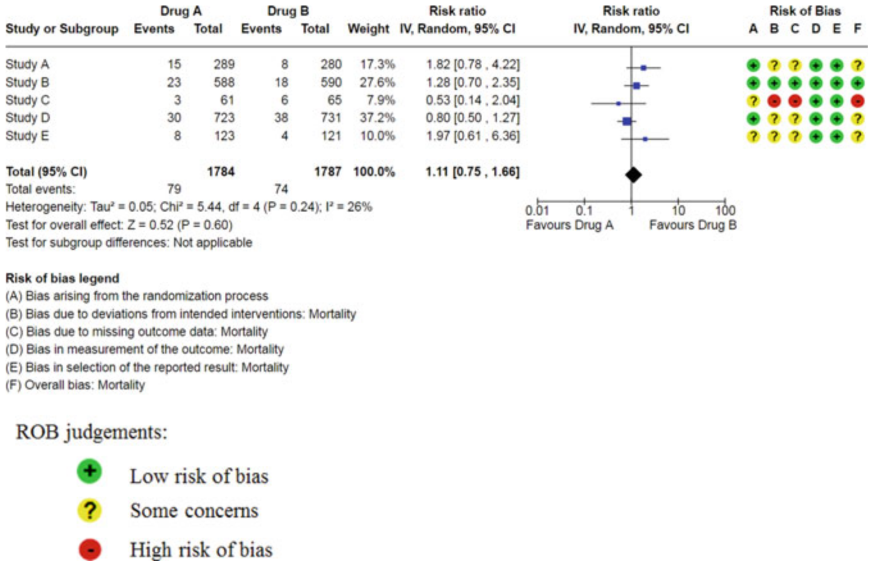
Industry sponsorship and conflict of interest: A recent Cochrane review and other systematic reviews have found that sponsorship of drug and device studies by the manufacturing company leads to more favourable results and conclusions than sponsorship by other sources (Lundh et al. 2012, 2017, 2018). Hence they proposed that the Cochrane ROB tool should include funding source as a standard item (Bero 2013). However, other experts argue that there is little evidence that trial methods are more likely to be flawed in industry-funded trials (Sterne 2013). They also raise the concern that adding a source of funding as a bias domain in the ROB tool would send an extremely negative message to the pharmaceutical industry, and it might have the unintended consequence of labelling high-quality trials as biased while diverting attention from appropriate solutions to problems associated with pharmaceutical industry-sponsored trials (Sterne 2013). Currently, the Cochrane handbook recommends this information to be included in the "Characteristics of included studies" table, rather than the ROB table (Higgins et al. 2019).

Early stopping of trials due to benefit: Another area of debate is whether to include information regarding "early stopping of trials due to benefit" as a ROB domain. It is well known that RCTs ceased early for the benefit is associated with greater effect size than RCTs not stopped early (Bassler et al. 2010; Montori et al. 2005). Researchers have shown that about half of the trials stopped early for benefit were followed by subsequent trials addressing a similar question, suggesting that future trialists may have been sceptical about the premature termination of the prior trials (Murad et al. 2017). Other experts argue that early termination of clinical trials, for either apparent efficacy or harm, is a cornerstone of efficient and ethical trial design, and it does not lead to substantive bias in the estimation of treatment effects (Berry et al. 2010). The current Cochrane handbook does not comment on this issue (Higgins et al. 2019).

Guidelines for Systematic Reviewers to Assess ROB in the Included Studies

The Cochrane handbook's guidelines are a useful tool for systematic reviewers. The ROB is classified as low risk, some concerns, or high risk, based on the information available from the RCTs. The recently updated RoB-2 tool includes algorithms that map responses to signalling questions to a proposed ROB judgement for each domain (Sterne et al. 2019). Systematic reviewers should make an effort to contact the authors of RCTs if relevant information is not available or is unclear in the published articles.

The ROB assessment should include a minimum of two independent reviewers with an unbiased reconciliation method such as a third-person serving as arbitrator (Research 2011). Systematic reviewers should present the "ROB graph" adjacent to



	Drug A		Drug B			Risk ratio	Risk of Bias					
Study or Subgroup	Events	Total	Events	Total	Weight	IV, Random, 95% CI	A	B	C	D	E	F
Study A	15	289	8	280	17.3%	1.82 [0.78 , 4.22]	+	?	?	+	+	?
Study B	23	588	18	590	27.6%	1.28 [0.70 , 2.35]	+	+	+	+	+	+
Study C	3	61	6	65	7.9%	0.53 [0.14 , 2.04]	?	-	-	+	+	-
Study D	30	723	38	731	37.2%	0.80 [0.50 , 1.27]	+	?	?	+	+	?
Study E	8	123	4	121	10.0%	1.97 [0.61 , 6.36]	?	?	?	+	+	?
Total (95% CI)		**1784**		**1787**	**100.0%**	**1.11 [0.75 , 1.66]**						
Total events:	79		74									

Heterogeneity: Tau² = 0.05; Chi² = 5.44, df = 4 (P = 0.24); I² = 26%
Test for overall effect: Z = 0.52 (P = 0.60)
Test for subgroup differences: Not applicable

Risk of bias legend
(A) Bias arising from the randomization process
(B) Bias due to deviations from intended interventions: Mortality
(C) Bias due to missing outcome data: Mortality
(D) Bias in measurement of the outcome: Mortality
(E) Bias in selection of the reported result: Mortality
(F) Overall bias: Mortality

ROB judgements:

+ Low risk of bias
? Some concerns
- High risk of bias

Fig. 1 Forest plot with the risk of bias graph

the forest plot in the review (Figure 1). Stern et al. (2019) provide a useful template for systematic reviewers for judging the ROB in the included studies.

In summary, bias is defined as a systematic error, or deviation from the truth, in results (Higgins 2019). Biases can result in under-estimation or over-estimation of the true intervention effect. Therefore, systematic reviewers should carefully assess and report the ROB in the RCTs included in the review and explore its impact on the meta-analysis.

References

Altman DG, Schulz KF. Statistics notes: Concealing treatment allocation in randomised trials. BMJ (Clinical research ed). 2001;323(7310):446–7.

Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomised trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA. 2010;303(12):1180–7.

Bero LA. Why the Cochrane risk of bias tool should include funding source as a standard item [editorial]. Cochrane Database Syst Rev. 2013;12.

Berry SM, Carlin BP, Connor J. Bias and trials stopped early for benefit. JAMA. 2010;304(2):156; author reply 8-9.

Cai HW, Xia JL, Gao DH, Cao XM. Implementation and experience of a web-based allocation system with Pocock and Simon's minimisation methods. Contemp Clin Trials. 2010;31 (6):510–3.

Clark L, Schmidt U, Tharmanathan P, Adamson J, Hewitt C, Torgerson D. Allocation concealment: a methodological review. J Eval Clin Pract. 2013;19(4):708–12.

da Costa BR, Beckett B, Diaz A, Resta NM, Johnston BC, Egger M, et al. Effect of standardised training on the reliability of the Cochrane risk of bias assessment tool: a prospective study. Syst Rev. 2017;6(1):44.

Day SJ, Altman DG. Statistics notes: blinding in clinical trials and other studies. BMJ (Clinical Research Ed). 2000;321(7259):504.

Dettori J. The random allocation process: two things you need to know. Evid Based Spine-care J. 2010;1(3):7–9.

Hewitt CE, Kumaravel B, Dumville JC, Torgerson DJ. Assessing the impact of attrition in randomised controlled trials. J Clin Epidemiol. 2010;63(11):1264–70.

Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2019.

Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials: a systematic review of trials randomising patients to blind and nonblind sub-studies. Int J Epidemiol. 2014;43(4):1272-83.

Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ (Clinical Research Ed). 2012;344:e1119.

Ioannidis JP, Caplan AL, Dal-Re R. Outcome reporting bias in clinical trials: why monitoring matters. BMJ (Clinical Research Ed). 2017;356:j408.

Jones CW, Keil LG, Holland WC, Caughey MC, Platts-Mills TF. Comparison of registered and published outcomes in randomised controlled trials: a systematic review. BMC Med. 2015;13:282.

Kahan BC, Rehal S, Cro S. Risk of selection bias in randomised trials. Trials. 2015;16:405.

Karanicolas PJ, Farrokhyar F, Bhandari M. Practical tips for surgical research: blinding: who, what, when, why, how? Can J Surg. 2010;53(5):345–8.

Kennedy ADM, Torgerson DJ, Campbell MK, Grant AM. Subversion of allocation concealment in a randomised controlled trial: a historical case study. Trials. 2017;18(1):204.

Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ (Clinical Research Ed). 2010;340:c365.

Kirkham JJ, Altman DG, Chan AW, Gamble C, Dwan KM, Williamson PR. Outcome reporting bias in trials: a methodological approach for assessment and adjustment in systematic reviews. BMJ (Clinical Research Ed). 2018;362:k3802.

Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomised trials in meta-analyses. Ann Intern Med. 2001;135(11):982–9.

Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev. 2017;2:Mr000033.

Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev. 2012;12:Mr000033.

Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome: systematic review with meta-analysis. Intensive Care Med. 2018;44(10):1603–12.

McCambridge J, Sorhaindo A, Quirk A, Nanchahal K. Patient preferences and performance bias in a weight loss trial with a usual care arm. Patient Educ Couns. 2014;95(2):243–7.

McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ (Clinical Research Ed). 2002;324(7351):1448–51.

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet (London, England). 1998;352(9128):609–13.

Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomised trials stopped early for benefit: a systematic review. JAMA. 2005;294(17):2203–9.

Morice V. RandoWeb, an online randomisation tool for clinical trials. Comput Methods Prog Biomed. 2012;107(2):308–14.

Murad MH, Guyatt GH, Domecq JP, Vernooij RWM, Erwin PJ, Meerpohl JJ, et al. Randomised trials addressing a similar question are commonly published after a trial stopped early for benefit. J Clin Epidemiol. 2017;82:12–9.

Nuesch E, Reichenbach S, Trelle S, Rutjes AW, Liewald K, Sterchi R, et al. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. Arthritis Rheum. 2009;61(12):1633–41.

Nunan D, Heneghan C, Spencer EA. Catalogue of bias: allocation bias. BMJ Evid Based Med. 2018a;23(1):20–1.

Nunan D, Aronson J, Bankhead C. Catalogue of bias: attrition bias. BMJ Evid Based Med. 2018b;23(1):21–2.

Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schunemann H, et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev. 2011(4): Mr000012.

Page MJ, McKenzie JE, Kirkham J, Dwan K, Kramer S, Green S, et al. Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions. Cochrane Database Syst Rev. 2014(10):Mr000035.

Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC. Impact of allocation concealment on conclusions drawn from meta-analyses of randomised trials. Int J Epidemiol. 2007;36(4):847–57.

Randomization PN. Part 1: sequence generation. Am J Orthod Dentofac Orthop. 2011;140(5):747–8.

Research AFH, Quality. Methods guide for effectiveness and comparative effectiveness reviews: Agency for Healthcare Research and Quality; 2011.

Savovic J, Jones H, Altman D, Harris R, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technol Assess (Winchester, England). 2012;16(35):1–82.

Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet (London, England). 2002a;359(9306):614–8.

Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet (London, England). 2002b;359(9307):696–700.

Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet (London, England). 2002c;359(9308):781–5.

Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomised trials. Ann Intern Med. 2002;136(3):254–9.

Setia MS. Methodology Series Module 5: Sampling Strategies. Indian J Dermatol. 2016;61 (5):505–9.

Sterne J. Why the Cochrane risk of tool should not include funding source as a standard item [editorial]. Cochrane Database Syst Rev. 2013;12.

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 (Clinical Research Ed). 2019;366:l4898.

Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomised trials? Stat Methods Med Res. 2016:962280216683570.

Wartolowska K, Beard D, Carr A. Blinding in trials of interventional procedures is possible and worthwhile. F1000 Research. 2017;6:1663.

White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ (Clinical Research Ed). 2011;342:d40.

Assessment of Publication Bias

Sven Schulzke

Abstract Ideally, all methodologically sound clinical studies should be published and be included in a systematic review if they adequately address the question at hand. In reality, only a proportion of all initiated studies are completed, and only a proportion of these are published within a reasonable time. Some of the completed studies are never published, and their results tend to systematically differ from those that are published. The reporting bias arising from this phenomenon is termed *publication bias*. Systematic reviewers predominantly rely on data from published studies. Meta-epidemiological studies have shown publication bias as a significant issue in meta-analyses of clinical trials. Publication bias may lead to an over- or under-estimation of the effects of intervention because completed, but unpublished studies are not included in the analysis. This chapter covers the reasons for publication bias and its influence on effect estimates, and the methods (e.g. funnel plot, statistical tests) for assessing and handling this risk.

Keywords Effect estimates · Over-estimation · Funnel plot · Publication bias · Statistical tests · Under-estimation · Visual assessment

Introduction

Ideally, all methodologically sound studies should be published and be included in a systematic review if they adequately address the question at hand. In reality, only a proportion of all initiated studies are completed, and only a proportion of these are published within a reasonable time. Some of the completed studies are never published, and their results tend to systematically differ from those that are published.

S. Schulzke (✉)
Neonatologist, Director of Research, University Children's Hospital Basel UKBB,
Spitalstrasse 33, 4056 Basel, Switzerland
e-mail: sven.schulzke@ukbb.ch

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_5

The reporting bias arising from this phenomenon is termed *publication bias* (Dickersin et al. 1987; Easterbrook et al. 1991). Systematic reviewers predominantly rely on data from published studies. A recent study investigating the frequency of publication bias in meta-analyses published in four major general medical journals (BMJ, JAMA, Lancet, and PLOS Medicine) indicates strong evidence of publication bias in 36% (10/28) of meta-analyses of clinical trials (Kicinski 2013). Further, meta-epidemiological studies demonstrate that publication bias is not assessed in 31% (36/116) of systematic reviews published in the top 10 high-impact factor journals of the general medical literature (Onishi and Furukawa 2014).

Publication bias in meta-analysis of studies assessing treatment effects may lead to an over- or under-estimation of the effects of intervention because completed, but unpublished studies are not included in the analysis (Easterbrook et al. 1991; Kicinski et al. 2015). Typically, published studies are more likely to report beneficial effects of treatment. In contrast, unpublished studies are more likely to confer negative, i.e., non-beneficial or even harmful effects of an intervention (Dickersin et al. 1987; Easterbrook et al. 1991; Kicinski et al. 2015). Current data suggest that in meta-analyses from the Cochrane Database of Systematic Reviews outcomes favouring treatment were 27% more likely to be included than those not favouring treatment and outcomes showing no evidence of adverse effects of intervention were 78% more likely to be included than those reporting adverse effects (Kicinski et al. 2015). This issue is called ‘the file drawer problem’, symbolically describing that studies with negative results remain in ‘the file drawer’, or, at least, stay there for longer while those with positive results are being published fast (Rosenthal 1979). The main reason for this is that investigators tend to refrain from submitting negative studies to scientific journals (Dickersin et al. 1987). To a smaller extent, some journal editors may be less likely to publish negative studies because those studies may not appear to be interesting enough to attract readers (Dickersin 1990).

Another reason for publication bias relates to the funding source of clinical trials. Published industry-sponsored trials tend to more frequently report beneficial effects of treatment compared to studies not funded by industry sponsors (Dickersin 1990). In other words, they are more likely to report efficacy or lack of adverse effects of an intervention –with the intervention typically being a product of the sponsor.

Several instruments related to the conduct and reporting of clinical trials have been developed to reduce the impact of publication bias; for example, numerous international study registries have been established, policies of academic journals have changed to accept reports of clinical trials only if they were registered before enrolling patients and with public access to the study protocol, some journals and authors accept industry-sponsoring only if the sponsor agrees not to have a role in analysis and reporting of results, and data sharing procedures aiming at public access to raw data from clinical trials are currently being discussed.

How to Assess Publication Bias

Funnel Plot

Funnel plots are used to visually detect publication bias (Figs. 1 and 2) (Egger et al. 1997). A funnel plot is a scatter plot displaying the effect estimates of each of the studies included in a meta-analysis. As a rule of thumb, at least 5-10 studies are required in a funnel plot to investigate publication bias (Sedgwick and Marston 2015). The x-axis shows the sample effect estimate of each study. In Figs. 1 and 2, this is given as the risk ratio (RR) of an outcome in treatment vs. control group. The y-axis shows the standard error (SE) of the effect estimate. The SE is a measure of the precision of the RR as an estimate of the population parameter. Typically, small trials with small sample size result in imprecise effect estimates. With increasing sample size, the precision of the effect estimate increases and the size of the SE decreases. The y-axis in the funnel plot is inverted, i.e., smaller studies with less precise effect estimates typically scatter closer to the bottom of the plot. The vertical dotted line represents the summary effect estimate of the meta-analysis. As sample size and precision of effect estimates increase, the horizontal distance of the effect estimates from the summary effect estimate decreases. In the absence of publication bias, we expect the effect estimates of included studies to scatter symmetrically around the overall effect estimate of the meta-analysis due to sampling error, creating a funnel-like shape (Fig. 1). Asymmetrical scattering of studies in the funnel plot indicates potential publication bias (Fig. 2). However, it is impossible to discriminate between publication bias and other types of reporting bias, causing asymmetry of the funnel plot (Sedgwick and Marston 2015). Furthermore, substantial heterogeneity between small and large trials in a meta-analysis may also lead to asymmetry of a funnel plot as studies may measure different effects and, consequently, may scatter in different regions of the plot. When SE is used as a measure of precision, the dotted lines limiting the funnel define the area where 95% of studies are expected to scatter. Alternative measures of precision are in use, e.g., the inverse variance of the effect estimate or sample size, however, using SE is the preferred measure as the expected shape of the plot in the absence of bias is a symmetrical funnel, the plot emphasizes smaller studies which are more prone to bias, and, as mentioned above, the limits of the funnel correspond to the 95% confidence interval (Sterne and Egger 2001).

In Fig. 1, the x-axis shows the risk ratio (RR) of the effect estimate in the treatment vs. control group. The inverted y-axis displays the standard error (SE) of the effect estimate. The 11 studies included in this meta-analysis scatter symmetrically around the vertical dotted line representing the overall effect estimate. Thus, publication bias or other types of reporting bias are not apparent from the funnel plot.

Please note that in Fig. 2, the x-axis shows the risk ratio (RR) of the effect estimate in the treatment vs. control group. The inverted y-axis displays the standard error (SE) of the effect estimate. The ten studies included in this meta-analysis

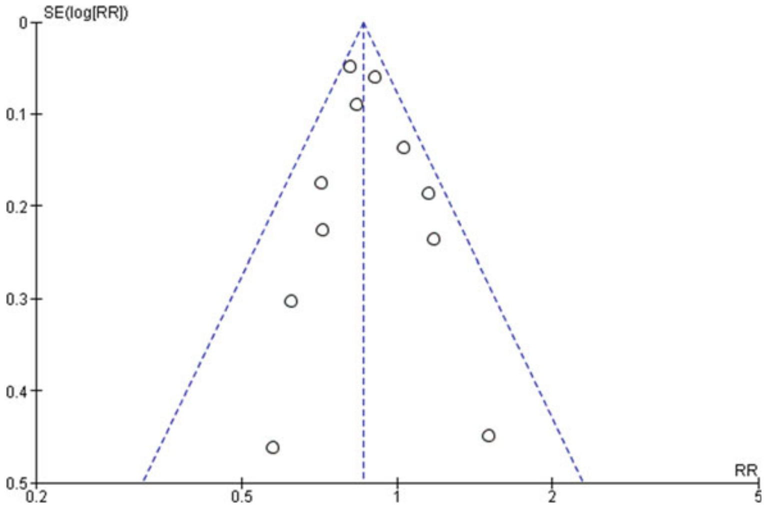


Fig. 1 Funnel plot

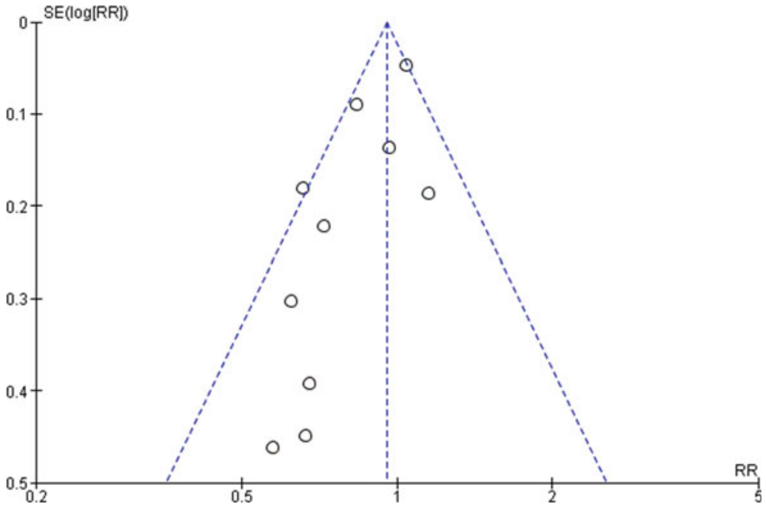


Fig. 2 Funnel plot indicating potential publication bias

scatter asymmetrically around the vertical dotted line representing the overall effect estimate. Thus, the funnel plot indicates potential publication bias.

Formal statistical tests exist for detecting asymmetry in a funnel plot. Most tests statistically determine if there is an association between effect estimate and trial size. Major disadvantages of most statistical tests for detection of publication bias include their low statistical power, and, in the case of regression-based methods, that they mainly assess the impact of small studies rather than publication bias per se. Thus, statistical tests for funnel plot asymmetry should be used with caution and should not be overinterpreted (Mavridis and Salanti 2014).

Egger's Test

The most commonly cited test is the Egger's test (Egger et al. 1997). Egger's test is based on a weighted regression of the effect estimate on its SE with weights inversely proportional to the variance of the effect. The null hypothesis for Egger's test is that the funnel plot is symmetrical in shape. The p-value of Egger's test for the example shown in Fig. 2 is 0.042, i.e., we reject the null hypothesis that the funnel plot is symmetrical at the 5% significance level. We conclude that due to asymmetry in the funnel plot, there is apparent bias in the studies included in this meta-analysis. Simulation studies have challenged the performance of Egger's test, particularly when the summary effect estimate is expressed as the natural logarithm of the odds ratio (lnOR) (Peters et al. 2006). In these cases, alternative regression tests based on a modified Macaskill's test may be more appropriate; details on such alternatives are given in Peters et al. (Peters et al. 2006).

Begg and Mazumdar Rank Correlation Test

The rank correlation test introduced by Begg and Mazumdar is frequently used in meta-analysis (Begg and Mazumdar 1994). This test is a direct statistical analogue of the funnel plot and based on the fact that publication bias will tend to induce a correlation between observed treatment effects and their variances. The test correlates standardized treatment effect with the variance of the treatment effect using Kendall's tau as the measure of association. It is powerful for large meta-analyses (≥ 75 studies) and has moderate power for meta-analyses with 25 studies (Begg and Mazumdar 1994). The test should be interpreted with caution as sensitivity is low, particularly in meta-analysis with few studies. Bias cannot be ruled out if the test is not significant. It is thus considered as a formal, exploratory tool complementing funnel plot inspection for assessment of publication bias. Modified rank correlation tests with improved sensitivity are available and may be particularly helpful in fixed-effects meta-analysis (Gjerdevik and Heuch 2014). When conducting a meta-analysis of observational studies, modified regression methods using a smoothed variance to estimate the precision of a study seem to offer a more robust performance compared to other statistical tests assessing publication bias (Jin et al. 2014).

Dealing with Publication Bias

A variety of approaches exist for dealing with publication bias.

(1) *Ignore publication bias*

One could consider ignoring publication bias. This is not recommended as such an approach may influence the overall effect estimate and, consequently, interpretation of a meta-analysis.

(2) *Do not pool studies for meta-analysis*

Denoting the other extreme, authors might consider not conducting a meta-analysis due to potential publication bias. While this option may be reasonable in cases of an extraordinary level of publication bias, it is too radical in many other circumstances.

(3) *Explore and potentially adjust for publication bias*

Recommended approaches include describing potential reasons and impact of publication bias in the results and interpretation of a meta-analysis and, potentially, adjusting the results of a meta-analysis to quantify the effects of publication bias (Mavridis and Salanti 2014).

The trim and fill method is an approach that attempts to identify and adjust the results of a meta-analysis for publication bias (Duval and Tweedie 2000). In the first step, it omits small studies (trimming) until the funnel plot becomes symmetrical. Thereby, an adjusted overall effect estimate is produced from the remaining studies. Then, the funnel plot is redrawn with the omitted studies replaced and their 'missing' equivalents added on the opposite side of the plot (filling). The funnel plot is now symmetrical around the adjusted overall effect estimate. Appealingly, the trim-and-fill method provides a summary effect adjusted for publication bias and also estimates the number of unpublished studies. However, it is based on the potentially incorrect assumption that asymmetry in the funnel plot is solely caused by publication bias and performs poorly in the presence of substantial heterogeneity (Mavridis and Salanti 2014; Terrin et al. 2003). Furthermore, the mechanism causing publication bias is unknown, and we do not know whether the 'filled' studies really exist. More complex techniques for exploring publication bias are available. For example, **selection models** attempt to simulate the publication process and produce observed effect sizes acknowledging that the overall effect is conditional to the observed studies being published (Mavridis and Salanti 2014). Afterwards, they calculate the so-called marginal effect size, which is the effect size unconditional to the publication status. Thus, they provide adjusted estimates via sensitivity analysis (Copas and Shi 2000; Copas and Shi 2001). However, their theoretical background and practical applications are not easily accessible to non-statisticians.

References

Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50(4):1088–101.

Copas JB, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. Biostatistics. 2000;1:247–62.

Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. Stat Meth Med Res. 2001;10:251–65.

Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA. 1990;263(10):1385–9.

Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H Jr. Publication bias and clinical trials. Control Clin Trials. 1987;8(4):343–53.

Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455–63.

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet. 1991;337(8746):867–72.

Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–34.

Gjerdevik M, Heuch I. Improving the error rates of the Begg and Mazumdar test for publication bias in fixed effects meta-analysis. BMC Med Res Methodol. 2014;14:109.

Jin ZC, Wu C, Zhou XH, He J. A modified regression method to test publication bias in meta-analyses with binary outcomes. BMC Med Res Methodol. 2014;14:132.

Kicinski M. Publication Bias in Recent Meta-Analyses. PLoS ONE. 2013; 8(11):e81823.

Kicinski M, Springate DA, Kontopantelis E. Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. Stat Med. 2015;34(20):2781–93.

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.

Onishi A, Furukawa TA. Publication bias is underreported in systematic reviews published in high-impact-factor journals: metaepidemiologic study. J Clin Epidemiol. 2014;67(12):1320–6.

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. JAMA. 2006;295(6):676–80.

Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull. 1979;86(3):638–41.

Sedgwick P, Marston L. How to read a funnel plot in a meta-analysis. BMJ. 2015;351.

Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol. 2001;54(10):1046–55.

Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. Stat Med. 2003;22(13):2113–26.

Data Extraction from Included Studies

Kwi Moon and Shripada Rao

Abstract Accurate data extraction and their synthesis form the basis of appropriate conclusions of a systematic review. Systematic reviewers should extract ALL data relevant to the review question, not just the outcome data. Data to be extracted include baseline characteristics of study participants, information related to study methodology and outcomes and other relevant information. If published articles have given the results using figures instead of actual numbers, specialised software that convert images to pixel values may be utilised to obtain the actual data values. Tools such as Plot Digitizer, WebPlotDigitizer, Engauge, Dexter, Ycasd and GetData Graph Digitizer can be used for this purpose. When unable to extract data from available reports or to seek clarifications, the reviewers could contact the original investigators. Data extraction should be performed using pre-piloted forms independently by at least two reviewers to ensure accuracy. A high level of diligence is required to minimise errors during the stage of data extraction.

Keywords Comparisons · Covidence · Descriptive · Entry · Error · Interventions · Outcomes · Participant · Source

K. Moon (✉)
Department of Pharmacy, Perth Children's Hospital, 15 Hospital Ave, Nedlands,
Perth, WA 6009, Australia
e-mail: kwi.moon@health.wa.gov.au

S. Rao
Neonatal Directorate, Perth Children's Hospital, 15 Hospital Ave, Nedlands, Perth,
WA 6009, Australia
e-mail: shripada.rao@health.wa.gov.au

K. Moon · S. Rao
School of Medicine, University of Western Australia, Perth, WA 6009, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_6

Introduction

Data extraction is an important aspect of a systematic review because accurate data and their synthesis form the basis of appropriate conclusions (Li et al. 2015). Data collected for systematic reviews should be accurate, complete, and accessible for future updates of the review and data sharing (https://training.cochrane.org/handbook/current/chapter-05#section-5-1). It is essential to pilot the data collection form before beginning data entry. The completed data collection forms should be provided to the editors of journals or the Cochrane review group upon request.

Which Data to Extract

Systematic reviewers should extract ALL data relevant to the review question, and not just the outcome data.

a. *Descriptive data of individual included studies:* Information on authors, settings, study design, characteristics of participants, details of the intervention, outcomes, sample size, funding source and the reason for inclusion or exclusion should be collected. Such detailed extraction and reporting of the descriptive data will enable clinicians to establish the generalizability of the results. Descriptive data are also important to the reviewer and enables them to understand and explore heterogeneity (Munn et al. 2014).
b. *Outcome data:* The outcome data should be collected separately for each outcome. The details should include the raw numbers (numerator and denominator), statistical measures such as relative risks, odds ratios, weighted means, standard deviations and confidence intervals. A standardised approach should be used while extracting data. Examples of data collection templates are available from organisations such as Cochrane Collaboration, Joanna Briggs Institute (JBI) and BMJ group.
c. *Data to assess the risk of bias:* It is vital to collect information to assess the risk of bias in the included studies. For example, while conducting a systematic review of RCTs, it is essential to gather information on methods used for generation of random sequences, allocation concealment, blinding, completeness of follow up and any other sources of bias. Studies with a high risk of bias decrease their internal validity leading to erroneous conclusions. The details have been covered in the chapter titled “Assessment of Risk of Bias”.

How to Minimise Errors in Data Extraction

Errors in data extraction can alter the results and conclusions of the review, and hence utmost diligence is required during this stage.

Jones et al. retrospectively repeated the data extraction in all systematic reviews conducted by the Cochrane Cystic Fibrosis and Genetic Disorders Group using the same articles that were used by the original Cochrane reviewers (Jones et al. 2005). They reported that errors were found in 20 of 34 reviews, including incorrect calculations made when converting data in primary articles into data required for the review and misinterpretation of data that were reported in the primary article (Jones et al. 2005). In another study, Carroll et al. (2013) evaluated differences in the data extracted by three different systematic reviews comparing total hip arthroplasty versus hemiarthroplasty in osteoarthritis. The authors reported that 8–42% of the data differences between the reviews resulted from the selection of alternative reported data, while 8–17% of the differences resulted from data errors. They concluded that systematic reviewers should use double-data extraction to minimise error and make every effort to clarify or explain their choice of data (Carroll et al. 2013). Buscemi et al. found that single data extraction resulted in more errors than double data extraction (relative difference: 21.7%, P = .019) (2006). Mathes et al. identified six studies that had addressed the issue of errors in data extraction (2017). They found a high rate of extraction errors (up to 50%), and often the errors influenced effect estimates.

The Institute of Medicine (IOM) recommends that review authors should, "*at a minimum, use two or more independent researchers to extract quantitative and other critical data from each study*" (Eden et al. 2011). The Cochrane Handbook also makes similar recommendations (Higgins et al. 2019). Any disagreements between authors are to be resolved by discussion among all authors or by consulting a senior author.

Sources to Obtain Data

Reports from included studies are the major source of data for systematic reviews. Such reports may be published or unpublished (e.g. Journal articles, conference abstracts, dissertation, and online clinical trial registries). It is important to be aware that conference abstracts may have preliminary findings only. Sometimes outcome data may be given only as figures in the published manuscripts. Specialised software that converts images to pixel values may be utilised to obtain more accurate data values (Vucic et al. 2015). Tools such as Plot Digitizer, WebPlotDigitizer, Engauge, Dexter, ycasd and GetData Graph Digitizer can be used for this purpose. The software takes an image of a figure and then digitising the data points off the figure using the axes and scales set by the users (https://training.cochrane.org/handbook/current/chapter-05#section-5-5-8).

When unable to extract information from available reports or to seek clarifications, the reviewers need to contact the original investigators. Young and Hopewell (2011) reported that email correspondence with authors resulted in the greatest response (Young and Hopewell 2011). The Cochrane handbook recommends that obtaining unpublished data is highly desirable and potentially increases precision and minimises the impact of reporting biases (https://training.cochrane.org/handbook/current/chapter-05#section-5-2-3). **It is vital to pay special attention to 'errata' from published studies.** Hauptman et al. reviewed the frequency and significance of published errata in 20 general medicine and cardiovascular journals (median impact factor 5.52) over 18 months. They found that 557 articles were associated with errata reports (overall errata report rate 4.2 per 100). At least one significant error that materially altered data interpretation was present in 24.2% of articles with errata (Hauptman et al. 2014).

Where to Enter the Data

Data collection forms promote standardised approach for data extraction and address the review question/assessment criteria directly, providing a clear summary. Furthermore, the forms create a historical record of data collection and decisions made throughout the review process, including the final statistical data for meta-analyses. Depending on the author's preferences, data collection forms can be electronic (e.g. Microsoft Excel) or paper-based. A generic template may be used at the beginning for testing and then updated by reviewers to ensure that the form meets their needs. ***Covidence*** is primary screening and data extraction tool recommended by Cochrane for Cochrane authors (https://www.covidence.org/home). It allows authors to upload search results, screen abstracts and full text, complete data collection, conduct risk of bias assessment and export data into Revman or Excel. For complex systematic reviews, **EPPI-Reviewer** is useful for data collection and other aspects of a systematic review. The Covidence and EPPI-Reviewer can be accessed free of cost by the Cochrane reviewers. For independent systematic reviewers, they are subscription-based. Example of data collection items to be included in the systematic review is shown in Table 1.

Automating Data Extraction

Manual extraction of the data is slow, costly and subject to human error (Bui et al. 2016). Automating or semi-automating this step has the potential to decrease the time taken to complete systematic reviews and thus decrease the time lag for research evidence to be translated into clinical practice (Jonnalagadda et al. 2015). Natural language processing (NLP), including text mining, involves information extraction, which is the discovery by computer of information by automatically extracting

Table 1 Data collection items for studies included in the systematic review

• **Name of data extractor/s, date of data extraction**
• **Source**: Journal name, year of publication; Conference name, year
• **Setting**, **Country**
• **Title** of the article
• **Inclusion and exclusion criteria**
• **Study design**: RCT, cluster RCT, case-control, cohort, others
• **Years of conduct**
• **Duration of follow up**
Methods
• For a generation of random sequence numbers; concealment of allocation sequence, blinding
• For statistical analysis
Participants: Baseline characteristics (e.g. age, sex, weight, comorbidity, socioeconomic status)
Intervention: Details of intervention (e.g. drug dose, frequency, route, and duration)
Control: Details (e.g. no intervention, placebo, standard care)
Description of co-interventions
Outcomes
For each pre-specified outcome (e.g. mortality, morbidity) in the systematic review:
• Definition, Timing of measurements
• Adverse outcomes
Results
For each group, and for each outcome at each time point:
• Number of participants assigned
• Number of participants included in the analysis
• Number of participants who withdrew or excluded
• Number who were lost to follow-up
• Summary data for each group (e.g. 2 × 2 table for dichotomous data; means and standard deviations for continuous data)
• Between-group effect size estimates (e.g. risk ratio, odds ratio, mean difference)
Miscellaneous
• Study authors' conclusions
• Correspondence required for clarification
• Comments from the study authors or by the review authors
• Funding source
• Authors' potential conflicts of interest

information from different written resources (Hearst 1999). NLP techniques have been used to automate the extraction of genomic and clinical information from biomedical literature. Similarly, automation of the data extraction step of the systematic review process through NLP may decrease the time necessary to complete a systematic review (Jonnalagadda et al. 2015). In a recent study, Bui et al. developed a computer system that used machine learning and natural language processing approaches to generate summaries of full-text scientific publications (Bui et al. 2016) automatically. The summaries at the sentence and fragment levels were evaluated in finding common clinical SR data elements such as sample size, group size, and PICO values. They compared the computer-generated summaries with human-written summaries (title and abstract) in terms of the presence of necessary information for the data extraction as presented in the Cochrane review's study characteristics tables. They found that at the sentence level, the computer-generated summaries covered more information than humans do for systematic reviews. They concluded that machine learning and natural language processing are promising approaches to the development of such an extractive summarisation system (Bui et al. 2016). In the long run, these new approaches to evidence synthesis, which use human effort and machine automation in mutually reinforcing ways, can enhance the feasibility and sustainability of "*living systematic reviews*" (Thomas et al. 2017).

Conclusions

Data extraction is a critical step while conducting a systematic review. A high level of diligence is required to minimise errors during this stage.

References

Bui DDA, Del Fiol G, Hurdle JF, Jonnalagadda S. Extractive text summarisation system to aid data extraction from full text in systematic review development. J Biomed Inform. 2016;64:265–72.

Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol. 2006;59 (7):697–703.

Carroll C, Scope A, Kaltenthaler E. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. BMC Res Notes. 2013;6:539.

Eden J, Levit L, Berg A, Morton S. committee on standards for systematic reviews of comparative effectiveness research; Institute of Medicine. In: Finding what works in health care: standards for systematic reviews. Washington, DC: The National Academies Press; 2011.

Hauptman PJ, Armbrecht ES, Chibnall JT, Guild C, Timm JP, Rich MW. Errata in medical publications. Am J Med. 2014;127(8):779–85.

Hearst MA. Untangling text data mining. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics; College Park, Maryland, 1034679: Association for Computational Linguistics; 1999. pp 3–10.

Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. John Wiley & Sons; 2019.

Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol. 2005;58(7):741–2.

Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev. 2015;4:78.

Li T, Vedula SS, Hadar N, Parkin C, Lau J, Dickersin K. Innovations in data collection, management, and archiving for systematic reviews. Ann Intern Med. 2015;162(4):287–94.

Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol. 2017;17(1):152.

Munn Z, Tufanaru C, Aromataris E. JBI's systematic reviews: data extraction and synthesis. Am J Nursing. 2014;114(7):49–54.

Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol. 2017;91:31–7.

Vucic K, Jelicic Kadic A, Puljak L. Survey of Cochrane protocols found methods for data extraction from figures not mentioned or unclear. J Clin Epidemiol. 2015;68(10):1161–4.

Young T, Hopewell S. Methods for obtaining unpublished data. Cochrane Database Syst Rev. 2011(11):Mr000027.

Fixed and Random-Effects Models for Meta-Analysis

Ravisha Srinivasjois

Abstract Results of a randomised controlled trial (RCT) may differ from other similar RCTs despite best efforts in study design and conduct. This is because some heterogeneity is inevitable as no two individuals are identical, and responses to interventions vary. A meta-analysis of 'more or less similar' studies generates a more reliable summary estimate to better predict the true population effect because of the improved power and precision. Meta-analysis involves assigning 'weight' to each included study based on various factors, including the sample size, and observed variance. The weight assigned to each study differs based on the model chosen to generate the pooled effect estimate. Judging the effect of heterogeneity on the results of included studies is crucial for selecting the right model for meta-analysis. The choice of the model affects the outcomes of the summary estimate. This chapter covers the key assumptions, characteristics and rationale for selection of the fixed effect and random effects model for analysis.

Keywords Meta-analysis · Fixed effect · Random effects · Heterogeneity · Pooled estimate · Forest plot

Introduction

The next step after finalising the studies eligible for pooling the data is meta-analysis to obtain the summary estimate. Selection of the model for meta-analysis is important in this context (Deeks et al. 2008). Thorough understanding of this process is necessary before discussing the models available for meta-analysis.

Randomised controlled trials (RCTs) are considered as the gold standard for clinical research (Sibbald and Roland 1998). RCTs evaluate the effect of an intervention in the selected study population and *analyse how it differs from the true*

R. Srinivasjois (✉)
School of Medicine, University of Western Australia, Perth, WA 6027, Australia
e-mail: Srinivasjoisr@ramsayhealth.com.au

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_7

effect observed in the *entire* population (Sibbald and Roland 1998). Explicit and rigid inclusion and exclusion criteria, and the process of randomisation try to ensure that the characteristics of participants are similar and the distribution of known as well as unknown confounders are balanced between the intervention and control arms of the trial. Despite this, the observed results of one RCT may be different from other similar RCTs (Kendall 2003). This is because some heterogeneity is inevitable as no two individuals are identical, and responses to interventions vary. Factors such as differences in participant characteristics (e.g. severity of illness, age, and gender), mode of delivery of the intervention, and unique aspects of the settings result in clinical heterogeneity (Deeks et al. 2008). On the statistical side, small RCTs do not have an adequate power to estimate the clinically significant minimum true effect of the intervention in the study population if an effect does exist. Large RCTs are hence expected to provide an estimate that is closer to the truth - 'the true population estimate'. Conducting large adequately powered RCTs is challenging; the reason they are less frequent compared to small trials. A meta-analysis of *'more or less similar'* studies helps in generating a more reliable summary estimate to predict the true population effect because of the improved power and precision (Schmidt et al. 2009; Zwahlen et al. 2008).

Meta-analysis involves assigning 'weight' to each included study based on various factors including the sample size, standard error of the mean, and observed variance. The weight assigned to each study differs based on the model chosen to generate the pooled effect estimate (Borenstein et al. 2010). Judging the effect of clinical and statistical heterogeneity on the results of studies included in the systematic review is crucial for selecting the model for meta-analysis. The choice of the model affects the outcomes of the summary estimate. This chapter covers the key assumptions, characteristics and rationale for the selection of the fixed-effect model and random-effects model for meta-analysis (Table 1).

Fixed Effect Model

The fixed-effect model assumes that the true effect size is identical for all included studies (Borenstein et al. 2010). Under this model (Nikolakopoulou et al. 2014), any differences in individual study results are assumed to be due to a random error or sampling error (Schmidt et al. 2009). The model suggests that differences in participant characteristics and intervention have minimal or no effect on the observed result. Hence, during meta-analysis, larger studies are assigned higher weight because they are possibly closer to the true effect than the smaller studies. Thus large studies have a stronger influence on the summary effect using the fixed effect model (Schmidt et al. 2009).

Table 1 Models for meta-analysis

	Fixed effect model	Random effects model
Assumption	True effect size is identical for all studies	True effect is different for each study
Weighting	Larger studies are assigned higher weightage, and smaller studies are assigned lower weightage	Large studies are assigned relatively lower weight, and small studies are assigned a relatively higher weight
Null hypothesis	Tests for the null effect in each included study	Tests for the 'mean effect' derived from the observed results
Choice	Choose if the included studies are homogeneous (i.e. no significant clinical or statistical heterogeneity)	Choose if included studies are heterogeneous
Pooling method	Mantel-Haenszel, Peto or inverse variance method	DerSimonian and Laird inverse variance method
Confidence intervals of the summary estimate	Usually tight	Usually wide

Random Effects Model

In this model, each individual study is considered different from other studies. In addition to the random error, the differences in observed effect sizes are considered to be due to variation in true effect (Nikolakopoulou et al. 2014). Each study is different because of the differences in the study population, intervention (e.g. timing, mode of delivery), and outcomes. This means that the studies are considered 'heterogeneous'. Under this model, the pooled effect estimate assesses the *mean of the distribution of effects* observed in individual studies. Hence, even if the study is small, it contributes to the summary effect in a way that no other study can contribute. Thus, smaller studies receive a higher weight in a meta-analysis using the random effects model (Schmidt et al. 2009; Nikolakopoulou et al. 2014).

The summary estimates computed in the random-effects model vs. the fixed-effect model differ partly because of the differences in the weight given to the small studies. Because the random-effects model assumes that the individual studies are heterogeneous, the variance, standard error and confidence intervals for the summary estimate are wider compared with those generated by the fixed effect model. The summary estimate generated by the random-effects model is more generalisable to other similar populations than the fixed-effect model. The two models also differ with regards to two other aspects of meta-analysis as follows:

(1) **The null hypothesis on the summary estimate**: The random-effects model tests the hypothesis that the 'mean effect' derived from the observed results is null, whereas the fixed effect model tests for the null effect in each study. Although this is not routinely considered while choosing the model for meta-analysis, it is important to know these differences in computation.

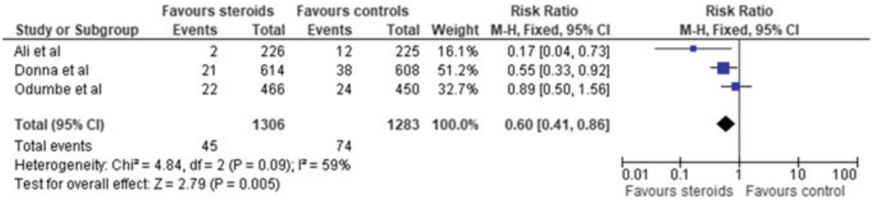
(2) **The method used for combining the dichotomous outcomes**: The fixed effect model uses the Mantel-Haenszel (Mantel and Haenszel 1959), Peto (Yusuf et al. 1985) or inverse variance method (Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019; Fleiss 1993) for meta-analysis whereas the random-effects model uses the DerSimonian and Laird inverse variance method (DerSimonian and Laird 1986) for this purpose. Each of these methods has specific advantages and disadvantages. The Peto method (Yusuf et al. 1985) can only combine odds ratios while the other three methods can combine odds ratios, risk ratios or risk differences. Mantel-Haenszel method is the most commonly used method in fixed-effect (Mantel and Haenszel 1959).

As mentioned earlier, selecting the appropriate model for summary estimates is a critical decision. Aiming to choose the 'best' model is not the correct strategy. It is crucial to decide which model best suits the question being addressed in the systematic review. If the studies included in the meta-analysis are essentially similar, the fixed-effect model is appropriate for deriving the summary estimate. If the studies are considered to be heterogeneous, the random-effects model is preferred (Deeks et al. 2008). If the included studies are 'too dissimilar' temptation for a meta-analysis should be avoided.

In general, if the studies have statistical (significant values for Chi-square test and I^2 statistic) and clinical (as discussed above) heterogeneity, the random-effects model should be the appropriate choice. Some researchers first analyse the data using the fixed-effect model and then cross-check the results with a random-effects model. This provides the readers with information from both models, helping to derive their conclusions. However, experts have discouraged this practice (Borenstein et al. 2010). At times, the summary estimates can be statistically significant under one model but remain non-significant in the other model. In such situations, the reviewer needs to decide which model is appropriate for the question and discuss this in the paper. The difference in the summary effects of the two models is demonstrated below using an example.

Example: Investigators assessed the risk of admission to the neonatal intensive care unit (NICU) of infants born by an elective caesarean after antenatal glucocorticoid exposure at or beyond 37 weeks of gestation.

a. **Fixed effect model**

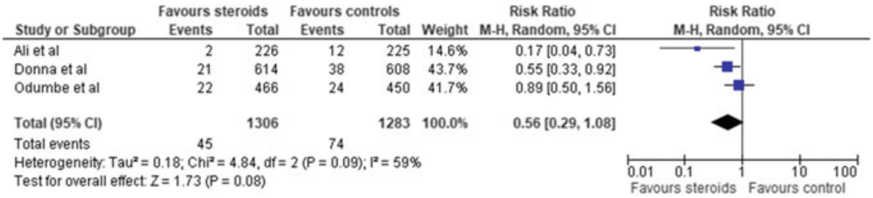


Study or Subgroup	Favours steroids Events	Favours steroids Total	Favours controls Events	Favours controls Total	Weight	Risk Ratio M-H, Fixed, 95% CI
Ali et al	2	226	12	225	16.1%	0.17 [0.04, 0.73]
Donna et al	21	614	38	608	51.2%	0.55 [0.33, 0.92]
Odumbe et al	22	466	24	450	32.7%	0.89 [0.50, 1.56]
Total (95% CI)		**1306**		**1283**	**100.0%**	**0.60 [0.41, 0.86]**
Total events	45		74			

Heterogeneity: Chi² = 4.84, df = 2 (P = 0.09); I² = 59%

Test for overall effect: Z = 2.79 (P = 0.005)

b. **Random effects model**



Study or Subgroup	Favours steroids Events	Favours steroids Total	Favours controls Events	Favours controls Total	Weight	Risk Ratio M-H, Random, 95% CI
Ali et al	2	226	12	225	14.6%	0.17 [0.04, 0.73]
Donna et al	21	614	38	608	43.7%	0.55 [0.33, 0.92]
Odumbe et al	22	466	24	450	41.7%	0.89 [0.50, 1.56]
Total (95% CI)		**1306**		**1283**	**100.0%**	**0.56 [0.29, 1.08]**
Total events	45		74			

Heterogeneity: Tau² = 0.18; Chi² = 4.84, df = 2 (P = 0.09); I² = 59%

Test for overall effect: Z = 1.73 (P = 0.08)

The three studies included in the meta-analysis are RCTs. Participants were full-term pregnant mothers with gestation above 37^{+0} weeks. The study group received two doses of antenatal corticosteroids. The interval from completing the corticosteroids to the time of delivery was not available. Clinical information on maternal comorbidities (e.g. gestational diabetes, preeclampsia), and reasons for the infant's admission to the NICU was not available in all included studies, resulting in clinical heterogeneity. Statistical heterogeneity assessed by I^2 statistic was high (59%).

The two models resulted in different weightage for the included trials (Fig. a, b). In the random-effects model, the study by Donna et al. received lower, and Odumbe et al. received higher weightage compared with the fixed effect model. The fixed-effect model meta-analysis, which does not account for heterogeneity, resulted in an overall summary estimate (Risk ratio) of 0.60, demonstrating a reduction in the risk of infant's admission to NICU. Statistical significance was achieved with a 95% confidence interval of 0.41–0.86. The random-effects model showed marginally better-pooled effect estimate for the antenatal corticosteroid exposed group. However, the confidence interval was wider (0.29 to 1.08), rendering the summary estimate non-significant compared with the statistically significant results of the fixed-effect model.

In summary, judgement about heterogeneity (participants, intervention, control, outcome, design, and setting) of the studies selected for pooling the data is important in selecting the model for meta-analysis. A random-effects model is preferred in the presence of clinical heterogeneity. The temptation of choosing the model based on the statistical significance of the summary effect is best avoided as it can provide a wrong estimate of the effect size, its confidence intervals, and significance. Experts have pointed out that understanding the caveats of both, the fixed effect model and the random-effects model is important for conducting meta-analysis (Borenstein et al. 2010). Pending further research for developing a suitable alternative, presenting results of both, the random effects and fixed-effect model meta-analysis and letting the readers be the judge may be an appropriate strategy (Sera et al. 2019; Bakbergenuly and Kulinskaya 2018).

References

Bakbergenuly I, Kulinskaya E. Meta-analysis of binary outcomes via generalised linear mixed models: a simulation study. BMC Med Res Methodol. 2018;18:70. https://doi.org/10.1186/s12874-018-0531-9.

Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Meth. 2010;1(2):97–111.

Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane; 2019. www.training.cochrane.org/handbook.

Deeks JJ, Higgins JT, Altman DG. Analysing data and understanding meta-analysis. Wiley-Blackwell; 2008.

DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88.

Fleiss JL. the statistical basis of meta-analysis. Stat Meth Med Res. 1993;2:121–45.

Kendall JM. Designing a research project: randomised controlled trials and their principles. Emerg Med J. 2003;20:164–8.

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22(4):719–48.

Nikolakopoulou A, Mavridis D, Salanti G. How to interpret meta-analysis models: fixed effect and random effects meta-analyses. Evidence-Based Mental Health. 2014;17(2):64.

Schmidt FL, Oh IS, Hayes TL. Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. Br J Math Stat Psychol. 2009;62(Pt 1):97–128.

Sera F, Armstrong B, Blangiardo M, Gasparrini A. An extended mixed-effects framework for meta-analysis. Stat Med. 2019;38(29):5429–44. https://doi.org/10.1002/sim.8362 Accessed 16 Sep 2020.

Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? BMJ. 1998;316(7126):201.

Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomised trials. Prog Cardiovasc Dis. 1985;27(5):335–71.

Zwahlen M, Renehan A, Egger M. Meta-analysis in medical research: potentials and limitations. Urologic Oncol. 2008;26(3):320–9.

Forest Plots in a Meta-Analysis

Sanjay Patole

Abstract Generating forest plots is the next step after extracting data from studies eligible for meta-analysis. A forest plot displays the effect estimates and confidence intervals of individual studies and their meta-analysis. The key feature of the forest plot is the pooled effect estimate represented by the much sought after'diamond'. However, it is important to note that meta-analysis is not justified unless all potentially eligible studies have comparable clinical and methodological characteristics, and are addressing the question at the core of the systematic review. Failure to pay attention to this vital consideration leads to what is commonly called *"garbage in and garbage out"*. Assuring that the extracted data are in a suitable format and choosing the appropriate model (fixed effect vs random effects) for meta-analysis are other important considerations. This chapter is focussed on forest plots in a meta-analysis and provides a 10-point checklist for their assessment and interpretation.

Keywords Pooled effect estimate · Confidence intervals · Fixed effect model · Random-effects model · Meta-analysis · Checklist

Introduction

Having extracted the data from studies eligible for meta-analysis, the next step is to generate the forest plots for deriving the pooled estimates of the outcome of interest using meta-analysis software such as the Review manager (RevMan, Cochrane Collaboration, Nordic Cochrane Centre). For dichotomous outcomes, the extracted data will be the number of participants with the event and the number analysed in each treatment group of each study. For continuous outcomes, it will be the mean

S. Patole (✉)
School of Medicine, University of Western Australia, Perth, WA 6009, Australia
e-mail: sanjay.patole@health.wa.gov.au

S. Patole
Neonatal Directorate, King Edward Memorial Hospital for Women,
Perth, WA 6008, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_8

and the standard deviation. If required, the mean and the standard deviation could be derived from median and range or the median and interquartile range by using the Hozo and Wan formula respectively (Hozo et al. 2005; Wan et al. 2014).

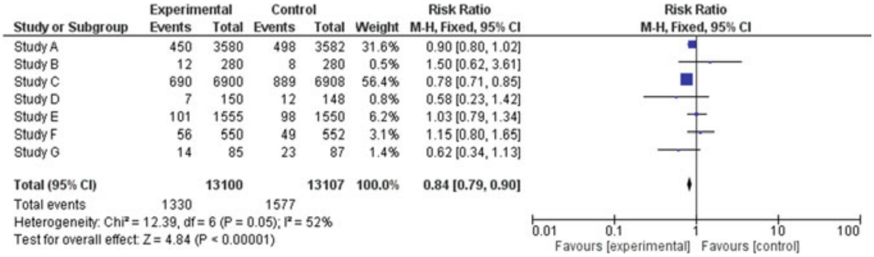
As discussed elsewhere in this book, selecting the model for meta-analysis (Fixed effect vs Random effects) is an important decision. Careful consideration of the pre-stated criteria for study selection, and characteristics (PICOS) of the included studies helps in this process. Meta-analysis of the data from studies included in the systematic review is not justified unless all have comparable PICOS characteristics, and are addressing the question at the core of the systematic review. Failure to pay attention to this vital consideration leads to what is commonly called *"garbage in and garbage out"*.

Generating the Forest Plot

The technical aspects of entering the data to generate a forest plot are not difficult. However, it is important to check for transcription errors and avoid transposition errors (i.e. correct data entered in the wrong column!)

What Is a Forest Plot?

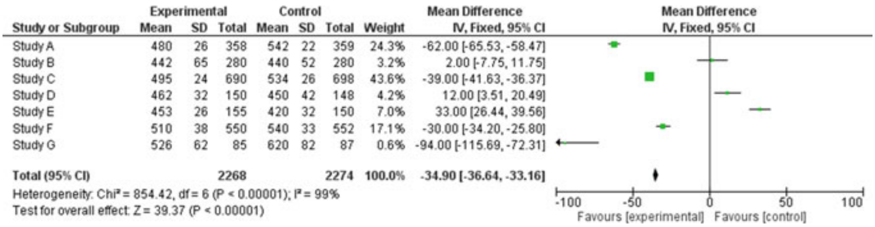
A forest plot displays the effect estimates and confidence intervals (CI) of individual studies and their meta-analysis (Lewis and Clarke 2001) (Figs. 1, 2, 3 and 4) Each study is represented by a square block at the point estimate of the effect of the intervention with a horizontal line extending on either side. The size of the block indicates the weight assigned to that study in the meta-analysis while the horizontal line represents the uncertainty (95% CI) around the effect estimate. The length of the horizontal lines reflects whether the CI is tight or wide. The result of each of the included studies is thus a '*whisker plot'* in the forest plot. The pooled effect estimate is represented by a diamond. The centre of the diamond shows the point estimate, and its tips show the 95% CI around the summary estimate.



	Experimental		Control			Risk Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Fixed, 95% CI
Study A	450	3580	498	3582	31.6%	0.90 [0.80, 1.02]
Study B	12	280	8	280	0.5%	1.50 [0.62, 3.61]
Study C	690	6900	889	6908	56.4%	0.78 [0.71, 0.85]
Study D	7	150	12	148	0.8%	0.58 [0.23, 1.42]
Study E	101	1555	98	1550	6.2%	1.03 [0.79, 1.34]
Study F	56	550	49	552	3.1%	1.15 [0.80, 1.65]
Study G	14	85	23	87	1.4%	0.62 [0.34, 1.13]
Total (95% CI)		**13100**		**13107**	**100.0%**	**0.84 [0.79, 0.90]**
Total events	1330		1577			

Heterogeneity: $Chi^2 = 12.39$, $df = 6$ ($P = 0.05$); $I^2 = 52\%$
Test for overall effect: $Z = 4.84$ ($P < 0.00001$)

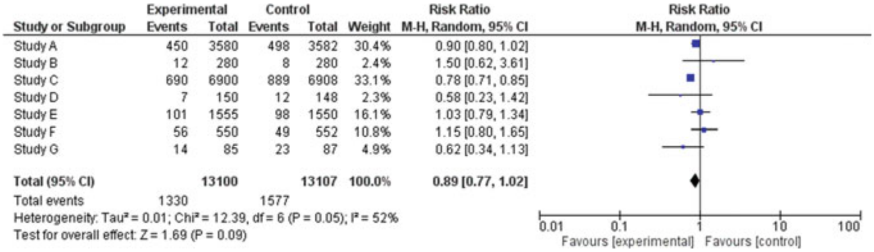
Fig. 1 Forest plot for a categorical outcome (FEM)*. (FEM: Fixed effect model) *Note* M-H: Mantel-Haenszel, df: Degree of freedom



	Experimental			Control				Mean Difference
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Fixed, 95% CI
Study A	480	26	358	542	22	359	24.3%	-62.00 [-65.53, -58.47]
Study B	442	65	280	440	52	280	3.2%	2.00 [-7.75, 11.75]
Study C	495	24	690	534	26	698	43.6%	-39.00 [-41.63, -36.37]
Study D	462	32	150	450	42	148	4.2%	12.00 [3.51, 20.49]
Study E	453	26	155	420	32	150	7.0%	33.00 [26.44, 39.56]
Study F	510	38	550	540	33	552	17.1%	-30.00 [-34.20, -25.80]
Study G	526	62	85	620	82	87	0.6%	-94.00 [-115.69, -72.31]
Total (95% CI)			**2268**			**2274**	**100.0%**	**-34.90 [-36.64, -33.16]**

Heterogeneity: Chi² = 854.42, df = 6 (P < 0.00001); I² = 99%
Test for overall effect: Z = 39.37 (P < 0.00001)

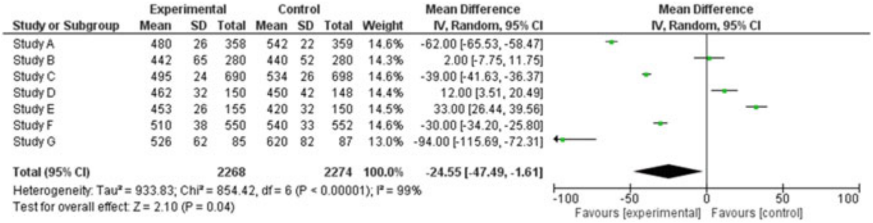
Fig. 2 Forest plot for a continuous outcome using (FEM)* (FEM: Fixed effect model) *Note* IV: Inverse variance, df: Degree of freedom



	Experimental		Control			Risk Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Random, 95% CI
Study A	450	3580	498	3582	30.4%	0.90 [0.80, 1.02]
Study B	12	280	8	280	2.4%	1.50 [0.62, 3.61]
Study C	690	6900	889	6908	33.1%	0.78 [0.71, 0.85]
Study D	7	150	12	148	2.3%	0.58 [0.23, 1.42]
Study E	101	1555	98	1550	16.1%	1.03 [0.79, 1.34]
Study F	56	550	49	552	10.8%	1.15 [0.80, 1.65]
Study G	14	85	23	87	4.9%	0.62 [0.34, 1.13]
Total (95% CI)		**13100**		**13107**	**100.0%**	**0.89 [0.77, 1.02]**
Total events	1330		1577			

Heterogeneity: Tau² = 0.01; Chi² = 12.39, df = 6 (P = 0.05); I² = 52%
Test for overall effect: Z = 1.69 (P = 0.09)

Fig. 3 Forest plot for a categorical outcome using (REM)* (REM: Random-effects model) *Note* M-H: Mantel-Haenszel, df: Degree of freedom



	Experimental			Control				Mean Difference
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Random, 95% CI
Study A	480	26	358	542	22	359	14.6%	-62.00 [-65.53, -58.47]
Study B	442	65	280	440	52	280	14.3%	2.00 [-7.75, 11.75]
Study C	495	24	690	534	26	698	14.6%	-39.00 [-41.63, -36.37]
Study D	462	32	150	450	42	148	14.4%	12.00 [3.51, 20.49]
Study E	453	26	155	420	32	150	14.5%	33.00 [26.44, 39.56]
Study F	510	38	550	540	33	552	14.6%	-30.00 [-34.20, -25.80]
Study G	526	62	85	620	82	87	13.0%	-94.00 [-115.69, -72.31]
Total (95% CI)			**2268**			**2274**	**100.0%**	**-24.55 [-47.49, -1.61]**

Heterogeneity: Tau² = 933.83; Chi² = 854.42, df = 6 (P < 0.00001); I² = 99%
Test for overall effect: Z = 2.10 (P = 0.04)

Fig. 4 Forest plot for a continuous outcome using (REM)* (REM: Random-effects model) *Note* IV: Inverse variance, df: Degree of freedom

The *studies included* in the meta-analysis are arranged *on the left side* of the forest plot in alphabetical or chronological order or by the weight assigned to them. The *type of model* used for meta-analysis is mentioned on the *top right-hand corne*r of the forest plot, just under the *type of outcomes* such as the risk ratio (RR), odds ratio (OR), or the mean difference (MD).

Important to note is the *"line of no effect"* in the *middle of the forest plot*. The whisker plots of included studies are spread on either side of this line. Any study whose CI touches this line (=1 on the x-axis) is where the intervention is interpreted as having no significant effect. This could also be interpreted as 'no significant difference' (ratio = 1) in the outcome in participants who received the intervention under study vs those in the control group.

Noting on which side of the '*line of no effect'* the results favour the intervention vs control, and considering the expected effect of the intervention (Benefits vs Reduced harm) is important for correct interpretation of the forest plot. Depending on the type of outcome (Categorical vs continuous), the forest plot will display the *number of events* (numerator) and the *number analysed* (denominator) or the *mean* and *standard deviation* for each study included in the meta-analysis.

At the bottom of the left side where the description of included studies ends, the forest plot for categorical outcomes will display the *total number of events and the total denominator* in the intervention vs control group. For continuous outcomes, it will display the *total denominator* in the intervention vs control group. Opposite this data will be the *'pooled effect estimate"* (e.g. RR and 95% CI for categorical outcomes or the MD and 95% CI for continuous outcomes) represented by a *'Diamond'* as the net output of the meta-analysis. As mentioned earlier, the size of the diamond reflects the pooled effect size and its boundaries represent its 95% CI [51]. If the boundaries of the diamond touch the line of no effect, overall, the intervention is interpreted as having no significant effect on the condition under study. The *"weight"* given to each of the included studies will differ depending on whether a fixed effect model (FEM) or random effects model (REM) was used.

At the *bottom of the left side*, *under the pooled effect estimates* will be the data on *heterogeneity (i.e. Tau^2, Chi^2, degree of freedom, I^2 value)* and the *test for overall effect* (i.e. Z). Figures 1 and 2 display the forest plots for categorical (RR) and continuous outcomes (MD) using the FEM, respectively. Figures 3 and 4 display the forest plots for categorical (RR) and continuous outcomes (MD) respectively using the REM.

It is important not to confuse between the p-value for the test of heterogeneity and that for the test for overall effect (pooled effect estimate).

FEM Vs REM for Meta-Analysis

The methods for meta-analysis, and importantly, the assumptions are different in FEM vs REM. The FEM uses the Mantel-Haenszel, Peto or inverse variance method for meta-analysis, whereas the REM uses the DerSimonian and Laird inverse variance method for this purpose. The Peto method can only combine ORs. The other three can combine ORs, RRs or risk differences. The inverse variance method minimises the uncertainty of the pooled effect estimate by assigning weight to each included study as the inverse of the variance of the effect estimate (i.e. one divided by the square of its standard error). Compared with smaller studies which have large standard errors, the larger studies are assigned more weight.

The FEM assumes that the intervention is equally effective across all studies, and the only reason that the effect size varies between studies is the *within-studies* error in estimating the effect size. Hence it gives a confident assumption. When assigning weights to the different studies, it ignores the information provided by the smaller studies because better information about the effect size is available from larger studies (Borenstein et al. 2010). In contrast, the REM allows for inter-study variability in effectiveness, giving a conservative assumption. Being less confident, it usually has wider CIs with due consideration to smaller studies. It is important to appreciate that the REM estimates the *'mean of a distribution of effects'* and not the 'true effect'. It does so by ensuring that small studies are not ignored while the large ones are not given undue weightage (Borenstein et al. 2010). Borenstein et al. recommend that the only criteria for model selection should be whether it fits the distribution of effect sizes (Borenstein et al. 2010). They suggest that REM is suitable for published studies, and discourage the strategy of starting with a FEM and moving to a REM due to significant heterogeneity (Borenstein et al. 2010).

It is preferable to compare the FEM and REM estimates of the treatment effect. If REM estimate appears more beneficial, treatment was more effective in smaller studies because the weight given to each study by REM is less influenced by sample size. If there is no evidence of heterogeneity between studies, the FEM and REM estimates will be identical. Knowing there would always be some heterogeneity (e.g. PICOS characteristics, baseline severity of illness, mode of delivery of the intervention) in included studies, many investigators in the field of medicine prefer the REM for meta-analysis (Higgins et al. 2003).

The Chi-squared Statistic

The Chi-squared statistics (Cochrane's Q test) is the conventional test for assessing heterogeneity in meta-analyses. It tests the null hypothesis that the true effect of the intervention is the same across studies and variations are simply due to chance (West et al. 2010). It is calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies, with k (number of studies) minus 1 degrees of freedom (df) and study weights as per the meta-analysis model (https://www.statsdirect.com/help/meta_analysis/heterogeneity.htm.

A low P-value suggests variation in effect estimates beyond chance. Considering that the test has low power with a small number or small sample sizes of included studies, a non-significant result should not be interpreted as evidence of no heterogeneity. This is the reason why, instead of the conventional 0.05, a P-value of 0.10 is used to assess statistical significance in such situations. In contrast, the test has a high power to detect a small, perhaps clinically non-significant heterogeneity, when there are many studies in a meta-analysis (Deeks et al. 2019).

The I-Squared Statistic

The I^2 is not a statistical test but an estimate of the percentage of variation in results across the included studies that is due to real differences and not simply due to chance. For example, an I^2 value >50% suggests that more than half of the heterogeneity from *between-study variance* cannot be explained by chance alone. It is important to note that I^2 does not provide information on the causes of heterogeneity. Exploring the reasons for significant heterogeneity (e.g. by meta-regression or subgroup analyses), and questioning the rationale for meta-analysis is important when the I^2 is > 50%. The importance given to I^2 depends on the magnitude and direction of effects, and the strength of evidence for heterogeneity (e.g. P-value from the Chi^2 test, or the 95% CI for I^2) (Higgins et al. 2002, 2003). It is calculated as follows:

I^2*= [(Q minus df) divided by Q]* × *100* where Q is the Chi^2 statistic and df is its degrees of freedom.

The rough interpretation of I^2 values is as follows: 0 to 40%: might not be important; 30 to 60%: may represent moderate heterogeneity; 50 to 90%: may represent substantial heterogeneity; 75 to 100%: considerable heterogeneity (Higgins et al. 2002, 2003).

What is Tau^2?

Tau squared (τ^2 or Tau^2) is an estimate of the *between-study* variance in REM meta-analysis. The square root of Tau^2 (i.e. tau) is the estimated standard deviation of underlying effects across studies (Aromataris and Munn 2020).

Assessment and Interpretation of the Forest Plot

For a detailed discussion on the assessment and interpretation of the forest plot, please refer to Chap. 10. Briefly, the process involves scrutiny of the following issues:

(1) *Number of studies, sample sizes of individual studies, and total sample size*

Inclusion of at least a few thousand participants in a randomised controlled trial (RCT) is considered essential for the results to have optimal validity and certainty for guiding clinical practice and research (Guyatt et al. 2011). Based on this assumption, the cumulative sample size of the studies included in a meta-analysis should be at least a few thousand. Considering the strengths and weakness of the study design (e.g. RCTs vs. non-RCTs) is important in judging the risk of bias affecting the pooled estimate provided by the meta-analysis.

(2) *Check the weightage given to different studies; is any study driving the results? Are there any outliers?*

(3) *Check the number of events in the intervention vs control* group

The event rates affect the ability of included studies to influence the pooled estimate of the effect under evaluation. In FEM meta-analysis, the weightage given to individual studies depends on their sample size as well as the event rates (Werre et al. 2005; Deeks et al. 2019; Xu et al. 2020). A study with a large sample size will not influence the results significantly if the event rate is low.

(4) *Assessment of heterogeneity: Overlap of CIs*

(5) *Tests for heterogeneity: Chi^2 (Q statistics) and its P-value, I^2: (%)*

Visual inspection of the forest plot to check for overlap of the CIs is useful to assess heterogeneity (Mohan and Adler 2019; Viechtbauer 2007; Coulson et al. 2010). reasons for significant heterogeneity should be explored (Higgins and Thompson 2002; Higgins et al. 2002; Melson et al. 2014; IntHout et al. 2015; Ioannidis 2008; Evangelou et al. 2007; von Hippel 2015; Rücker et al. 2008; Huedo-Medina et al. 2006; Bowden et al. 2011). The rationale for meta-analysis could be questioned if there is significant clinical heterogeneity.

(6) *Pooled effect (Z) size, P-value, and statistical vs. clinical significance*

It is vital to assess the pooled effect size as well as the certainty around it. Clinical significance is more important than statistical significance (Ranganathan et al. 2015).

(7) *RR vs. OR, absolute risk ratio (ARR) or difference (ARD) and the numbers needed to treat (NNT)*

The correct interpretation of RR and OR and the clinical significance of ARR and ARD is vital to avoid misinterpretation of results (Balasubramanian et al. 2015). NNT is the reciprocal of the ARD between treatment and control groups in an RCT. It is sensitive to PICOS characteristics and other factors that can affect the baseline risk. Consideration of the baseline risk and severity of illness is essential for optimal interpretation of NNTs (Ebrahim 2001).

(8) *Models used for meta-analysis, and concordance/discordance of results*

As discussed earlier, revisiting the key assumptions and characteristics of FEM vs REM is important when interpreting the forest plot (Nikolakopoulou et al. 2014; Borenstein et al. 2010; Schmidt et al. 2009; Sanchez-Meca and Marin-Martinez 2008; Hunter and Schmidt 2000; Jackson and Turner 2017; Shuster 2010; Stanley and Doucouliagos 2015). Discordance between FEM vs REM results indicates the need for exploring heterogeneity.

(9) *The strength of evidence for the pooled estimates*

Check the number of studies as well as their design, sample size, and risk of bias, contributing to the pooled estimate. The CI helps in assessing the precision of the estimate based on the total sample size available for assessing the outcome of interest. Other elements such as event rates, baseline severity of the underlying condition, setting, duration of follow up, and adverse effects are also important for judging the strength and external validity of an intervention.

(10) *Human errors in data extraction, entry and interpretation*

Check for errors in sample sizes, event rates (numerator and denominator) from included studies and their allocation to the intervention vs. control group. Transposition errors can have severe consequences for results and their interpretation. Standard error can be confused with standard deviation, and a 'minus' sign can be missing or confused with a hyphen. Be careful in the interpretation of RR vs OR.

In summary, a forest plot displays the effect estimates and confidence intervals of individual studies and their meta-analysis. Assuring that the studies selected for meta-analysis are reasonably similar to each other with regards to the PICOS characteristics is a critical step before generating forest plots. Explicit pre-stated eligibility and exclusion criteria are essential in this context. Skills in assessment and interpretation of forest plot are essential for critical appraisal of systematic reviews and meta-analysis.

References

Aromataris E, Munn Z (eds). JBI manual for evidence synthesis. JBI; 2020. https://synthesismanual.jbi.global, https://doi.org/10.46658/JBIMES-20-01. Accessed 30 Aug 2020.

Balasubramanian H, Ananthan A, Rao S, Patole S. Odds ratio vs risk ratio in randomised controlled trials. Postgrad Med. 2015;127:359–67.

Borenstein M, Hedgesb LV, Julian PT. Higgins JPT, Rothsteind HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Syn Methods. 2010; 1: 97–111. https://doi.org/10.1002/jrsm.12.

Borenstein M, Hedges LV, Higggins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010;1:97–111.

Bowden J, Tierney JF, Copas AJ, et al. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Med Res Methodol. 2011;11:41. https://doi.org/10.1186/1471-2288-11-41.

Coulson M, Healey M, Fidler F, Cumming G. Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. Front Psychol. 2010 Jul 2;1:26. https://doi.org/10.3389/fpsyg.2010.00026. eCollection 2010.

Deeks JJ, Higgins JPT, Altman DG (eds) Chapter 10: analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (eds) Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane; 2019. www.training.cochrane.org/handbook.

Deeks JJ, Higgins JPT, Altman DG; on behalf of the Cochrane Statistical Methods Group. Chapter 10, Section 10.10.2: identifying and measuring heterogeneity. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (eds) Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane; 2019. www.training.cochrane.org/handbook.

EBM notebook: Weighted event rates. Werre SR, Walter-Dilks C. BMJ Evidence-based Medicine, June 2005;10:70. http://dx.doi.org/10.1136/ebm.10.3.70.

Ebrahim S. The use of numbers needed to treat derived from systematic reviews and meta-analysis: caveats and pitfalls. Eval Health Prof. 2001;24:152–64.

Evangelou E, Ioanidis JPA, Patsopoulos NA. Uncertainty in heterogeneity estimates in meta-analyses. BMJ. 2007;335:914–6.

Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence and imprecision. J Clin Epidemiol. 2011; 64: 1283e–1293.

Heterogeneity in Meta-analysis (Q, I-square)—StatsDirect. https://www.statsdirect.com/help/meta_analysis/heterogeneity.htm. Accessed 30 Aug 2020.

Higgins J, Thompson S, Deeks JJ, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: A critical appraisal of guidelines and practice. J Health Service Res Policy. 2002; 7:51–61.

Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;15 (21):1539–58.

Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. J Health Serv Res Policy. 2002;7:51–61.

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003a;327:557–60.

Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003b;327(7414):557–60. https://doi.org/10.1136/bmj.327.7414.557.

Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol. 2005;5:13.

Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods. 2006;11:193–206.

Hunter JE, Schmidt FL. Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. Int J Sel Assess. 2000; 8: 275–292.

IntHout J, Ioannidis JP, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. J Clin Epidemiol. 2015;68:860–9.

Ioannidis JP. Interpretation of tests of heterogeneity and bias in meta-analysis. J Eval Clin Pract. 2008;14:951–7.

Jackson D, Turner R. Power analysis for random-effects meta-analysis. Res Synth Methods. 2017;8:290–302.

Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ. 2001;322:1479–80.

Melson WG, Bootsma MCJ, Rovers MM, Bonten MJM. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. Clin Microbiol Infec. 2014;20:123–9.

Mohan BP, Adler DG. Heterogeneity in systematic review and meta-analysis: how to read between the numbers. Gastrointest Endosc. 2019;89:902–3.

Nikolakopoulou A, Mavridis D, Salanti G. How to interpret meta-analysis models: fixed effect and random effects meta-analyses. Evid Based Mental Health. 2014;17(2):64. https://doi.org/10.1136/eb-2014-101794.

Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: Clinical versus statistical significance. Perspect Clin Res. 2015;6:169–70.

Rücker G, Schwarzer G, Carpenter JR, et al. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8:79.

Sanchez-Meca J, Marin-Martinez F. Confidence intervals for the overall effect size in random-effects meta-analysis. Psychol Methods. 2008;13:31–48.

Schmidt FL, Oh IS, Hayes TL. Fixed-versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. Br J Math Stat Psychol. 2009;62:97–128.

Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. Stat Med. 2010;30 (29):1259–65.

Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. Stat Med. 2015;15(34):2116–27.

Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. Stat Med. 2007;26:37–52.

von Hippel PT. The heterogeneity statistic I(2) can be biased in small meta-analyses. BMC Med Res Methodol. 2015;14(15):35.

Wan X, Wang W, Liu J, et al. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol. 2014;14:135.

West SL, Gartlehner G, Mansfield AJ, et al. Comparative effectiveness review methods: clinical heterogeneity [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2010 Sep (Table 7, summary of common statistical approaches to test for heterogeneity). https://www.ncbi.nlm.nih.gov/books/NBK53317/table/ch3.t2/.

Xu C, Li L, Lin L, et al. Exclusion of studies with no events in both arms in meta-analysis impacted the conclusions. J Clin Epidemiol. 2020; 123: 91–99.

Sensitivity and Subgroup Analyses

Mangesh Deshmukh

Abstract Systematic reviews involve a sequence of decisions and assumptions ranging from the definition of a particular variable and use of statistical methods to the type of model chosen for meta-analysis. If incorrect, these decisions and assumptions can influence the conclusions of the systematic review. Sensitivity and subgroup analyses play an important role in addressing these issues in meta-analysis. Sensitivity analysis helps in checking the sensitivity of the overall conclusions to various limitations of the data, assumptions, and approach to analysis. Consistency between the results of primary analysis and sensitivity analysis strengthens the conclusions and credibility of the findings. Effects of an intervention may not be homogeneous across all participants in a clinical trial. They may vary based on participant characteristics such as age, gender, and severity of illness. Subgroup analyses help in identifying subgroups of participants with most benefits (or adverse effects) of the intervention compared with others. This chapter covers the principles, practice, and pitfalls, of sensitivity and subgroup analyses in systematic reviews and meta-analysis.

Keywords Sensitivity · Subgroup · Analysis · Heterogeneity · Systematic review · Meta-analysis

Introduction

Sensitivity and subgroup analyses play an important role in systematic reviews and meta-analysis. Understanding the principles, practice and pitfalls of sensitivity and subgroup analyses in systematic reviews and meta-analysis is hence necessary.

M. Deshmukh (✉)
Department of Neonatology, Fiona Stanley Hospital, School of Medicine,
Curtin and University of Western Australia, Perth, WA 6150, Australia
e-mail: mangesh.deshmukh@health.wa.gov.au

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_9

Sensitivity Analysis

Systematic reviews involve a sequence of assumptions and decisions during the process of deriving the primary results. Many of these assumptions and decisions are objective and non-controversial, but some could be subjective and debatable (Jonathan et al. 2020). The assumptions or decisions can range from the definition of a particular variable and use of statistical methods to the type of model chosen for meta-analysis. It is essential to check if the results of the systematic review are not dependent on subjective, arbitrary, unclear, or changed assumptions or decisions.

Sensitivity analysis is a repetition of the primary analysis with an altered dataset or statistical method to check if altering any of the assumptions changes the pooled effect estimate, and hence the final conclusions (Viel et al. 1995). To put it simply, *sensitivity analysis checks the sensitivity of the overall findings to limitations of the data, assumptions, and the approach to analysis*.

The results of a systematic review are considered robust if they remain unchanged despite changes in the assumptions. On the other hand, their robustness is questionable if variations in assumptions significantly change the findings. Consistency between the results of primary analysis and sensitivity analysis enhances the credibility of the findings (Thabane et al. 2013)

Sensitivity analysis is a widely used technique to support decision-makers. The US Food and Drug Administration (FDA), European Medicines Association (EMEA), National Institute of Health and Clinical Excellence (NICE) recommend its use for checking the robustness of results, exploring alternative scenarios and assessing the uncertainty in cost-effectiveness (Thabane et al. 2013)

Scope for Sensitivity Analysis in a Systematic Review and Meta-analysis

Ideally, all meta-analyses should include a sensitivity analysis. It can be pre-specified, but often many issues suitable for sensitivity analysis are identified only during the review process. Decisions are generally based on the quality of the included studies, heterogeneity, and publication bias. Results of a survey of major medical and health economics journals by Thabane et al. showed that the point prevalent use of sensitivity analysis was very low at 26.7% (36/135) (Thabane et al. 2013). Compared with medical journals, a higher percentage of publications in health economics journals (20.3 vs 30.8%) reported some sensitivity analysis. Assessing the robustness of findings to different methods of analysis was the most common type of sensitivity analysis (Thabane et al. 2013).

Conducting Sensitivity Analysis

Sensitivity analysis can be conducted at different stages of a systematic review as follows:

1. **Literature search**

Sensitivity analysis can be conducted at literature search level, based on the decision to include studies published only as an abstract. The primary analysis can be conducted using the data from the abstracts. However, a sensitivity analysis can be performed to check the robustness of the results by excluding the studies published only as abstracts.

2. **Study design**

Sensitivity analysis can be conducted by various approaches at this stage. In the case of randomised controlled trials (RCTs), sensitivity analysis can be performed based on the risk of bias (ROB) of included studies. The ROB is typically assessed in the domains of random sequence generation, allocation concealment, blinding of participants and investigators, incomplete outcome data, selective reporting and other biases (Julian et al. 2020). Sensitivity analysis can be conducted by performing reanalysis at every domain. In the case of non-RCTs, sensitivity analysis can be conducted by excluding studies that have not used regression analysis to adjust for confounders. In cluster-RCTs, the intra-class correlation coefficient values can be used when analyses are not adjusted for clustering (Jonathan et al. 2020). In cross-over trials, the within-subject correlation coefficient values can be used when this is not available in primary reports (Jonathan et al. 2020).

3. **Eligibility criteria**

Sensitivity analysis can be conducted based on the characteristics of participants (e.g. age, gender), intervention (e.g. dose, duration, and route), control (e.g. standard treatment, placebo), outcomes (e.g. primary or secondary), and time (e.g. old vs. new studies).

4. **Type of data**

Sensitivity analysis can be conducted based on assumptions of the distribution of censored data in the case of time-to-event studies. In the case of continuous data, it can be based on whether standard deviations are missing, and when and how should they be imputed (Jonathan et al. 2020). It is also important to consider whether sensitivity analyses should be based on a change in scores from baseline or on final scores. In case of ordinal scales, such analyses can be conducted depending on the cut-off points used to dichotomise short ordinal scales into two groups (Jonathan et al. 2020).

5. **Type and method of analysis**

One of the common methods to conduct sensitivity analysis depends on the approach to analysis in the included studies (e.g. intention to treat vs. per-protocol analysis). Sensitivity analysis can be based on the model used for meta-analysis (e.g. fixed-effect vs. random-effects model). It can be based on the parameters for reporting the effect estimates such as the odds ratio (OR), risk ratio (RR) or risk difference (RD) in case of dichotomous outcomes. In the case of continuous outcomes, it can be done based on the standardised mean difference (SMD) across all scales or as the mean differences (MD) individually for each scale.

6. **Other issues**

Sensitivity analysis can be conducted based on the variation in inclusion criteria involving a numerical value (e.g. age at enrolment). The choice of value could be arbitrary (e.g. defining old age as >60, >70 or >75 years). They can also be conducted if an included study does not provide or did not obtain the required data (e.g. loss to follow-up).

Issues After Sensitivity Analysis

Specific assumptions or missing data may significantly influence the results of the systematic review. Reviewers are expected to address this issue by contacting authors for additional and/or individual patient data. Results must be interpreted with caution if this is not possible. Such findings may generate proposals for further investigations and research.

Reporting of Sensitivity Analysis

A sensitivity analysis is generally reported in a summary table. Individual forest plots for each sensitivity analysis are usually unnecessary.

Example of a Sensitivity Analysis

Chou et al. evaluated the efficacy and safety of statins for prevention of cardiovascular disease in adults. (Chou et al. 2016) A meta-analysis of all ten studies showed that statins significantly reduced cardiovascular mortality (RR: 0.69(95% CI: 0.54 to 0.88). The authors noted that two of the large trials (JUPITER and ASCOT-LLA) were terminated prematurely at two years of follow up as per the recommendations of the data and safety monitoring committee based on the significantly lower mortality in the statin group. It is well known that trials terminated prematurely can result in erroneous conclusions. The authors hence conducted a sensitivity analysis by excluding those two trials. The results remain significant (RR: 0.70 (95% CI: 0.52 to 0.94) after excluding the JUPITER and ASCOT-LLA trials. The authors also observed that 5/10 included trials were of low quality (i.e. carried a high ROB). Hence, they conducted a sensitivity analysis by excluding those five trials. The results remained significant after excluding five studies with high ROB; (RR: 0.64(95% CI: 0.44 to 0.93). Three of the studies had included patients with pre-existing cardiovascular disease. Sensitivity analysis excluding these studies showed no difference from the primary analysis (RR: 0.62 (95% CI: 0.46 to 0.85). Four studies had a relatively short duration of follow up for less than three years. Results of the sensitivity analysis performed by excluding these four studies remained robust (RR: 0.71 (95% CI: 0.53 to 0.96). Overall, these sensitivity analyses showed that the primary results were robust (Chou et al. 2016).

Subgroup Analysis

Significant resources are involved in conducting clinical trials to assess if an intervention works. However, it is essential to appreciate that the intervention may not work homogeneously across all participants in a trial. Effects of the intervention under study may vary based on patient characteristics such as age, gender, geographic location, the severity of illness, and comorbidities. Hence, there may be subgroups of participants with greater benefits following the intervention. Subgroup analyses are helpful in identifying the subgroup of participants with more beneficial effects (or adverse effects) related to the intervention (Tanniou et al. 2016).

What Is Involved In Subgroup Analyses?

Subgroup analysis involves splitting data from all participants into groups and analysing it separately to make comparisons to obtain further information regarding the efficacy of the intervention under study (Jonathan et al. 2020). Similar to sensitivity analysis, a subgroup analysis attempts to pick out lost information. Conducting a subgroup analysis is vital if a specific group of participants in the study is expected to respond differently to an intervention based on biologic plausibility.

Subgroup analysis can be conducted based on the study design (RCT, Non-RCT), subsets of participants (e.g. males and females), and intervention (e.g. mode, route, type, dose), or subsets of studies (e.g. different geographical locations). It helps in investigating heterogeneous results or answering specific questions about particular participant groups, and the intervention or study type (Jonathan et al. 2020).

Subgroup analysis involves reanalysis after dividing the entire dataset according to some factor/characteristic of interest. *This is a different approach compared with sensitivity analysis, which requires reanalysis after adding or removing studies.* It is important to note that subgroup analysis is not for searching a group showing benefit in the study population when the overall results are negative. Conducting a subgroup analysis under such circumstances is inappropriate.

What are the Types of Subgroup Analyses?

Subgroups can be pre-planned or post hoc. Pre-planned subgroups, as the name suggests, are planned *a priory* during the design phase of the study. They generally include analysis based on biological plausibility or, as mentioned before, some relevant factor such as gender, age, and comorbidities. On the contrary, a post hoc analysis is undertaken after the results of the study are known. Pre-planned subgroup analysis is considered superior to post hoc analysis if it is based on stratified randomisation.

What is Data Mining and Data Dredging in Subgroup Analyses?

Data mining and data dredging are two specific terms used in the context of subgroup analysis. Data mining implies appropriate use of subgroup analysis where the study has achieved the primary outcome, and the data is further scrutinised to find the subset of study population where the intervention works the best (Martin 1984). Data dredging means trying to stretch the data by inappropriate use of subgroup analysis to find the subgroup which showed beneficial effects when the study had not achieved the primary endpoint. To put it simply, this means creating many subgroups until the significant effect is detected. Data dredging is associated with a higher risk of getting false-positive results (Jonathan et al. 2020; Martin 1984; Gebski and Keech 2003; Lee 2004)

Subgroup Analysis in Clinical Trials and Meta-analysis

Subgroup analysis has become an integral part of the analysis of clinical trials. It attempts to provide more information from the available dataset of the trial. As discussed above, pre-planned and post hoc subgroup analyses can be conducted using the entire dataset of a trial based on baseline characteristics of included participants or the outcome. Outcome-based subgroup analyses frequently involve specific outcomes such as the severity as against the occurrence of the condition and quality of life as against survival. Subgroup analyses based on participant characteristics are generally prefixed and carry more weightage than the outcome-based subgroup analysis, which are usually post hoc (Hirji 2009).

Strategies for Optimising Subgroups

Experts suggest that (1) subgroups should be based on participant characteristics (e.g. young vs old, male vs female, the severity of illness) which may be associated with different response to the intervention, (2) the level of significance should be stringent, with $p < 0.01$ or <0.005, (3) post hoc subgroups should be supported by biological plausibility, and prior evidence, (4) and methodology for analysis (e.g. test of interaction) should be reported in detail (Jonathan et al. 2020; Gebski and Keech 2003; Dijkman et al. 2009; Lagakos 2006).

Interpreting the Results of Subgroup Analyses

Results of subgroup analyses should be interpreted with caution given the risk of bias because the subgroup population is different from the one that was initially randomised. Furthermore, reduction in sample size is an inherent major drawback of subgroup analyses. The power to detect clinically meaningful effects, especially when the study has a moderate or small sample size to start with, is reduced considerably (Fagerland 2009). The probability of obtaining positive results by chance alone is high when too much emphasis is put on subgroup analysis (Lagakos 2006) At best, the results of subgroup analyses can only help generate hypothesis for future clinical trials when the original study has not reached the primary endpoint.

Examples of a Subgroup Analysis

(1) Dermyshi et al. conducted a meta-analysis of RCTs evaluating the effects of probiotic supplementation in preterm neonates (gestation < 34 weeks and birth weight < 1500 g). (Dermyshi et al. 2017) Pooled estimate from 30 RCTs showed a significant reduction in necrotising enterocolitis (NEC Stage $\geq$ II) in the probiotic supplemented group. However, subgroup analysis in ELBW (extremely low birth weight: Birth weight $\leq$ 1000gms) neonates showed no such benefit (Table 1). The pre-stated subgroup was justified considering that the incidence, mortality and morbidity related to NEC Stage $\geq$ II are significantly higher in ELBW neonates

Table 1 Probiotics for preventing necrotising enterocolitis in preterm neonates*

	Probiotics	Placebo	RR (95% CI)	P-value
Preterm neonates: Gestation < 34 weeks and BW < 1500gms	146/4304	253/4231	0.57, (0.47, 0.70)	<0.00001
ELBW neonates (BW < 1000 g)	62/599	69/617	0.93 (0.67,1.27)	0.64

*Dermyshi et al. (2017)

RR: Relative risk, CI: Confidence interval, BW: Birth weight, ELBW: Extremely low birth weight

Table 2 Meta-analysis of randomised trials assessing the effect of Aspirin on pre-eclampsia*

	Aspirin	Placebo	RR (95% CI)	P-value
Commenced at ≤ 16 weeks' gestation	221/2564	354/2549	0.57, (0.43, 0.75)	<0.001
Commenced at > 16 weeks' gestation	517/7701	586/7669	0.81 (0.66,0.99)	0.04

*Roberge et al. (2017)

RR: Relative risk, CI: Confidence interval

(2) Roberge et al. conducted a systematic review of RCT's evaluating the efficacy of Aspirin to reduce pre-eclampsia in women at risk (Roberge et al. 2017). Based on the biological plausibility (i.e. placentation occurs within first 16 weeks of gestation and disorders of placentation lead to pre-eclampsia, a subgroup analysis was conducted based on the timing (before and after 16 weeks gestation) of commencement of Aspirin. Results of the subgroup analysis suggested that Aspirin was more beneficial if commenced early (Table 2).

Sensitivity Versus Subgroup Analyses

Sensitivity and subgroup analyses involve exploring the nuances of the dataset. Understanding the key differences between the two is essential (Table 3). The former does not estimate the effect in the group of studies removed from the analysis, whereas the latter estimates the effect in the subgroup of interest. Sensitivity analyses compare the effect of different methods of estimation on the same outcome of interest. In contrast, subgroup analyses involve comparisons across the subgroups. The limitation common to both analyses is the reduction in sample size (Jonathan et al. 2020).

Table 3 Differences between Sensitivity and subgroup analysis

Sensitivity analysis	Subgroup analysis
Can be pre-specified or post hoc	Usually pre-specified
Results of excluded studies are not reported	Results of each sub-group are reported
Forest plots are not necessary	Forest plots are useful

References

Chou R, Dana T, Blazina I, Daeges M, Jeanne TL. Statins for Prevention of Cardiovascular Disease in Adults: Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA. 2016 Nov 15;316(19):2008–24. PMID: 27838722. https://doi.org/10.1001/jama.2015.15629.

Dermyshi E, Wang Y, Yan C, Hong W, Qiu G, Gong X, et al. The "Golden Age" of probiotics: a systematic review and meta-analysis of randomised and observational studies in preterm infants. Neonatology. 2017;112(1):9–23. PMID: 28196365. https://doi.org/10.1159/000454668.

Dijkman B, Kooistra B, Bhandari M. How to work with a subgroup analysis. Can J Surg. 2009;52 (6):515–22 PMID: 20011190.

Hirji KF, Fagerland MW. Outcome based subgroup analysis: a neglected concern. Trials. 2009 May 20;10:33. PMID: 19454041. https://doi.org/10.1186/1745-6215-10-33.

Gebski VJ, Keech AC. Statistical methods in clinical trials. Med J Aust. 2003;178(4):182–4 PMID: 12580749.

Jonathan J, Deeks JPH, Douglas GA. Sensitivity analyses. In: Julian Higgins JT, editor. Cochrane handbook for systematic reviews of interventions version 61; 2020.

Jonathan J Deeks JPH, Douglas GA. What are subgroup analyses?. In: Julian Higgins JT, editor. Cochrane handbook for systematic reviews of interventions version 50; 2020.

Julian PT, Higgins JS, Matthew JP, Elbers RG, Sterne JAC. Assessing risk of bias in a randomised trial. In: Higgins JT, editor. Cochrane handbook for systematic reviews of interventions version 61;2020.

Lagakos SW. The challenge of subgroup analyses–reporting without distorting. N Engl J Med. 2006 Apr 20;354(16):1667–9. PMID: 16625007. https://doi.org/10.1056/nejmp068070.

Lee SK. On Classification and regression trees for multiple responses. In: Banks DMF, Arabie P, Gaul W, editors. Classification, clustering, and data mining applications. Berlin: Springer; 2004.

Martin G. Munchausen's statistical grid, which makes all trials significant. Lancet (London, England). 1984;2(8417–8418):1457.

Roberge S, Nicolaides K, Demers S, Hyett J, Chaillet N, Bujold E. The role of aspirin dose on the prevention of pre-eclampsia and fetal growth restriction: systematic review and meta-analysis. Am J Obstet Gynecol. 2017 Feb;216(2):110–20.e6. PMID: 27640943. https://doi.org/10.1016/j.ajog.2016.09.076.

Tanniou J, van der Tweel I, Teerenstra S, Roes KC. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. BMC Med Res Methodol. 2016 Feb 18;16:20. PMID: 26891992. https://doi.org/10.1186/s12874-016-0122-6.

Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol. 2013 Jul 16;13:92. PMID: 23855337. https://doi.org/10.1186/1471-2288-13-92.

Viel JF, Pobel D, Carré A. Incidence of leukaemia in young people around the La Hague nuclear waste reprocessing plant: a sensitivity analysis. Stat Med. 1995 15–30;14(21–22):2459–72. PMID: 8711281. https://doi.org/10.1002/sim.4780142114.

Rating Certainty of the Evidence Using GRADE Guidelines

Abhijeet Rakshasbhuvankar

Abstract Systematic reviews in healthcare should review and synthesise all available evidence, and provide information regarding certainty (quality) of evidence to inform readers about the amount of confidence they can place in the evidence. Many international organisations, such as the World Health Organisation, National Institute for Health and Care Excellence (NICE) and the Cochrane Collaboration have recommended GRADE (The Grading of Recommendations Assessment, Development, and Evaluation) guidelines to rate the certainty of (a body of) evidence in systematic reviews. These guidelines provide a structured and transparent process to rate the certainty of evidence considering critical factors which may decrease (risk of bias, inconsistency, indirectness, imprecision, and reporting bias) or increase (a very large effect, dose-response relation, and bias that would decrease effect estimate) our confidence in effect estimates. The process of rating certainty of the evidence is presented as a Summary of Findings table in a systematic review. This chapter covers the use of GRADE guidelines for rating certainty of evidence in a systematic review.

Keywords Certainty · Evidence · GRADE · Imprecision · Inconsistency · Indirectness · Publication bias · Quality · Reporting bias

Introduction

Systematic reviews aim to synthesise the available evidence to help clinicians, guideline developers, and researchers make evidence-based decisions, develop clinical care guidelines, and identify the gaps in knowledge, respectively. The synthesis should include not only the effect estimates but also the level of confidence in them. The level of confidence in the effect estimate decides its usefulness and is determined by the certainty (quality) of evidence (Guyatt et al. 2008). The

A. Rakshasbhuvankar (✉)
School of Medicine, University of Western Australia, Perth, WA 6008, Australia
e-mail: Abhijeet.rakshasbhuvankar@health.wa.gov.au

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_10

certainty of the evidence is defined as the extent to which one can be confident that an estimate of effect is correct (Atkins et al. 2004). Various systems have been used to grade the certainty of evidence.

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) guidelines were developed by the GRADE Working Group and are recommended by the Cochrane collaboration to rate the certainty of evidence (Puhan et al. 2014; Schunemann et al. 2019). The GRADE system has an advantage over other systems. It explicitly considers multiple vital components that determine the evidence's certainty, provides a structured and explicit approach for reviews to make their judgments, and enables readers to understand the reasoning behind the decisions.

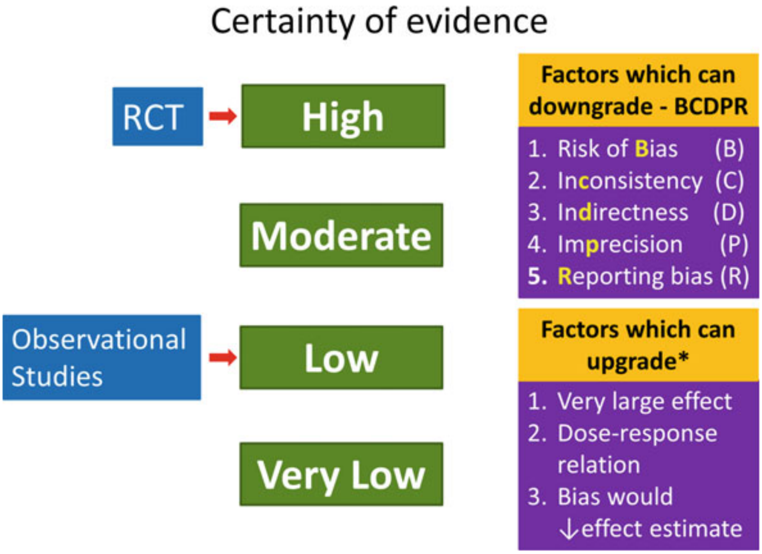
In addition to rating the certainty of the evidence, GRADE guidelines are also used for rating strength of recommendations. The strength of recommendation indicates the extent to which one can be confident that adherence to the recommendation will do more good than harm (Atkins et al. 2004). The judgment regarding the strength of recommendation in addition to the certainty of the evidence requires careful consideration of the balance between beneficial versus harmful effects, baseline risk, and available resources. This chapter covers the approach to the judgment about the certainty of evidence using the GRADE system.

GRADE Levels of Evidence

GRADE classifies certainty of evidence in four levels: high, moderate, low, and very low (Puhan et al. 2014). The level of confidence progressively decreases as we move stepwise from "high" to "very low" category (Fig. 1) (Guyatt et al. 2008). In general, the certainty of evidence generated from randomised controlled trials (RCTs) is considered as "High," and that from observational studies is regarded as "low." However, significant concerns regarding any of the following factors may downgrade certainty of evidence: risk of bias (ROB), inconsistency, indirectness, imprecision, and publication bias. The certainty of evidence may be upgraded, although rarely, in observational studies in the presence of large effect size, dose-response gradient, or plausible confounders or biases that increase the confidence in the estimated effect (Guyatt et al. 2008; Balshem et al. 2011). The details regarding the factors which can downgrade or upgrade certainty of evidence in a systematic review are described below.

Risk of Bias (ROB)

Bias is a systematic error in results and arises from methodological flaws in a study (Higgins et al. 2019). The reliability of RCT results depends on the extent to which potential sources of biases have been avoided. Bias may arise from the



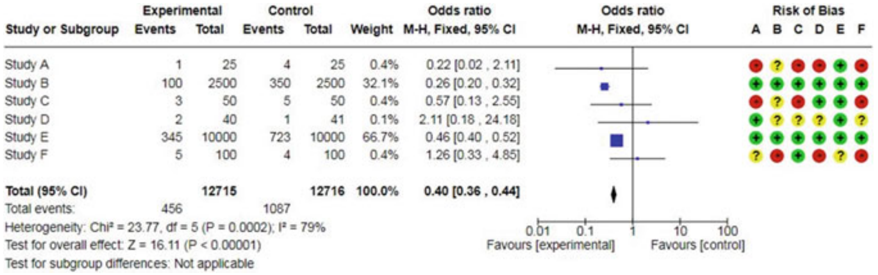
*Limited mainly to observational studies

Fig. 1 Levels of certainty of evidence and the factors which downgrade and upgrade it

randomisation process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Risk of bias (ROB) assessment is an integral part of the systematic review methodology. GRADE requires the systematic reviewers to decide the (overall) ROB for each outcome across all studies and all domains. The judgment demands careful consideration of ROB in the individual studies for the outcome under consideration and the extent to which the study contributes to the effect estimate (weightage).

(1) *ROB assessment*: Each outcome under consideration is assessed for five sources (domains) of ROB. The risk in each domain is judged as Low risk, Some concerns, or High risk. The details of evaluating an individual study for the ROB are provided elsewhere in this book.

(2) *Contribution (weightage) of the study to the effect estimate*: The contribution of a study for the ROB in a systematic review is proportional to the contribution the study makes for the effect estimate. For example, Fig. 2 shows a forest plot and ROB for a hypothetical systematic review of drug A for pancreatic cancer for the outcome of five-year survival. Studies A, C, and F have high ROB from multiple sources; however, the forest plot indicates that the studies contribute to a negligible extent to the pooled effect estimate. In contrast, Studies B and E, which add the most to the pooled estimate, have low ROB. Hence the reviewers may judge ROB for drug A in pancreatic cancer for the outcome of five-year survival as “Low”.



	Experimental		Control			Odds ratio	Risk of Bias					
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Fixed, 95% CI	A	B	C	D	E	F
Study A	1	25	4	25	0.4%	0.22 [0.02 , 2.11]	-	?	-	-	+	-
Study B	100	2500	350	2500	32.1%	0.26 [0.20 , 0.32]	+	+	+	+	+	+
Study C	3	50	5	50	0.4%	0.57 [0.13 , 2.55]	-	?	-	+	+	-
Study D	2	40	1	41	0.1%	2.11 [0.18 , 24.18]	+	?	?	?	+	?
Study E	345	10000	723	10000	66.7%	0.46 [0.40 , 0.52]	+	+	+	+	+	+
Study F	5	100	4	100	0.4%	1.26 [0.33 , 4.85]	?	-	+	-	?	-
Total (95% CI)		**12715**		**12716**	**100.0%**	**0.40 [0.36 , 0.44]**						
Total events:	456		1087									

Heterogeneity: Chi² = 23.77, df = 5 (P = 0.0002); I² = 79%
Test for overall effect: Z = 16.11 (P < 0.00001)
Test for subgroup differences: Not applicable

Fig. 2 Hypothetical forest plot and risk of bias—judgement regarding overall risk of bias

Suggestions for downgrading for ROB: (1) If most information is from studies at low ROB: Do not downgrade; (2) If most information is from studies with some concerns: Downgrade by one level, (3) If most information is from studies at high ROB: Downgrade by one or two levels based on the seriousness of limitations.

Reviewers need to apply judgement while deciding overall ROB. In close-call situations, reviewers should be conservative in the decisions of rating down the evidence, should consider ROB judgement in the context of other limitations, and make explicit statements regarding the reasoning behind their judgement (Guyatt et al. 2011).

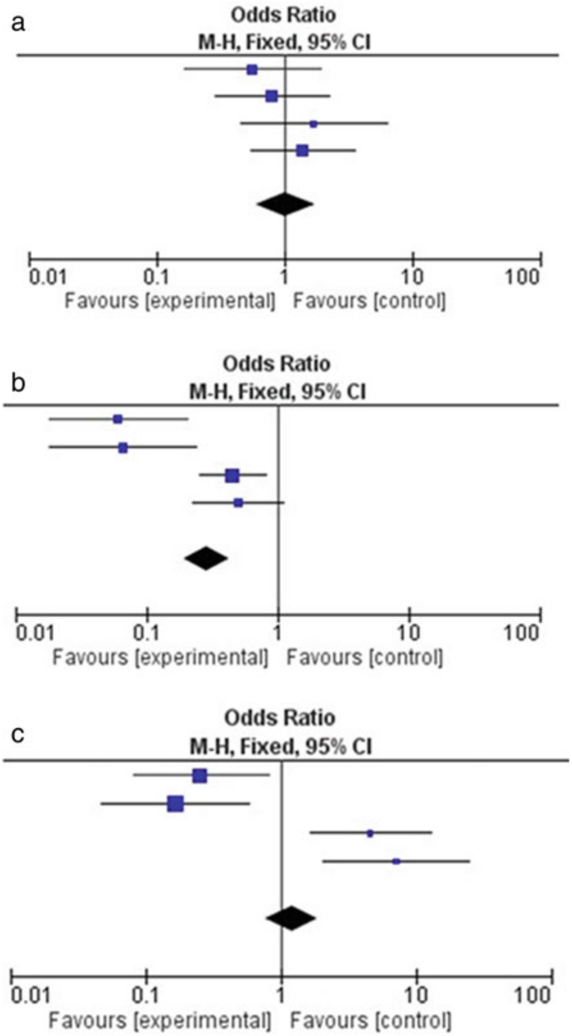
Inconsistency (Heterogeneity)

Consistency in a systematic review refers to the similarity in the magnitude of effect estimates of the studies. The study results are inconsistent when the variations in the effect estimates between the studies cannot be explained based on chance alone. Inconsistency which cannot be explained by a priori hypotheses may decrease our confidence in the results. Inconsistency is important only when it reduces our confidence in the effect estimates. Assessment of inconsistency of effects across the studies is an integral part of a meta-analysis and grading of evidence (Higgins et al. 2003).

Judgement regarding inconsistency is based on visual inspection of forest plot and statistical tests (Guyatt et al. 2011).

(1) *Forest plot*. The direction of effect and overlap of confidence intervals between the trials are two critical factors in the forest plot, which help in the judgement regarding inconsistency. The impact of these two factors on the judgment of inconsistency is explained with the help of hypothetical forest plots in Fig. 3. In the forest plot A, the directions of effects in the first two studies are different from those in the second two studies. However, the magnitude of the difference is small, and the confidence intervals of the trials overlap. Therefore, the forest plot does not show inconsistency, and our confidence in the pooled effect

Fig. 3 Hypothetical forest plots—judgement regarding inconsistency



estimate remains intact; hence, we should not downgrade for inconsistency. In the forest plot B, all the four trials have the same direction of effect; however, the magnitudes of effect estimates vary, and there is little overlap between the confidence intervals between first and second two studies. Therefore, the forest plot shows inconsistency. However, the inconsistency probably does not decrease our confidence in the pooled estimate. Hence, we may not downgrade for inconsistency. In the forest plot C, the magnitude of difference in the effect estimates between the first two and second two studies is similar to that in the forest plot B, but the direction of effects are opposite. The first two studies favour intervention while the latter two studies favour control. Therefore, the forest plot shows inconsistency. Does the inconsistency decrease the confidence in the pooled estimate? Probably yes, and hence, we should downgrade certainty of the evidence for inconsistency.

(2) *Statistical tests*. The two commonly used statistical tests for inconsistency (heterogeneity) are the Chi-squared test (test for heterogeneity) and the I^2 test. The Chi-squared test examines the null hypothesis that all studies evaluate the same effect. A p-value of < 0.05 for Chi-squared test indicates heterogeneity. I^2 test quantifies heterogeneity and can be used to compare heterogeneity across meta-analyses of different sizes, of different types of studies, and different types of outcome data (Higgins et al. 2003). I^2 value of < 40, 30–60, 50–90, 75–100% indicate low, moderate, substantial, and considerable heterogeneity respectively. The disadvantage of the I^2 test is that the cut-off values are not established, and judgement is required when the values fall in the overlapping zone. Chi-squared test and I^2 values are calculated in RevMan 5.4 software, and the values are displayed at the bottom of the forest plot (Review Manager. Version 5.4. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2020).

Suggestions for downgrading: The judgement regarding inconsistency requires careful evaluation of the forest plot and statistical tests. Downgrade for inconsistency only if it decreases our confidence in the pooled effect estimate.

Indirectness

Direct evidence comes from research that directly compares the interventions in which we are interested when applied to the populations in which we are interested, and measures outcomes important to patients (Guyatt et al. 2011). Indirectness refers to the extent to which the people, interventions, and outcome measures are different from those of interest. The fourth cause of indirectness results when there is no direct comparison between the two interventions of interest.

(1) ***Indirectness resulting from differences in the population of interest***: Systematic reviews will include only those studies which fulfil criteria with regards to population. However, indirectness can still result in some situations. For example, systematic review plans to investigate the effect of drug A in a patient population of individuals > 60 years. After performing a literature search, the reviewers notice that many studies examining drug A had 70 years or more eligibility criteria. In this case, the studies recruiting patients exclusively above 70 years of age still satisfy the inclusion criteria for the systematic review. Still, the age criteria of the included studies and the systematic review are not identical. Therefore, the effect of indirectness must be considered when concluding such situations. In this example, the reviewers may consider downgrading the level of evidence by one level if (a) there is a physiological

basis to assume that the effect of drug A in population >60 years is likely to be significantly different from the effect in a population exclusively >70 years of age, and (b) the studies with population exclusively >70 years' of age contribute a significant amount (weightage) of information to the pooled effect estimate.

(2) ***Indirectness resulting from differences in intervention***: Indirectness results when reviewers want to compare drug A to drug B; however, there is no direct comparison of drug A to drug B. Instead, the studies have compared drug A to drug C, and drug C to drug B. This type of indirectness is uncommon in systematic reviews. A more common reason for indirectness maybe when the studies have used only a part of rather than whole intervention. For example, a systematic review aims to investigate the effect of a group of interventions A-B-C-D for expediting post-operative recovery. The reviewers find that many studies have used only interventions A-C-D. The reviewers must consider the effect of indirectness if they include the studies with intervention A-C-D in the systematic review. The decision regarding downgrading certainty of evidence depends on whether the difference in the interventions (A-B-C-D versus A-C-D) is likely to have a significant effect on the outcome of interest (post-operative recovery) and amount of information (weightage) contributed by the studies with A-C-D intervention.

(3) ***Indirectness resulting from differences in the outcome***: This is a common reason for indirectness in systematic reviews. It may result for two reasons:

   (i) *Differences in the time frame*: For example, if the reviewers are interested in the intervention effect at 12 months but include studies that have reported effect only at six months. Suppose there is evidence that for other similar interventions, the effect decreases significantly from 6 months to 12 months, and a significant amount of information comes from the studies which have reported effect only until six months. In that case, the reviewers may decrease the level of certainty for indirectness.

   (ii) *Use of surrogate outcome*: Indirectness results when studies report only surrogate markers of the clinically meaningful outcomes; for example, HbA1c for symptoms of diabetes, C-reactive protein for sepsis. In such scenarios, reviewers should consider indirectness resulting from the difference in the outcome while grading the level of evidence

(4) ***Indirectness when there is no direct comparison between two interventions of interest***: This type of indirectness results when reviewers want to compare intervention A versus intervention B; however, the studies have compared intervention A versus intervention C and Intervention B versus intervention C. The indirect comparison requires assumption to be made that the population characteristics, co-interventions, outcome measurement, and the methodological qualities are not significantly different between the studies to result in different effects (Song et al. 2009). Because this assumption is always in some doubt, indirect comparisons always warrant rating down the quality of evidence by one level.

Suggestions for downgrading: The reviewers should consider rating down the certainty of evidence if indirectness is likely to influence the outcome of interest, and the significant amount of information comes from the studies with indirectness. Reviewers may rate down by one level when indirectness comes from a single factor, and by two if it comes from multiple factors. The decision requires judgement and consideration of the overall impact of the indirectness on the effect estimate.

Imprecision

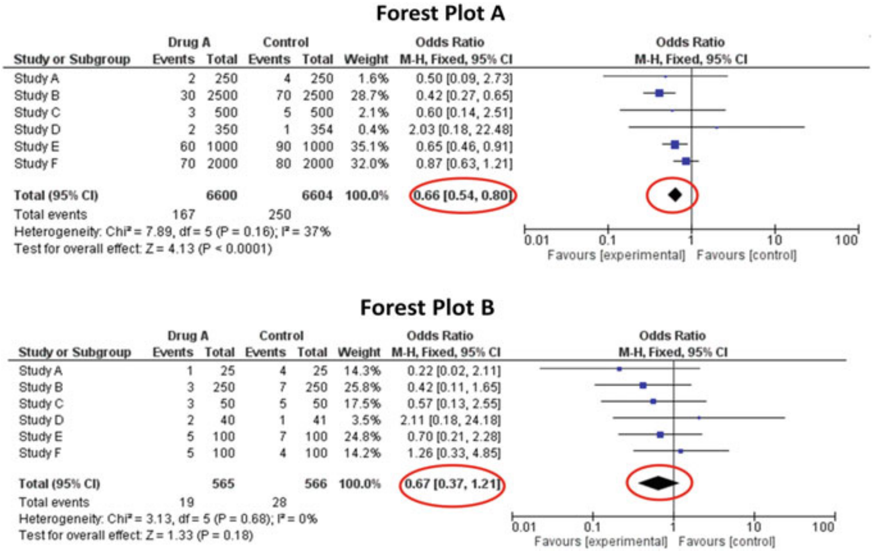
Precision refers to the degree of agreement between repeated measurements. If repeated measures are close together, our confidence in the results increases as they are more likely to be close to the real population value. Thus precision is a surrogate marker of accuracy. The judgement regarding precision is based on 95% confidence intervals and sample size.

(1) *Confidence intervals:* Confidence intervals represent a range of values based on sample data, in which the population value is likely to lie. Confidence intervals are the measure of the precision of a mean. In general, for systematic reviews, precision is adequate if 95% confidence intervals exclude no effect.

In hypothetical forest plots (Fig. 4) of two systematic reviews A and B, the confidence interval does not cross the line of no effect in the forest plot A indicating "no imprecision". In contrast, it crosses the line of no effect in systematic review B, indicating "imprecision".

(2) *Optimal information size*: The results of a systematic review are reliable only when the confounding factors which influence the outcome are balanced between the intervention and control groups. The confounding factors to be balanced between the two groups require a minimal number of patients, often referred to as "Optimal information size", randomised to either intervention or control group. Optimal information size equals to the number of patients required to conduct an adequately powered RCT.

The importance of fulfilling criteria for optimal information size is evident from the following example: A systematic review and meta-analysis compared intravenous magnesium versus placebo in patients with suspected myocardial infarction for prevention of death (Fig. 5) (Teo et al. 1991; Guyatt et al. 2011). The meta-analysis showed a significant beneficial effect of the intervention with an odds ratio of 0.44 and confidence intervals 0.27 to 0.71. Even though the effect estimate's confidence interval did not cross the line of no effect, one may not be confident in the results because of the small sample size and fewer events. In such



Forest Plot A

Study or Subgroup	Drug A Events	Drug A Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Study A	2	250	4	250	1.6%	0.50 [0.09, 2.73]
Study B	30	2500	70	2500	28.7%	0.42 [0.27, 0.65]
Study C	3	500	5	500	2.1%	0.60 [0.14, 2.51]
Study D	2	350	1	354	0.4%	2.03 [0.18, 22.48]
Study E	60	1000	90	1000	35.1%	0.65 [0.46, 0.91]
Study F	70	2000	80	2000	32.0%	0.87 [0.63, 1.21]
Total (95% CI)		6600		6604	100.0%	0.66 [0.54, 0.80]
Total events	167		250			

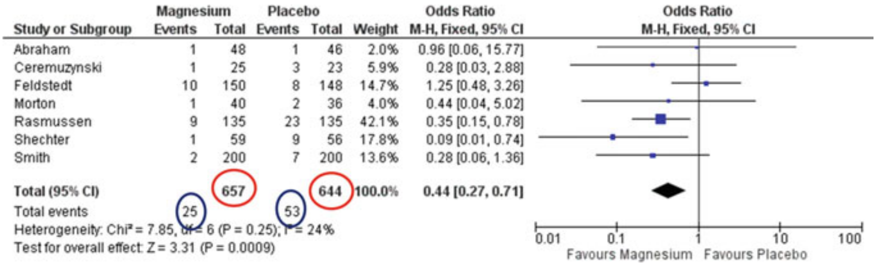
Heterogeneity: Chi² = 7.89, df = 5 (P = 0.16); I² = 37%
Test for overall effect: Z = 4.13 (P < 0.0001)

Forest Plot B

Study or Subgroup	Drug A Events	Drug A Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Study A	1	25	4	25	14.3%	0.22 [0.02, 2.11]
Study B	3	250	7	250	25.8%	0.42 [0.11, 1.65]
Study C	3	50	5	50	17.5%	0.57 [0.13, 2.55]
Study D	2	40	1	41	3.5%	2.11 [0.18, 24.18]
Study E	5	100	7	100	24.8%	0.70 [0.21, 2.28]
Study F	5	100	4	100	14.2%	1.26 [0.33, 4.85]
Total (95% CI)		565		566	100.0%	0.67 [0.37, 1.21]
Total events	19		28			

Heterogeneity: Chi² = 3.13, df = 5 (P = 0.68); I² = 0%
Test for overall effect: Z = 1.33 (P = 0.18)

Fig. 4 Hypothetical forest plots—judgement regarding imprecision



Study or Subgroup	Magnesium Events	Magnesium Total	Placebo Events	Placebo Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Abraham	1	48	1	46	2.0%	0.96 [0.06, 15.77]
Ceremuzynski	1	25	3	23	5.9%	0.28 [0.03, 2.88]
Feldstedt	10	150	8	148	14.7%	1.25 [0.48, 3.26]
Morton	1	40	2	36	4.0%	0.44 [0.04, 5.02]
Rasmussen	9	135	23	135	42.1%	0.35 [0.15, 0.78]
Shechter	1	59	9	56	17.8%	0.09 [0.01, 0.74]
Smith	2	200	7	200	13.6%	0.28 [0.06, 1.36]
Total (95% CI)		657		644	100.0%	0.44 [0.27, 0.71]
Total events	25		53			

Heterogeneity: Chi² = 7.85, df = 6 (P = 0.25); I² = 24%
Test for overall effect: Z = 3.31 (P = 0.0009)

Fig. 5 Forrest plot comparing intravenous magnesium versus placebo in patients with suspected myocardial infarction for prevention of death (Teo 1991)

situations, it may be reasonable to downgrade the certainty of the evidence for imprecision because of small information size.

Suggestions for downgrading: Do not downgrade for imprecision if optimal information size criterion is met, and confidence interval excludes no effect (i.e., relative risk (RR) of 1.0). Downgrade by one level if optimal information size criterion is not met or if the confidence interval fails to exclude significant benefit or harm (e.g., overlaps RR of 1.0). Reviewers may consider rating down by two if both the criteria (Confidence interval and optimal information size) are not met or when the confidence interval is very wide (Guyatt et al. 2011).

Publication/Reporting Bias

Publication bias is a reporting bias that results from failure to identify all relevant trials. Publication bias occurs from the publication or non-publication of relevant trials, depending on the nature and direction of the results (Sedgwick 2015). Trials with positive findings are more likely to be published than trials with negative or null findings (RR 1.78, CI 1.58 to 1.95) (Hopewell et al. 2009). Therefore, a meta-analysis in the presence of publication bias is likely to over-estimate the treatment effect. If a systematic review contains studies predominantly with small sample sizes or studies sponsored by the pharmaceutical industry, it increases publication bias. The pharmaceutical sector discourages publication of trials they supported, which have negative findings (Egger and Smith 1998).

The other sources of reporting bias include time-lag bias (delay in the publication of trials with negative findings), language bias (not including studies published in languages other than English), and bias arising from publication of trial in "grey literature" (e.g., theses, conference abstracts, un-indexed journals). These sources of bias prevent an eligible study from being identified and included in the systematic review.

The presence of reporting bias in a systematic review is assessed by visual inspection of the funnel plot for symmetry and Egger's test. Funnel plots are scatter-plots of the studies in a meta-analysis, with the treatment effect on the horizontal axis and some measure of weight, such as the inverse variance, the standard error, or the sample size, on the vertical axis (Lau et al. 2006). Generally, effect estimates from large studies will be more precise and will be near the apex of an imaginary funnel. In contrast, results from smaller studies will be less precise and would lie towards the funnel base evenly distributed around the vertical axis. Asymmetric distribution of the studies around the vertical axis raises the possibility of publication bias. However, apart from publication bias, a skewed funnel plot may result from other causes: by chance, true heterogeneity in the intervention effect, and statistics used to measure effect size.

Suggestions for downgrading: Consider rating down the evidence if the evidence is based mainly on multiple small trials, especially when industry-sponsored or investigators have conflicts of interest. Consider rating down the evidence when publication bias is strongly suspected based on funnel plot asymmetry. As there is no full-proof method to prove or rule out publication bias or to determine a threshold for publication bias, GRADE suggests systematic reviewers to decide whether publication bias was "undetected" or "strongly suspected" in a systematic review. Because of the uncertainty in assessing the likelihood of publication bias, GRADE suggests rating down by a maximum of one level when publication bias is strongly suspected (Guyatt et al. 2011).

Factors that can improve the certainty of evidence in systematic reviews of observational studies: Generally, evidence generated from observational studies is considered as "Low" certainty. However, in the following rare circumstances, observational studies can produce moderate or high certainty evidence.

(1) When methodologically robust observational studies show large or very large and consistent treatment effect, the treatment and effect relationship is likely to be stronger. In these situations, the reviewers may consider upgrading the certainty of evidence by one level.
(2) When studies show a dose-response effect, the effect is more likely related to the intervention. Hence the reviewers may consider upgrading certainty of evidence by one level.
(3) When plausible biases or confounding factors are likely to decrease the effects of an intervention, reviewers may consider upgrading the certainty of evidence by one level.

Summary of Findings Table

A Summary of Findings (SoF) table summarises the critical results of a systematic review. It also informs the readers about the level of reviewer's confidence in the results based on the GRADE approach. The SoF table allows reviewers to make explicit judgements regarding the certainty of evidence and readers to understand the reasoning behind the judgements. The GRADEpro Guideline Development Tool (GRADEpro GDT) is online software (available at https://gradepro.org/) used to create a summary of findings table for systematic reviews.

Summary

GRADE offers a system for rating certainty of evidence in systematic reviews. In this chapter, we have discussed the critical aspects that systematic review authors need to consider while grading the certainty of evidence. The GRADE process requires judgement and is not objective, but it does provide a transparent and well-defined method for developing and presenting evidence summaries for systematic reviews.

References

Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.

Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401–406.

Egger M, Smith GD. Bias in location and selection of studies. BMJ. 1998;316(7124):61–6.

Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence–imprecision. J Clin Epidemiol. 2011;64(12):1283–1293.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence–inconsistency. J Clin Epidemiol. 2011;64(12):1294–1302.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence–indirectness. J Clin Epidemiol. 2011;64(12):1303–1310.

Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence–publication bias. J Clin Epidemiol. 2011;64(12):1277–1282.

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence–study limitations (risk of bias). J Clin Epidemiol. 2011;64 (4):407–415.

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):924–6.

Higgins JPT SJ, Page MJ, Elbers RG, Sterne JAC. Chapter 8: Assessing risk of bias in a randomised trial. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook of systematic reviews of interventions, 2nd edn. Chichester (UK): John Wiley and Sons; 2019. p. 205–228.

Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–60.

Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009; (1): Mr000006.

Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ. 2006;333(7568):597–600.

Puhan MA, Schünemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ. 2014;349:g5630.

Schunemann HJ HJ, Vist GE, Glasziou P, Akl EA, Skoetz N, Guyatt GH. Chapter 14: completing 'summary of findings' tables and grading the certainty of the evidence. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions, 2nd edn. Chichester (UK): John Wiley and Sons; 2019. p. 375–402.

Sedgwick P. What is publication bias in a meta-analysis? BMJ. 2015;351:h4419.

Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ. 2009;338:b1147.

Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. BMJ. 1991;303(6816):1499–503.

Reporting of Meta-Analysis (PRISMA)

Sam Athikarisamy and Sanjay Patole

Abstract Reporting is the final step of the systematic review process. An accurate and reliable reporting of systematic reviews assists the end-users (clinicians, policymakers, funding agencies, guideline developers) in making informed and evidence-based decisions. Poor quality of reporting of systematic reviews was recognised as an issue as far back as the late'80s. An international group hence drafted the QUOROM statement (Quality of Reporting of Meta-analyses) in 1996 to provide guidelines for improving the quality of reporting. These guidelines were updated after addressing the conceptual and methodological issues and renamed as PRISMA (Preferred Reporting Items of Systematic reviews and Meta-Analyses) by a group of authors. PRISMA is an evidence-based minimum set of items for reporting systematic reviews and meta-analyses of randomised controlled trials (RCTs). However, it can also be used for reporting systematic reviews of non-RCTs. PRISMA guidelines help to convey the information transparently. This chapter is focussed on PRSMA checklist and its 27 items under seven domains.

Keywords Bias · Meta-analysis · PRISMA · Randomised controlled trials · Reporting · Systematic reviews

S. Athikarisamy (✉) · S. Patole
Neonatal Directorate, King Edward Memorial Hospital for Women,
Perth, WA 6008, Australia
e-mail: sam.athikarisamy@health.wa.gov.au

S. Patole
e-mail: sanjay.patole@health.wa.gov.au

S. Athikarisamy · S. Patole
School of Medicine, University of Western Australia, Perth, WA 6009, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_11

Introduction

Reporting is the final step of the systematic review process. An accurate and reliable reporting of a systematic review assists the end-users (clinicians, policymakers, funding agencies, guideline developers) in making informed and evidence-based decisions (Moher et al. 2007a) Based on the inconsistent quality of reporting of the 300 systematic reviews evaluated, Moher et al. reiterated their call for adhering to the PRISMA guidelines for reporting systematic reviews and the supporting document with explanation and elaboration (Moher et al. 2007a, 2009; Liberati et al. 2009).

History of PRISMA

Poor quality of reporting of systematic reviews was recognised as an issue in the late'80 s (Mulrow 1987; Sacks et al. 1987). An international group drafted the QUOROM statement (Quality of Reporting of Meta-analyses) to provide guidelines for improving the quality of reporting and published it in 1999(Moher et al. 1999). It was updated after addressing the conceptual and methodological issues and renamed as PRISMA (Preferred Reporting Items of Systematic reviews and Meta-Analyses) by a group of 29 review authors (Liberati et al. 2009; Moher et al. 2000). PRISMA is an evidence-based minimum set of items for reporting systematic reviews and meta-analyses of randomised controlled trials (RCTs). However, it can also be used for reporting systematic reviews of other types of research such as non-RCTs. PRISMA guidelines help to covey the information transparently. Most peer-reviewed journals have made their use mandatory for manuscripts to be considered for publication. PRISMA guidelines can also be used for critical appraisal of systematic reviews.

PRISMA Checklist

The application of PRISMA guidelines involves going through the checklist containing 27 items under the seven domains (Liberati et al. 2009). Six of the seven domains relate to the standard sections of the manuscript (title, abstract, introduction, methods, results, discussion) whereas the seventh relates to funding. The other important component of the PRISMA statement is the four-phase flow diagram (Fig. 1). The PRISMA statement gives a clear explanation of the checklist and the rationale for each of the item listed. The statement and its extensions are available free of cost (http://prisma-statement.org/)

This chapter briefly discusses the various items under the seven domains of the PRISMA guidelines.

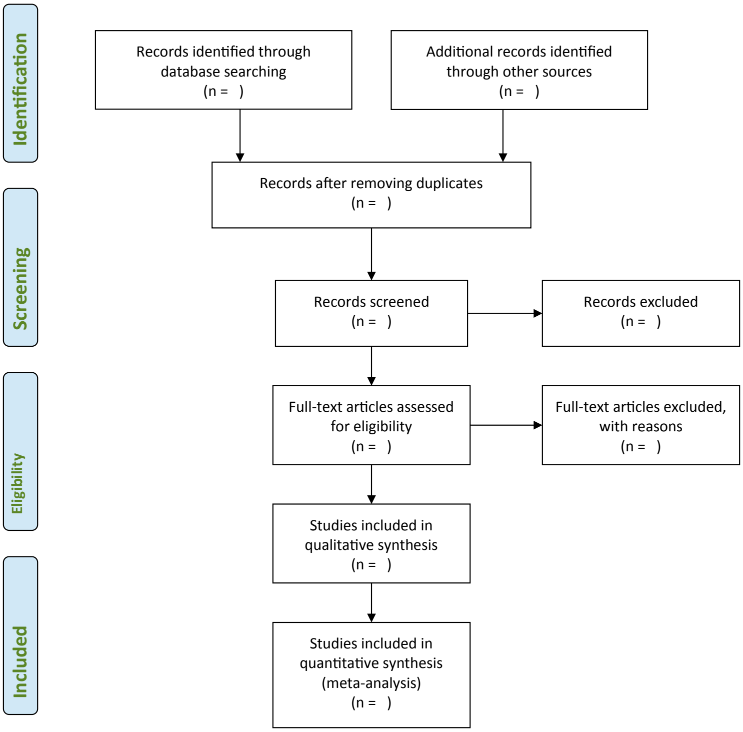

Identification
Screening
Eligibility
Included
Records identified through database searching (n =)
Additional records identified through other sources (n =)
Records after removing duplicates (n =)
Records screened (n =)
Records excluded (n =)
Full-text articles assessed for eligibility (n =)
Full-text articles excluded, with reasons (n =)
Studies included in qualitative synthesis (n =)
Studies included in quantitative synthesis (meta-analysis) (n =)

Fig. 1 Template for the study selection process

Sections/Topics and Checklist Items

As mentioned earlier, there are seven standard sections when writing a research manuscript. Under the PRISMA guidelines, each of these sections is assessed for accuracy of reporting using a checklist. For example, under the section 'Title' there is only one item to be checked, whereas for the section 'Methods' there are 12 items to be checked.

1. Title

Titles are important as they make the first impression of the article. Authors should identify their article by stating it as a 'systematic review' or 'meta-analysis' in the title itself (Moher et al. 2000). Moher et al. showed that only 50% of the reviews (n = 300) used the term "systematic review" or "meta-analysis" in the title or abstract (Moher et al. 2007b). As far as possible, the title should also reflect the PICOS approach (participants, intervention, comparators, outcomes and study design). Titles can be ‘indicative’ or ‘declarative’. An indicative title discloses the topic matter (e.g. *“Probiotics versus placebo for prevention of necrotising enterocolitis in preterm infants—a meta-analysis of randomised controlled trials”).* In contrast, the declarative title discloses the main conclusion of the review (e.g. *“Probiotics prevent necrotising enterocolitis in preterm infants—A meta-analysis of randomised controlled trials”*) (Deshpande et al. 2007). Evidence suggests that the style of the title (indicative vs declarative) influences the number of citations, probably because of the way electronic search of the literature is conducted (Jacques and Sebire 2010).

2. Structured Summary

A structured summary covering the rationale and objective of the review should be provided. The details of data sources (e.g. major databases, grey literature, trial registries, language restriction), study selection criteria (e.g. RCT vs non RCT) and method of data extraction (e.g. pre-defined data fields) need to be clarified. Results are a vital section in the abstract as readers often read only the results to draw a conclusion. The abstract should include whether the study results were pooled or not, and if pooled, the type of model used (Random effects vs fixed effect) for meta-analysis should be reported. The conclusion should link the study objective to the results. Despite the world limit, it is always possible to get the readers attracted towards reading the full article by a well written abstract. The abstract should provide a standalone summary, including the validity and applicability and should be easily identifiable in the literature search (Beller et al. 2013).

Introduction

The introduction section sets the stage for what is going to come and covers the rationale and the objective of the study.

3. Rationale

Every research proposal should start or end with a systematic review (Mulrow 1994). The rationale provides the reason for conducting the systematic review and clarifies how it will add information to what is already known in the field. It is important to clarify if it is the first review or update. In the case of the later, the reason for the update should be provided.

4. Objectives

The objective of the review should be based on the PICOS framework, and be simple enough to explain the scope of the review. It helps to determine other aspects of the review, namely the eligibility criteria and searching for relevant literature. Depending on the questions being asked the review can be further classified based on whether the focus is narrow or broad. Examples of such types would be (e.g. Narrow focus: Aspirin for prevention of myocardial infarction in elderly patients (>70y), with diabetes; Broad focus: Aspirin for prevention of myocardial infarction in adults). The advantage of the broad focus question is that it increases the generalizability of the results. But it may not answer an 'individual-specific' question.

Methods

This is the most important section of a manuscript and contains the maximum number (12) of checklist items. It should be presented in detail and with clarity so that readers are able to reproduce the results.

5. Protocol and Registration

Writing a protocol is an important part of a systematic review to pre-specify the objectives and methods for search. PRISMA-P provides a method for writing a protocol which can be registered with the PROSPERO (International Prospective Register of Ongoing Systematic Reviews) registry (Gallucci et al. 2018; PRISMA statement 2020). This critical step reduces the risk of duplication and publication bias (PRISMA statement 2020; Moher et al. 2015; PROSPERO 2020; Straus and Moher 2010). If applicable, the protocol details (registration number and registry) should be provided (Liberati et al. 2009).

6. Eligibility Criteria

Predetermined eligibility criteria are fundamental for assessing the validity of a systematic review (Moher et al. 2000; McCrae et al. 2015). The details of the included studies include the study type, years considered, language restrictions, type of patients (e.g. age restriction), and details of the intervention and outcomes.

7. Describe the Information Sources

Description of the databases (e.g. Embase, PsycINFO, Web of Science), platforms, provider (PubMed, MEDLINE), and the date of starting and ending the search should be mentioned. The number of authors who conducted the search and their role needs to be clarified. It is important to specify if any of the study authors were contacted for additional data or clarifications. Other sources of information like a non-English database (e.g. LILACS), (http://lilacs.bvsalud.org/en) trial registries (e.g. Cochrane Central, www.clinicaltrials.gov, WHO Trial registry), (http://apps.who.int/trialsearch) proceedings of conferences, and grey literature need to be reported.

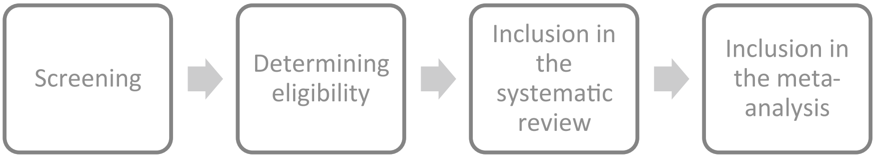
8. Search Methodology

It is important to report whether the search strategy was peer-reviewed. A detailed description of the full search strategy should be provided for at least one major database to facilitate reproducibility. The search terms (keywords and Mesh terminology) (https://learn.nlm.nih.gov/documentation/training-packets.T0042010P/)

and any constraints in the search (e.g. limited access, expertise, resources) should be disclosed. Some journals allow the full strategy to be uploaded as a supplement. It is important to save the search output for future updates.

Here is an example of search strategy: "*With consultation from a professional research librarian ...., we developed a search strategy to identify RA RCTs that were published in MEDLINE between January 1, 2008, and January 1, 2018. We used "Rheumatoid Arthritis" and "humans" as Medical Subject Headings (MeSH) terms and "randomised controlled trial" as a publication type to generate the following search phrase: "arthritis, rheumatoid" (MeSH) AND ("2008/01/01" [PDAT]: "2018/01/01" [PDAT] AND "humans" [MeSH Terms]) AND ("randomised controlled trial" [ptyp] AND "humans" [MeSH Term]). Filters were applied to identify English-language studies in adults 19 years and older. Our search was limited to RCTs published within ten years, since there was a tremendous and unprecedented increase in RA RCTs during this period, allowing us to examine a very large sample size. The English-language filter was applied to increase the likelihood that studies would have at least 1 US-based site*" (Strait et al. 2019).

9. Description of Study Selection


Screening
Determining eligibility
Inclusion in the systematic review
Inclusion in the meta-analysis

The final report should include how the reviewers selected studies as per the protocol. Authors choose from a large number of studies according to the eligibility criteria. Hence it is important to report the screening process, i.e. how many authors selected the articles, and how often it was necessary to go to full-text articles. The reviewers may disagree about the inclusion or exclusion of the studies at the screening stage. In such cases, the process for resolving the disagreement needs to be stated, including the level of inter-rater agreement, and how often arbitration about selection was required. The commonly reported methods for this purpose include the percent agreement and Cohen's kappa (≤ 0 = no agreement, 0.01–0.20 = none to slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = perfect agreement) (Park and Kim 2015; McHugh 2012; Chow et al. 2019).

10. Data collection process

Reporting the details of data extraction (e.g. pre-piloted data extraction form, number of authors who independently retrieved the data, criteria for selecting studies for data extraction, type of data collected) is important. The data extraction form can be provided as supplemental material (https://bmjopen.bmj.com/content/

bmjopen/7/6/e01562). It is important to report how discrepancies about data extraction were handled, and whether the authors of included studies were contacted for additional data or clarifications.

11. Data items

All individual data items extracted from each included study should be reported in the manuscript. The reasons for the unavailability of data should be reported to minimise bias and optimise transparency.

12. Describe methods used for assessing the risk of bias in individual studies

Assessing the risk of bias in included studies is a critical component of a systematic review as it gives strength to the body of evidence. The number of reviewers who independently assessed the adequacy of randomisation, concealment of allocation, and blinding of participants, health care providers and outcome assessors should be reported in the manuscript. The commonly used methods include the *"components approach"* which involves assessing individual items that reflect the methodological risk of bias, or other relevant considerations in the body of literature under study (e.g. whether randomisation sequence was concealed or not). The second method uses a *composite approach* that combines different components related to the risk of bias or reporting into a single overall score (Viswanathan et al. 2012).

13. Summary measures

Authors should clearly state the summary measures (e.g. risk ratio, the difference in means) and their 95% confidence intervals, and the model (fixed effect vs random effects) used for meta-analysis. Deriving summary measures by meta-analysis may not be possible if the units of measurements for the outcomes of interest are different in the included studies. The inability to do so should be clearly stated (Tripepi et al. 2007; Elwenspoek et al. 2019).

14. Planned methods of analysis

The methods for handling the data and pooling the results from the included studies and measures of consistency for each meta-analysis need to be described. Reporting the method for addressing/exploring heterogeneity between the included studies is important. The issue of heterogeneity has been covered in other chapters of this book.

15. Risk of bias across studies

Describing the methods used for assessment of publication bias, and statistical heterogeneity is important as it may affect the cumulative evidence. This includes describing the funnel plot if it was used and how it was assessed (informally by visual inspection and formally with a statistical test such as the Egger's test). The reasons for the risk of bias include missing studies (publication bias) and missing data from the included studies (selective reporting bias (van Enst et al. 2014; Oberoi et al. 2020)).

16. Additional analysis

Additional analyses such as sensitivity analysis, subgroup analysis and meta-regression may be required to increase the quality of the review. It is essential to report the rationale for additional analyses and whether they were pre-specified or not.

Results

17. Study selection

The study selection process should be reported using a flow diagram (Fig. 1). The four-phase flow diagram gives information on the total search output, removal of duplicates, number of excluded and included studies. The reasons for exclusion can be explained in a tabular form if the numbers are small. It may be preferable to use a different flow diagram for each outcome if multiple outcomes are assessed in a systematic review (Moher et al. 2007a).

18. Study characteristics

The characteristics of the included studies (e.g. author, year, site, sample size, PICOS, follow up) should be reported in detail, preferably in a tabular form (Table 1). Ideally, study level data should be summarised to compare the main characteristics of the included studies. Adequate details should be provided to assist the readers in judging the relevance of included studies. The reviewers are expected to contact the individual study authors for missing data or clarifications to assure that there are no missing data in this table. Reviewers must not assume anything about the missing data and acknowledge the facts clearly.

19. Risk of bias within studies

The risk of bias within included studies should be reported using a standard approach, as explained in item 12. The manuscript should contain the results of the risk of bias assessment for each outcome in every study. The methodological

Table 1 Characteristics of included studies

Author/ year/ Country	Sample size	Participants	Intervention	Comparator	Outcomes	Follow up	Author comments
XY 2020, Italy	1200						

limitations should be acknowledged. The limitations of each study should be presented in the tabular form in addition to a narrative summary in the text. The following markers measure the validity of the studies: (1) Concealment of randomisation (2) Whether the trial stopped early (3) Patients blinded or not (4) Health care providers blinded or not (5) Data collectors blinded or not (6) Outcome assessors blinded or not. The Cochrane collaboration has recently updated the ROB tool (Sterne et al. 2019).

20. Results of individual studies

The results sections should include all outcomes, including safety. For each study, the summary data for effect estimates and their confidence intervals, ideally with a forest plot (Fig. 2), should be provided.

21. Synthesis of results

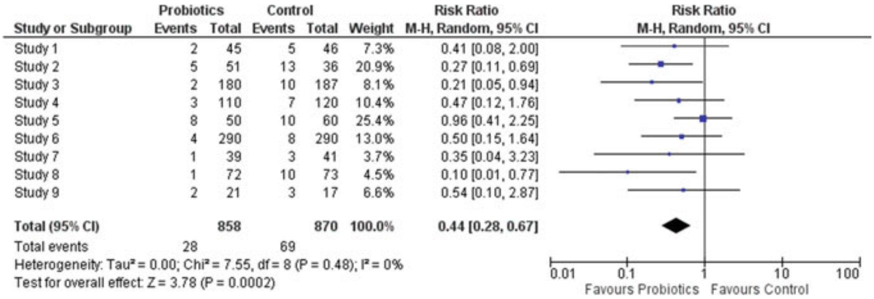
The synthesis of results should be reported systematically. Multiple outcomes should be presented with different forest plots. The main results of the review are presented using effect estimates and their confidence intervals. A qualitative narrative is useful when data on a particular outcome may not have been reported by all studies.

22. Risk of bias across studies

Presenting the results of risk of bias assessments across studies which can affect the cumulative evidence is important (e.g. publication bias, selective reporting). The number of studies should be adequate (>10) to generate a funnel plot (Sterne et al. 2011; Dwan et al. 2008).

23. Additional analysis

The results of all additional analyses, (e.g. sensitivity or subgroup analyses, meta-regression) should be provided to avoid selective reporting bias (Nelson et al. 2020; Bhangu et al. 2012).



	Probiotics		Control			Risk Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Random, 95% CI
Study 1	2	45	5	46	7.3%	0.41 [0.08, 2.00]
Study 2	5	51	13	36	20.9%	0.27 [0.11, 0.69]
Study 3	2	180	10	187	8.1%	0.21 [0.05, 0.94]
Study 4	3	110	7	120	10.4%	0.47 [0.12, 1.76]
Study 5	8	50	10	60	25.4%	0.96 [0.41, 2.25]
Study 6	4	290	8	290	13.0%	0.50 [0.15, 1.64]
Study 7	1	39	3	41	3.7%	0.35 [0.04, 3.23]
Study 8	1	72	10	73	4.5%	0.10 [0.01, 0.77]
Study 9	2	21	3	17	6.6%	0.54 [0.10, 2.87]
Total (95% CI)		**858**		**870**	**100.0%**	**0.44 [0.28, 0.67]**
Total events	28		69			

Heterogeneity: Tau² = 0.00; Chi² = 7.55, df = 8 (P = 0.48); I² = 0%

Test for overall effect: Z = 3.78 (P = 0.0002)

Fig. 2 Forest plot with summary data

Discussion

The section should include the summary of evidence, limitations and the conclusions of the review.

24. Summary of evidence

The summary should include the main findings and key messages. The strengths of the evidence should be stated clearly for all important outcomes, including safety. The evidence should be clearly directed to different stakeholders.

Example: "*Compared with vaginal delivery, cesarean delivery was associated with increased risk of dysbiosis in preterm infants*."

"*The strengths of our review include the robust methodology and comprehensive literature search with no restriction on language and set up, and with a low probability of publication bias*."

"*Future studies should carefully assess the long term implications of dysbiosis associated with cesarean delivery*" (Zhang et al. 2019).

25. Limitations

Addressing the limitations of the review is an integral part of the discussion. Future trials and reviewers should take these into consideration. The limitations can occur at a study and outcome level (e.g. risk of bias) or review level (e.g. publication bias, reporting bias). This part of the discussion should address the validity and limitations of the review process, including the limitations of search and applicability of the findings.

Example: Authors of a systematic review assessing the prevalence of zinc deficiency and associated factors among pregnant women and children in Ethiopia reported the following limitations which could have impacted their finding: *Small number of included studies, studies non-representative of all regions in Ethiopia, study design (cross-sectional), lack of information regarding the processing of specimens, inability to perform sub-group analysis due to the small number of studies, and non-compliance with recommendations in included studies* (Berhe et al. 2019).

26. Conclusions

The conclusions should include the general interpretation of results in the context of the review. Negative conclusions are as important as positive conclusions. The reviewers should clearly acknowledge if the findings are inconclusive. Gaps in knowledge should be identified to guide further research. A study assessing the ability of doctors and medical students to derive independent and appropriate conclusions from systematic reviews showed that the majority of the participants lacked this skill. The investigators concluded that authors, editors and reviewers should make an effort that the conclusions of a paper accurately reflect the results (Lai et al. 2011).

27. Funding

Reporting the source of funding or any other support (e.g. using the data from the manufacturing company) is essential to avoid any perception of conflict of interest. Funding should be acknowledged even if it is by a health care agency or an academic institution. The review by Moher et al. showed that at least 0% reviews did not mention the funding source Moher et al. (2007a). Bes-Rastrollo et al. have reported that systematic reviews with sponsorship or conflicts of interest with food or beverage companies were five times more likely to conclude no positive association between sugars sweetened beverages consumption and weight gain or obesity compared to reviews without sponsorship (Bes-Rastrollo et al. 2013).

In summary, reporting a systematic review based on the PRISMA guidelines allows the readers to understand the process and finding of the review systematically and transparently. While the PRISMA statement has improved the quality of the reporting, there is still scope for improvement. Adherence to PRISMA guidelines (Updated in 2020) by authors and journals is essential to enhance the transparency in reporting of a systematic review (Page et al. 2020).

References

Beller EM, Glasziou PP, Altman DG et al. PRISMA for Abstracts Group. PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. PLoS Med. 2013; 9: e1001419.

Berhe K, Gebrearegay F, Gebremariam H. Prevalence and associated factors of zinc deficiency among pregnant women and children in Ethiopia: a systematic review and meta-analysis. BMC Public Health. 2019;19:1663. https://doi.org/10.1186/s12889-019-7979-3.

Bes-Rastrollo M, Schulze MB, Ruiz-Canela M, Martinez-Gonzalez MA. Financial conflicts of interest and reporting bias regarding the association between sugar-sweetened beverages and weight gain: A systematic review of systematic reviews. PLoS Med. 2013;10: https://doi.org/10.1371/journal.pmed.1001578.

Bhangu A, Nepogodiev D, Gupta A, Torrance A, Singh P, West Midlands Research Collaborative. Systematic review and meta-analysis of outcomes following emergency surgery for Clostridium difficile colitis. Br J Surg. 2012; 99:1501–13.

Chow CHT, Rizwan A, Xu R, et al. Association of temperament with preoperative anxiety in pediatric patients undergoing surgery: a systematic review and meta-analysis. JAMA Netw Open. 2019;2: https://doi.org/10.1001/jamanetworkopen.2019.5614.

Data Collection Process. Supplementary file. https://bmjopen.bmj.com/content/bmjopen/7/6/e015626/DC1/embed/inline-supplementary-material-1.pdf?download=true. Accessed 25 Aug 2020.

Deshpande G, Rao S, Patole S. Probiotics for prevention of necrotising enterocolitis in preterm neonates with very low birthweight: a systematic review of randomised controlled trials. Lancet. 2007;369:1614–20.

Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS ONE. 2008;3:e3081.

Elwenspoek MMC, Sheppard AL, McInnes MDF, et al. Comparison of multiparametric magnetic resonance imaging and targeted biopsy with systematic biopsy alone for the diagnosis of prostate cancer: A systematic review and meta-analysis. JAMA Netw Open. 2019;2: https://doi.org/10.1001/jamanetworkopen.2019.8427.

Gallucci GO, Hamilton A, Zhou W, Buser D, Chen S. Implant placement and loading protocols in partially edentulous patients: a systematic review. Clin Oral Implants Res. 2018;29(Suppl 16):106–34.

International clinical trials registry Platform ICTRP Search Portal. Apps.who.int. 2020. http://apps.who.int/trialsearch. Accessed 25 Aug 2020.

Jacques TS, Sebire NJ. The impact of article titles on citation hits: an analysis of general and specialist medical journals. JRSM Short Rep. 2010;1:1–5.

Lai NM, Teng CL, Lee ML. Interpreting systematic reviews: are we ready to make our own conclusions? a cross-sectional study. BMC Med. 2011;9:30. https://doi.org/10.1186/1741-7015-9-30.

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6: https://doi.org/10.1371/journal.pmed.1000100.

LILACS. Lilacs.bvsalud.org. 2020. http://lilacs.bvsalud.org/en. Accessed 25 Aug 2020.

McCrae N, Blackstock M, Purssell E. Eligibility criteria in systematic reviews: a methodological review. Int J Nurs Stud. 2015;52:1269–76.

McHugh ML. Interrater reliability: the kappa statistic. Biochemia Med. 2012;22:276–82.

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: quality of reporting of Meta-analyses. Lancet. 1999; 354:1896–900.

Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009; 21;6 (7): e1000097.

Moher D, Shamseer L, Clarke M. et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015; 4:1. https://doi.org/10.1186/2046-4053-4-1.

Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med. 2007a;27:e78.

Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med. 2007b;4:e78.

Moher D, Coo DJ, Eastwood S, Olin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analysis of randomised controlled trials: The QUOROM Statement. Br J Surg. 2000;87:1448–54.

Mulrow CD. The medical review article: state of the science. Ann Intern Med. 1987;1(106):485–8.

Mulrow CD. Systematic reviews: rationale for systematic reviews. BMJ. 1994;309:597–9.

Nelson LF, Yocum VK, Patel KD, Qeadan F, Hsi A, Weitzen S. Cognitive outcomes of young children after prenatal exposure to medications for opioid use disorder: A systematic review and meta-analysis. JAMA Netw Open. 2020;3: https://doi.org/10.1001/jamanetworkopen.2020.1195.

Oberoi S, Yang J, Woodgate RL, et al. Association of mindfulness-based interventions with anxiety severity in adults with cancer: A systematic preview and meta-analysis. JAMA Netw Open. 2020;3: https://doi.org/10.1001/jamanetworkopen.2020.12598.

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71doi: https://doi.org/10.1136/bmj.n71 (Published 29 March 2021)

Park CU, Kim HJ. Measurement of inter-rater reliability in systematic review. Hanyang Med Rev. 2015;35:44–9.

PRISMA statement. http://www.prisma-statement.org/. Accessed 25 Aug 2020.

PROSPERO. Crd.york.ac.uk. 2020. https://www.crd.york.ac.uk/PROSPERO. Accessed 25 Aug 2020.

PubMed® Online Training. Learn.nlm.nih.gov. 2020. https://learn.nlm.nih.gov/documentation/training-packets.T0042010P/Accessed 25 Aug 2020.

Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomised controlled trials. N Engl J Med. 1987;316:450–5.

Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002.

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.

Strait A, Castillo F, Choden S, et al. Demographic characteristics of participants in rheumatoid arthritis randomized clinical trials: a systematic review. JAMA Netw Open. 2019;2: https://doi.org/10.1001/jamanetworkopen.2019.14745.

Straus S, Moher D. Registering systematic reviews. CMAJ. 2010;182:13–4.

Tripepi G, Jager KJ, Dekker FW, Wanner C, Zoccali C. Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat'. Kidney Int. 2007;72:789–91.

van Enst WA, Ochodo E, Scholten RJ et al. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. BMC Med Res Methodol. 2014; 70. https://doi.org/10.1186/1471-2288-14-70.

Viswanathan M, Ansari MT, Berkman ND, et al. Assessing the risk of bias of individual studies in systematic reviews of health care interventions 2002. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008. https://www.ncbi.nlm.nih.gov/books/NBK91433/.

Zhang T, Sidorchuk A, Sevilla-Cermeño L et al. Association of cesarean delivery with risk of neurodevelopmental and psychiatric disorders in the offspring: a systematic review and meta-analysis. JAMA Netw Open. 2019; 2:e1910236.

Critical Appraisal of Systematic Reviews and Meta-Analyses

Sanjay Patole

Abstract Systematic reviews are the most reliable and comprehensive statement about what works. They focus on a specific question and use clearly stated, prespecified scientific methods to identify, select, assess, and summarise the findings of similar but separate studies. A systematic review may or may not contain a meta-analysis for various reasons. Given the hierarchy of evidence-based medicine, a systematic review and meta-analysis are expected to provide robust evidence to guide clinical practice and research. However, the methodological rigour (design, conduct, analysis, interpretation, and reporting) of both, the systematic review and meta-analysis and the included studies deserve equal attention for judging the validity of the findings of a systematic review. Reproducibility is a critical aspect of science. Without transparency about what was done, and how it was done, it is difficult to reproduce the results, questioning the validity of any study. This chapter focuses on the critical appraisal of a systematic review and meta-analysis based on their principles and practice.

Keywords Systematic reviews · Meta-analysis · Critical appraisal · Validity · Reproducibility

Introduction

Before we proceed to discuss the critical appraisal of a systematic review and meta-analysis, a quick recap of their principles is necessary. As discussed before, systematic reviews are ‘the most reliable and comprehensive statement about what works (van der Knaap et al. 2008). They focus on a specific question and use clearly stated, prespecified scientific methods to identify, select, assess, and summarise the

S. Patole (✉)
School of Medicine, University of Western Australia, Perth, WA 6009, Australia
e-mail: sanjay.patole@health.wa.gov.au

S. Patole
Neonatal Directorate, King Edward Memorial Hospital for Women,
Perth, WA 6008, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_12

findings of similar but separate studies. A systematic review may or may not contain a meta-analysis for two reasons, either the data from available studies are not possible to be combined or, it does not make sense to combine the data from different studies together due to significant heterogeneity. The phrase, ‘combining apples with oranges’ is often used when faced with the latter scenario (Esteves et al. 2017; Purgato and Adams 2012). Following on to this principle, combining good apples with bad or rotten apples is also not appropriate.

It is essential to appreciate that a systematic review and meta-analysis can be conducted for various types of study, including randomised, non-randomised (cohort and case-control) and diagnostic accuracy studies. Even case reports prepared using a comprehensive, robust and transparent methodology, can be reviewed systematically (Jeong et al. 2019). Given the hierarchy of evidence-based medicine, a systematic review and meta-analysis of data from of randomised controlled trials (RCTs)—the gold standard of clinical research, is thus expected to provide robust data to guide clinical practice and further research. However, the methodological rigour (design, conduct, analysis, interpretation, and reporting) of both, the systematic review and meta-analysis and the included RCTs deserve equal attention for judging the validity of the findings of a systematic review (Smith and Hing 2011; Ioannidis 2016). Similar argument can be made in the interpretation of a systematic review of studies of other designs.

As for its characteristics, a *‘TRUE’* systematic review has transparent (T) and robust methodology for reproducibility (R) and is unbiased (U) with best precautions to minimise the risk of bias, with explicit (E) objective criteria for each step. Reproducibility is a critical aspect of science. Without transparency about what was done, and how it was done, it is difficult to reproduce the results, questioning the validity of any study (Lakens et al. 2016; Shokraneh 2019).

This chapter focuses on critical appraisal of a systematic review and meta-analysis based on their principles and methodology (Egger et al. 1997; da Costa and Juni 2014; Phan et al. 2015; Brown et al. 2012; Jones et al. 2008; Roever and Zoccai 2015).

Step 1: *Does the systematic review ask a focused, well defined, clinically useful, and importantly, an answerable question?*

Based on the conventional ‘PICO’ format, the title of a systematic review should clarify the four critical aspects of the question being asked (P: Patients/participants, I: Intervention, C: Control/Comparison, O: Outcome of interest, apart from the type of included studies (e.g. RCTs, non-RCTs) addressing the question (i.e. study design: S) (Richardson et al. 1995). The titles of some reviews will also convey the time frame (T) or the setting to which the results will be applicable, i.e. reflect the external validity of the effects of the intervention under study.

Step 2: *Have the reviewers justified the need for a systematic review?*

The number of systematic reviews published every year has been on the rise (Bastian et al. 2010; Fuhr and Hellmich 2015). It is essential to check if the systematic review was really needed. Every systematic review should clarify whether the question has not been addressed before or there are valid reasons for an update (Garner et al. 2016; Bashir et al. 2018).

Step 3: *Assessing the methodology for literature search*

(1) *Was the search strategy robust, comprehensive, and transparent?* In order to provide the best available evidence and assure reproducibility, every systematic review should have a clearly documented *and* comprehensive search strategy for tracing all relevant studies-published as well as unpublished (Cooper et al. 2018).

(2) *What type of studies were searched?* A detailed description of the type of studies (e.g. RCTs, non-RCTs, diagnostic accuracy studies) that will be searched is required.

(3) *Which databases were searched?* The comprehensiveness of any systematic review is limited only by human resources. It is possible that what is difficult to find may not be of good quality (van Driel et al. 2009). However, this may not be true. The obsession for significant 'p' values and the bias against studies with 'negative' findings mean many potentially valuable studies may not find a place in the conventional domain for academic publications (Simonsohn et al. 2014). At the minimum, at least four major databases (MEDLINE, Embase, Web of Science, Google Scholar) need to be searched for the optimal search of the literature (Bramer et al. 2017). However, it is important to note that literature goes way beyond conventional databases (Cooper et al. 2018). In the context of neonatology, the search commonly involves Medline, Embase, Emcare, Cochrane Central library, Google Scholar, and grey literature. Contacting experts in the field and crosschecking cross-references from reviews, and proceedings of relevant conferences are also necessary.

(4) *What were the search terms?*

Systematic reviews must provide details of the search strategy for at least one major database allowing readers to crosscheck/reproduce search output. Availability of open access makes it possible to submit the complete search strategy as online supplementary material.

(5) *Who conducted the literature search, and how?*

To avoid bias and human errors, it is essential that at least two reviewers independently search the literature and crosscheck with each other. Differences of opinion in the assessment of various aspects of the studies including simple issues such as deciding whether a study is a duplicate or has significant overlap with another publication needs robust discussion amongst the entire team, and if required, contact the authors of the studies in question. A clear description of these issues and processes is vital to assure the validity of a systematic review.

(6) *Do the numbers tally*

A robust, and comprehensive literature search and transparent decisions based on prespecified inclusion-exclusion criteria, finally leads to the number of studies included in the review. The PRISMA flow chart should provide the details of initial search output (potentially eligible studies) from each source, the number of studies excluded, the reasons for their exclusion, ending with the final number of studies included in the systematic review, and if applicable, meta-analysis (Liberati et al. 2009). The numbers in the PRISMA flow chart should tally and match those quoted in the core manuscript.

(7) *How was the risk of bias (ROB) assessed in the included studies?*

There should be a clear pre-stated strategy for ROB assessment appropriate for the study type included in the systematic review. For example, this will be the ROB tool recommended by the Cochrane Collaboration for Systematic Reviews of RCTs, and the New Castle Ottawa Scale or ROBINS-I tool for assessing ROB in non-RCTs (Wells et al. 2000; Sterne et al. 2016). Clear description of the reasons for judging the ROB is critical. For example, in a systematic review of RCTs judging the four crucial aspects of an RCT, i.e. randomisation, allocation concealment, blinding and completeness of follow up, requires meticulous attention. It is important to check if the reviewers have made every possible effort, including contacting the authors of the included trials, for accurate assessment of various domains of ROB tool.

Step 4: *Is the rationale for quantitative (i.e. meta-analysis) vs. qualitative synthesis of results provided?*

Deciding whether a meta-analysis is justified or not, is perhaps the most crucial step in systematic reviews. The markers of clinical heterogeneity such as PICO characteristics, study design, settings, and period, as well as the results of ROB assessment, are taken in consideration to make a clinical judgement whether it is sensible to pool the results from different studies for a meta-analysis. The reviewers should provide clear information in this context (Melson et al. 2014; Chess and Gagnier 2016; Kriston 2013; Gagnier et al. 2012; Malone et al. 2014).

A meta-analysis is possible if it is judged that the studies included in the systematic review are 'more or less similar'. However, it can be conducted only if the data on the outcome/s of interest are available in a format suitable for pooling. Therefore the methodology should pre-specify the format of the data required for meta-analysis for different types of outcomes (Categorical and continuous) and whether a reliable method of conversion was required to enable pooling (e.g. Hozo's formula for deriving the mean and standard deviation from median and range) (Hozo et al. 2005).

Step 5: *If the systematic review includes a meta-analysis, make an independent assessment of the rationale for pooling of data.* As mentioned earlier, crosscheck the PICO characteristics, study design, setting and period, and ROB assessment as a pooling of data is not appropriate if significant clinical heterogeneity is present.

Step 6: *Assessment and interpretation of the forest plot- the 10-point checklist*

A critical assessment of the forest plot showing results of the meta-analysis is essential. A 10-point checklist can be provided to make this process less complicated.

(1) *Number of studies, sample sizes of individual studies, and total sample size*

Inclusion of at least a few thousand participants in an RCT is considered essential for the results to have optimal validity and certainty for guiding clinical practice and research in the field (Guyatt et al. 2011). Based on this assumption, the cumulative sample size of the studies included in a meta-analysis should be at least a few thousand. Considering the strengths and weakness of the study design (e.g. RCTs vs. non-RCTs) is also important in judging the ROB affecting the evidence generated by the meta-analysis.

(2) *Check the weightage given to different studies; is any study driving the results? Any outliers?*
(3) *Check the number of events (numerator) and denominators in the intervention versus control group*

Based on the choice of the model selected for meta-analysis, the sample size of the included studies may or may not influence the pooled estimates. The event rates affect the ability of included studies to influence the pooled estimate of the effect under evaluation. In a meta-analysis using the fixed-effect model, the weightage given to individual studies depends on their sample size as well as the event rates (Werre SR et al. 2005; Deeks et al. 2019; Xu et al. 2020). A study may have a large sample size but will not influence the results significantly if the event rate is low. The duration of follow up is thus important in this context. Subject expertise, i.e. knowledge of the normal range of baseline/control group event rate, is essential for interpreting results in this context. An unexpectedly high or low baseline risk indicates the need for exploring the underlying reasons. It is also important to check for outliers, i.e. studies reporting unusual/conflicting results. Exploring the reasons for such significant heterogeneity is important. A post hoc sensitivity analysis can be helpful to judge the influence of outliers on the results (Baker and Jackson 2008). However, caution is required in interpreting the results of such analyses.

(4) *Assessment of heterogeneity: Overlap of confidence intervals*
(5) *Tests for heterogeneity: Chi^2 (Q statistics) and its P-value, I^2: (%)*

Visual inspection of the forest plot to check for overlap of the confidence intervals is a useful method to assess heterogeneity (Mohan and Adler 2019; Viechtbauer 2007; Coulson et al. 2010). As discussed above the potential reasons for the heterogeneity of outliers need to be explored. The results of the tests for heterogeneity need to be checked. Significant p-value (<0.05) of the Chi^2 (Q statistics) test and I^2 > 50% indicate significant heterogeneity that should be explored (Melson et al. 2014; Higgins and Thompson 2002; Higgins et al. 2002; IntHout et al. 2015; Ioannidis 2008; Evangelou et al. 2007; von Hippel 2015; Rücker et al. 2008; Huedo-Medina et al. 2006; Bowden et al. 2011). A meta-analysis is not justified if there is significant clinical heterogeneity. We will discuss the models used for meta-analysis under point 8 of this checklist.

(6) *Pooled effect (Z) size, P-value, and statistical vs. clinical significance*

The size of the diamond reflects the pooled effect size and its boundaries represent its 95% confidence intervals (Lewis and Clarke 2001). It is important to assess not only the effect size but also its 'certainty' (i.e. 'what are the results' and 'how confident/certain we are about them'). Considering the clinical significance (actual treatment effect and its certainty) is more important than focussing only on the "P" values and statistical significance (Ranganathan et al. 2015). A 10% reduction in mortality is more important than a 30% improvement in an outcome that has questionable importance in clinical practice.

(7) *Risk vs. odds ratio (RR vs. OR), absolute risk ratio (ARR) or difference (ARD) and the numbers needed to treat (NNT)*

Correct understanding and interpretation of RR and OR and the clinical significance of ARR and ARD is essential to avoid misinterpretation of results (Balasubramanian et al. 2015). NNT is the reciprocal of the ARD between treatment and control groups in an RCT. It is sensitive to PICO characteristics, setting and other factors that affect the baseline risk. Significant heterogeneity between included trials can result in misinterpretation of NNT in a meta-analysis. Consideration of the baseline risk/severity of illness is vital for optimal interpretation of NNTs (Ebrahim 2001).

(8) *Models used for meta-analysis, and concordance/discordance of results*

A quick recap of the critical assumptions and characteristics of the two models is important at this stage (Nikolakopoulou et al. 2014; Borenstein et al. 2010; Schmidt et al. 2009; Sanchez-Meca and Marin-Martinez 2008; Hunter and Schmidt 2000; Jackson and Turner 2017; Shuster 2010; Stanley and Doucouliagos 2015). The fixed effect model makes a confident assumption that intervention is equally effective across all studies, ignores "*between studies*" variation, and provides the *best estimate of the effect.* It gives weightage to the included studies based on their sample size (size of the square), and event rate. On the other hand, the random-effects model allows for '*within*' as well as '*between-study*' variability in effectiveness based on a conservative assumption. Being less confident, it usually has wider confidence intervals, gives adequate emphasis on smaller studies, and provides the estimated *average effect.* The validity of the results is supported well if the results of the meta-analysis by the two models are similar. Discordance indicates the need for exploring heterogeneity.

Check if the choice of model for meta-analysis (fixed effect vs. random effects model) was appropriate. It is important to appreciate that no two participants in any study are identical. Hence every individual's biologic response to an exposure/intervention is expected to be different even if everything else seems to be comparable. It is not uncommon to see the line differentiating between 'more or less similar', and 'significantly different' get blurred due to pre-existing conscious or subconscious biases. This is the reason the random-effects model is often advocated as the default model for meta-analysis. Others believe that starting with a fixed-effect model is acceptable if the studies are more or less similar. To safeguard from biased results, especially when significant statistical heterogeneity (indicated by the I^2 value) is noted, it is essential to compare the results of meta-analysis using both models.

(9) *The strength of evidence for the pooled estimates*

As discussed earlier, it is vital to check the number of studies as well as their design, sample size, and ROB, contributing to the meta-analysis to derive the pooled estimate of an effect/outcome. The confidence interval helps in assessing the precision of the estimate based on the total sample size available for assessing the outcome of interest. Other elements such as event rates, baseline severity of the underlying condition, setting, duration of follow up, and adverse effects are also crucial for judging the strength and external validity of an intervention.

(10) *Human errors in data extraction, entry, and interpretation*

Last but not least, it is essential to check for errors in sample sizes, event rates (numerator and denominator) from included studies and their correct allocation to the intervention vs. control group. Transposition errors can have severe consequences for results and their interpretation. Check if the selection and interpretation of labels (Favours intervention vs. Favours control) on both sides of the central line of no effect, is correct. Consideration of whether the outcome assessed is beneficial or adverse is important in this context. Standard error can be confused with standard deviation, and a 'minus' sign can be missing or confused with a hyphen!

Step 7: *Assessment of the funnel plot–checking for publication bias*

Publication bias occurs when published studies differ systematically from all conducted studies on a topic (Dickersin 1990). It arises when studies with statistically significant or positive results in a specific direction are more likely to be published compared to those without statistically significant or negative results. Careful visual inspection of the funnel plot is important as publication bias can seriously compromise the validity of systematic reviews (Sedgwick 2015). It is important to note that publication bias can never be ruled out. When it is less likely, the largest studies lie closest to the true value, and the smaller studies are spread on either side, creating the shape of a funnel. Check if the 'funnel' is challenging to visualise or incomplete with an area with missing studies.

Caution is required in the interpretation of a funnel plot as it is affected by many factors, including alternative explanations for the asymmetrical distribution of studies and inaccurate visual interpretation (Lau et al. 2006; Sterne et al. 2011). Potential reasons for funnel plot asymmetry other than publication bias include poor methodological quality leading to exaggerated effects in smaller studies, true heterogeneity, artefacts and chance (Sterne et al. 2011). As a general rule, at least ten studies are required for proper visual assessment of a funnel plot (Lau et al. 2006). Check if the reviewers have reported results of statistical tests for funnel plot asymmetry (publication bias) such as the Egger test, Begg test, and Harbord test. A statistically non-significant P-value for the asymmetry test does not exclude bias.

These tests are known to have low power (Sterne et al. 2011; Jin et al. 2015). The commonly used Egger's test has "inappropriate" type I error rate when heterogeneity is present, and the number of included studies is large (Jin et al. 2015). The Harbord Test has a better error rate compared to Egger's test in balanced trials with little or no heterogeneity (Jin et al. 2015). Considering there is no gold standard test for confirming publication bias, experts have cautioned about the risk of discrediting valid evidence following decisions based solely on asymmetrical funnel plots and positive statistical tests (Lau et al. 2006).

Step 8: *Is the interpretation of results appropriate?*

Check if the reviewers have provided a credible, unbiased, balanced interpretation of the results being as subjective as possible (Shrier et al. 2008; Liberati 1995; Tricco et al. 2011). Industry influence is a matter of concern in the interpretation of results of industry-sponsored studies (Jørgensen et al. 2008). The Cochrane collaboration includes industry involvement as one of the potential reasons for bias. It is hence essential to check for this possibility. Check if the reviewers have put research into context. Ideally, a systematic review should not tell what should be done for an individual patient. That process is left to the healthcare provider and the patient as a shared responsibility. Try not to be biased by the reviewer's conclusions!

Step 9: *Importance of safety as an outcome, and pitfalls related to subgroups, post hoc analyses, multiplicity, and 'trends'.*

Clinical trials often do not address/report safety as an important, or perhaps the most important outcome (Huang et al. 2011; Ioannidis and Lau 2001). Systematic reviews are therefore expected to address this issue. Subgroups in a systematic review and meta-analysis should be evidence-based and prespecified (Richardson et al. 2018; https://wiki.joannabriggs.org/display/MANUAL/3.3.7+Subgroups+in+meta-analysis). They are usually based on unique characteristics of participants (e.g. age, gender, the severity of illness), intervention (e.g. mode of delivery), comparisons (e.g. probiotics with vs without lactoferrin vs placebo) or outcomes in unique groups (e.g. extremely preterm infants). It is important to check if the reviewers have been careful in interpreting the results of subgroups and post hoc analyses. Systematic reviews are also not immune from the problem of multiplicity (Bender et al. 2008). Finally, it is important to avoid getting biased by 'trends' presented as positive findings (Gibbs and Gibbs 2015).

Step 10: *Is the method of reporting appropriate?*

Finally, every systematic review (and meta-analysis) should have a recommended structured format for reporting for the wide dissemination of results with clarity. Some of the standard reporting guidelines include the PRISMA statement (Checklist and a flow diagram) for Preferred Reporting Items for Systematic Reviews and Meta-Analyses, MOOSE guidelines for reporting meta-analysis of observational studies in epidemiology, and STROBE statement for the reporting of observational studies in epidemiology (Stroup et al. 2000; von Elm et al. 2008).

In summary, critical assessment of a systematic review and meta-analysis requires thorough knowledge of their principles, procedures, strengths and limitations and importantly, scientific expertise in the field of investigation. This is critical because their findings are expected to guide clinical practice and research. Systematic reviews and meta-analyses may not be an exact science and hence untrustworthy due to their dependence on the clinical judgement at every step (Haddaway and Rytwinski 2018; Thompson and Pocock 1991). However, when conducted and reported using a rigorous methodology, they remain the best available evidence. As for the validity of results of a meta-analysis of many small studies vs. those of a single adequately powered large RCT, the debate will continue (Glasziou et al. 2010; Scifres et al. 2009; Ioannidis et al. 1998). The fact is that systematic reviews and meta-analyses are critical components of the cycle of knowledge. The least they could do is to help design robust RCTs to know what works (Cooper et al. 2005; Clarke et al. 2010; Mahtani 2016).

References

Baker R, Jackson D. A new approach to outliers in meta-analysis. Health Care Manag Sci. 2008;11:121–31.

Balasubramanian H, Ananthan A, Rao S, Patole S. Odds ratio vs risk ratio in randomised controlled trials. Postgrad Med. 2015;127:359–67.

Bashir R, Surian D, Gunn AG. Time-to-update of systematic reviews relative to the availability of new evidence. Syst Rev. 2018;7:195. https://doi.org/10.1186/s13643-018-0856-9.

Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? PLoS Med. 2010;7: https://doi.org/10.1371/journal.pmed.1000326.

Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, Thorlund K. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol. 2008;61:857–65.

Borenstein M, Hedges LV, Higggins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010;1:97–111.

Bowden J, Tierney JF, Copas AJ, et al. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Med Res Methodol. 2011;11:41. https://doi.org/10.1186/1471-2288-11-41.

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: Article 245. https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-017-0644-y.

Brown PA, Harniss MK, Schomer KG, Feinberg M, Cullen NK, Johnson KL. Conducting systematic evidence reviews: core concepts and lessons learned. Arch Phys Med Rehabil. 2012;93:S177–84.

Chess LE, Gagnier JJ. Applicable or non-applicable: investigations of clinical heterogeneity in systematic reviews. BMC Med Res Methodol. 2016;17(16):19. https://doi.org/10.1186/s12874-016-0121-7.

Clarke M, Hopewell S, Chalmers I. Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. Lancet. 2010;376:20–1.

Cooper NJ, Jones DR, Sutton AJ. The use of systematic reviews when designing studies. Clin Trials. 2005;2:260–4.

Cooper C, Booth A, Varley-Campbell J, Britten N, Garside R. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. BMC Med Res Methodol. 2018;18:85. https://doi.org/10.1186/s12874-018-0545-3.

Coulson M, Healey M, Fidler F, Cumming G. Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. Front Psychol. 2010;1:26. https://doi.org/10.3389/fpsyg.2010.00026. eCollection 2010.

da Costa BR, Juni P. Systematic reviews and meta-analyses of randomised trials: principles and pitfalls. Eur Heart J. 2014;14(35):3336–45.

Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane, 2019. www.training.cochrane.org/handbook.

Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA. 1990;263:1385–9.

EBM notebook: Weighted event rates. Werre SR, Walter-Dilks C. BMJ Evid Based Med. 2005;10:70. http://dx.doi.org/10.1136/ebm.10.3.70.

Ebrahim S. The use of numbers needed to treat derived from systematic reviews and meta-analysis: caveats and pitfalls. Eval Health Prof. 2001;24:152–64.

Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. BMJ. 1997;315 (7121):1533–7.

Esteves SC, Majzoub A, Agarwal A. The problem of mixing 'apples and oranges' in meta-analytic studies. Transl Androl Urol. 2017;6:S412–3. https://doi.org/10.21037/tau.2017.03.23.

Evangelou E, Ioanidis JPA, Patsopoulos NA. Uncertainty in Heterogeneity Estimates in Meta-Analyses. BMJ. 2007;335:914–6.

Fuhr U, Hellmich M. Channelling the flood of meta-analyses. Eur J Clin Pharmacol. 2015;71:645–7.

Gagnier JJ, Moher D, Boon H, Beyene J, Bombardier C. Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. BMC Med Res Methodol. 2012;30(12):111. https://doi.org/10.1186/1471-2288-12-111.

Garner P, Hopewell S, Chandler J, et al. When and how to update systematic reviews: consensus and checklist. BMJ. 2016;354: https://doi.org/10.1136/bmj.i3507.

Gibbs NM, Gibbs SV. Misuse of 'trend' to describe 'almost significant' differences in anaesthesia research. Br J Anaesth. 2015;115:337–9.

Glasziou PP, Shepperd S, Brassey J. Can we rely on the best trial? A comparison of individual trials and systematic reviews. BMC Med Res Methodol. 2010;18(10):23. https://doi.org/10.1186/1471-2288-10-23.

Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence and imprecision. J Clin Epidemiol 2011; 64: 1283e–1293.

Haddaway NR, Rytwinski T. Meta-analysis is not an exact science: Call for guidance on quantitative synthesis decisions. Environ Int. 2018;114:357–9.

Higgins J, Thompson S, Deeks JJ, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. J Health Service Res Policy. 2002; 7:51–61.

Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;15 (21):1539–58.

Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol. 2005;5:13.

Huang HY, Andrews E, Jones J, Skovron ML, Tilson H. Pitfalls in meta-analyses on adverse events reported from clinical trials. Pharmacoepidemiol Drug Saf. 2011;20:1014–20.

Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods. 2006;11:193–206.

Hunter JE, Schmidt FL. Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. Int J Sel Assess. 2000; 8: 275–292.

IntHout J, Ioannidis JP, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. J Clin Epidemiol. 2015;68:860–9.

Ioannidis JP. Interpretation of tests of heterogeneity and bias in meta-analysis. J Eval Clin Pract. 2008;14:951–7.

Ioannidis JA. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q. 2016;94:485–514. https://doi.org/10.1111/1468-0009.12210.

Ioannidis JP, Lau J. Completeness of safety reporting in randomised trials: an evaluation of 7 medical areas. JAMA. 2001;285:437–43.

Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. JAMA. 1998;8(279):1089–93.

Jackson D, Turner R. Power analysis for random-effects meta-analysis. Res Synth Methods. 2017;8:290–302.

Jeong W, Keighley C, Wolfe R, et al. The epidemiology and clinical manifestations of mucormycosis: a systematic review and meta-analysis of case reports. Syst Rev. 2019;25 (1):26–34. https://doi.org/10.1016/j.cmi.2018.07.011.

Jin ZC, Zhou XH, He J. Statistical methods for dealing with publication bias in meta-analysis. Stat Med. 2015;34:343–60.

Jones JB, Blecker S, Shah NR. Meta-analysis 101: what you want to know in the era of comparative effectiveness. Am Health Drug Benefits. 2008;1:38–43.

Jørgensen AW, Maric KL, Tendal B, Faurschou A, Gøtzsche PC. Industry-supported meta-analyses compared with meta-analyses with non-profit or no support: differences in methodological quality and conclusions. BMC Med Res Methodol. 2008;9(8):60. https://doi.org/10.1186/1471-2288-8-60.

Kriston L. Dealing with clinical heterogeneity in meta-analysis. Assumptions, methods, interpretation. Int J Methods Psychiatr Res. 2013;22:1–15.

Lakens D, Hilgard J, Staaks J. On the reproducibility of meta-analyses: six practical recommendations. BMC Psychol. 2016;31(4):24. https://doi.org/10.1186/s40359-016-0126-3.

Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ. 2006;333(7568):597–600.

Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ. 2001;322:1479–80.

Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol. 2009; 62: e1–e34.

Liberati A. Meta-analysis: statistical alchemy for the 21st century: discussion. A plea for a more balanced view of meta-analysis and systematic overviews of the effect of health care interventions. J Clin Epidemiol. 1995;48:81–6.

Mahtani KR. All health researchers should begin their training by preparing at least one systematic review. J R Soc Med. 2016;109:264–8.

Malone DC, Hines LE, Graff JS. The good, the bad, and the different: a primer on aspects of heterogeneity of treatment effects. J Manage Care Special Pharm. 2014;20:555–63.

Melson WG, Bootsma MCJ, Rovers MM, Bonten MJM. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. Clin Microbiol Infec. 2014;20:123–9.

Mohan BP, Adler DG. Heterogeneity in systematic review and meta-analysis: how to read between the numbers. Gastrointest Endosc. 2019;89:902–3.

Nikolakopoulou A, Mavridis D, Salanti G. How to interpret meta-analysis models: fixed effect and random effects meta-analyses. Evid Based Mental Health. 2014. https://doi.org/10.1136/eb-2014-101794.

Phan K, Tian DH, Cao C, Black D, Yan TD. Systematic review and meta-analysis: techniques and a guide for the academic surgeon. Ann Cardiothorac Surg. 2015;4:112–22.

Purgato M, Adams CE. Heterogeneity: the issue of apples, oranges and fruit pie. Epidemiol Psychiatr Sci. 2012;21:27–9.

Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: clinical versus statistical significance. Perspect Clin Res. 2015;6:169–70. https://doi.org/10.4103/2229-3485.159943.

Richardson M, Garner P, Donegan S. Interpretation of subgroup analyses in systematic reviews: a tutorial. Clin Epidemiol Global Health. 2018 (Published: May 28, 2018) https://doi.org/10.1016/j.cegh.2018.05.005.

Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995; 123:A12–3.

Roever L, Zoccai GB. Critical appraisal of systematic reviews and meta-analyses. Evid Based Med Pract. 2015;1:1. https://doi.org/10.4172/EBMP.1000e106.

Rücker G, Schwarzer G, Carpenter JR, et al. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8:79.

Sanchez-Meca J, Marin-Martinez F. Confidence intervals for the overall effect size in random-effects meta-analysis. Psychol Methods. 2008;13:31–48.

Schmidt FL, Oh IS, Hayes TL. Fixed-versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. Br J Math Stat Psychol. 2009;62:97–128.

Scifres CM, Iams JD, Klebanoff M, Macones GA. Meta-analysis vs large clinical trials: which should guide our management? Am J Obstet Gynecol. 2009;200:484.e1–4845.

Sedgwick P. How to read a funnel plot in a meta-analysis. BMJ. 2015;351: https://doi.org/10.1136/bmj.h4718 (Published 16 September 2015).

Shokraneh F. Reproducibility and replicability of systematic reviews. World J Meta-Anal. 2019; 7 (3): 66–71. https://dx.doi.org/10.13105/wjma.v7.i3.66.

Shrier I, Boivin JF, Platt RW, et al. The interpretation of systematic reviews with meta-analyses: an objective or subjective process? BMC Med Inform Decis Making. 2008;19. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472–6947-8-19.

Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. Stat Med. 2010;30 (29):1259–65.

Simonsohn U, Nelson LD, Simmons JP. P-curve: A key to the file-drawer. J Exp Psychol Gen. 2014;143:534–47.

Smith TO, Hing CB. "Garbage in, garbage out"- the importance of detailing methodological reasoning in orthopaedic meta-analysis. Int Orthop. 2011;35:301–2.

Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. Stat Med. 2015;15(34):2116–27.

Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;22(343): https://doi.org/10.1136/bmj.d4002.

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;12(355): https://doi.org/10.1136/bmj.i4919.

Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting: Meta-analysis of observational studies in epidemiology (MOOSE) group. JAMA. 2000; 283: 2008–2012.

Subgroups in meta-analysis–Section 3.3.7: JBI Reviewer's Manual–JBI GLOBAL WIKI https://wiki.joannabriggs.org/display/MANUAL/3.3.7+Subgroups+in+meta-analysis.

Thompson SG, Pocock SJ. Can meta-analyses be trusted? Lancet. 1991;2(338):1127–30.

Tricco AC, Straus SE, Moher D. How can we improve the interpretation of systematic reviews? BMC Med. 2011: 31. https://bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-9-31.

van der Knaap LM, Leeuw FL, Bogaerts S, Nijssen LTJ. Combining Campbell standard and the realist evaluation approach: the best of two worlds? Am J Eval. 2008;29:48–57.

van Driel ML, De Sutter A, De Maeseneer J, Christiaens T. Searching for unpublished trials in Cochrane reviews may not be worth the effort. J Clin Epidemiol. 2009;62(838–44): https://doi.org/10.1016/j.jclinepi.2008.09.010.

Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. Stat Med. 2007;26:37–52.

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61:344–9.

von Hippel PT. The heterogeneity statistic I(2) can be biased in small meta-analyses. BMC Med Res Methodol. 2015;14(15):35.

Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of non-randomised studies in meta-analyses. The Ottawa Hospital Research Institute; 2000. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

Xu C, Li L, Lin L, et al. Exclusion of studies with no events in both arms in meta-analysis impacted the conclusions. J Clin Epidemiol. 2020; 123: 91–99.

Systematic Reviews and Meta-Analyses of Non-randomised Studies

Sanjay Patole

Abstract Randomised controlled trials (RCTs) are considered as the gold standard for clinical research because unlike other study designs, they control for known, and importantly, unknown confounders by randomisation. Evaluation of interventions should hence be ideally done by RCTs. However, RCTs are not always possible or feasible for various reasons, including ethical concerns and the need for time, effort, and funding. Difficulty in the generalisation of the findings of RCTs is also an issue given their rigid design. Non-randomised studies (non-RCTs) provide an alternative to RCTs in such situations. These include cohort, case-control and cross-sectional studies. Non-RCTs have the advantage of providing data from the real-life situation rather than that from the rigid framework of RCTs. The limitations of non-RCTs include selection bias and lack of randomisation that allow confounders to influence the results. At best, non-RCTs can only generate hypotheses for testing in RCTs. This chapter covers the methodology for conducting, reporting and interpreting systematic reviews and meta-analysis of non-RCTs.

Keywords Confounding · MOOSE guidelines · New castle ottawa scale · Non-randomised studies · Randomised controlled trials · Risk of bias · ROBINS-1 tool

Introduction

Randomised controlled trials (RCTs) are considered as the gold standard for clinical research because unlike other study designs, they control for known, and importantly, unknown confounders by randomisation. Allocation concealment protects randomisation. The core elements of the RCT (randomisation, allocation conceal-

S. Patole (✉)
School of Medicine, University of Western Australia, Perth, WA 6009, Australia
e-mail: sanjay.patole@health.wa.gov.au

S. Patole
Neonatal Directorate, King Edward Memorial Hospital for Women,
Perth, WA 6008, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_13

ment and blinding) minimise bias and optimise the internal validity of the results. Evaluation of interventions should hence be ideally done by RCTs. However, RCTs are not always possible or feasible for various reasons, including ethical issues, and importantly, the need for time, effort, and funding. Definitive trials particularly need significant resources considering their large sample sizes, complexity, logistics and the need for expertise in various aspects of the trial. Difficulty in generalisation (i.e. external validity) of the findings of RCTs with rigid designs is also an issue. Non-randomised studies (non-RCTs) provide an alternative to RCTs in such situations (Mariani and Pego-Fernandes 2014; Gershon et al. 2018; Gilmartin Thomas and Liew 2018; Heikinheimo et al. 2017; Ligthelm et al. 2007; Jepsen et al. 2004). These include cohort (Prospective or retrospective), case-control and cross-sectional studies. Non-RCTs have the advantage of providing data from the real-life situation rather than that from the rigid framework of RCTs.

Cohort studies allow estimation of the relative risk as well as the incidence and natural history of the condition under study. They can differentiate cause from an effect as they measure events in temporal sequence. When designed well, adequately powered prospective cohort studies provide the second-best option after RCT (Mann 2003). Both designs include two groups of participants and assess desired outcomes after exposure to intervention over a specified time in a setting (The PICOS approach). However, the critical difference is that unlike the RCT, the two groups (exposed vs not exposed) are not selected randomly in a cohort study. Retrospective cohorts are quick and cheaper to conduct, but the validity of their results is questionable considering the unreliable and often, inadequate retrospective data.

Unlike cohort studies that can assess common conditions and common exposures, case-control studies help in studying rare conditions/diseases and rare exposures (e.g. lung cancer after asbestos exposure). To put it simply, case-control studies assess the frequency of exposure in those with vs those without the condition/disease of interest. If the frequency of exposure is higher in those with the condition of interest than those without the condition; thus establishing an ‘association’. Hill’s criteria for associations are important in this context. Case-control studies estimate odds ratios (OR) rather than relative risk (RR). The difficulties in matching control groups for known confounders and a higher risk of bias are limitations of case-control studies. Cross-sectional studies are also relatively quick and cheap, can be used to estimate prevalence, and study multiple outcomes. However, they also cannot differentiate between cause and effect.

Overall, the major limitations of non-randomised studies include selection bias and lack of randomisation that allow confounders to influence the results (Gueyffier and Cucherat 2019; Gerstein et al. 2019). A confounder is any factor related to the intervention as well as the outcome and could affect both. Therefore, at best, non-randomised studies can only generate hypotheses that need to be tested in RCTs. They are useful for identifying associations that can then be more rigorously studied using a cohort study or ideally in an RCT. One of the commonly used statistical tools to address the issue of confounding is regression analysis which ‘adjusts/controls’ the results for known confounders. This is the reason why access to both, unadjusted as well as adjusted results (e.g. ORs), is important for interpreting the results of non-RCTs. Other techniques such as propensity scores and sensitivity analysis can reduce bias caused by the lack of randomisation in non-RCTs (Joffe and Rosenbaum 1999). Non-RCTs are known to overestimate the effects of an intervention. However, adequately powered, and well designed and conducted non-RCTs can provide effects estimates that are relatively close to those provided by RCT (Concato et al. 2000).

Despite their limitations, non-RCTs have a substantial and well-defined role in evidence-based practice. They are a crucial part of the knowledge cycle and complement RCTs (Faraoni and Schaefer 2016; Schillaci et al. 2013; Norris et al. 2010). Systematic reviews and meta-analyses of non-RCTs are hence common in all faculties of medicine. This section briefly covers the critical aspects of the process of systematic review and meta-analysis of non-RCTs compared with RCTs.

Conducting a Systematic Review of Non-RCTs

The initial steps in conducting a systematic review of non-RCTs are similar to those for a systematic review of RCTs. These include framing a clinically useful and answerable question using the PICO approach, deciding the type of studies to be searched (e.g. non-RCTs of an intervention), and conducting a comprehensive literature search for the best available evidence. The search is much broader compared to that for RCTs given the different study designs that come under the term "non-RCTs". To avoid wastage of resources and duplication, it is essential to check whether the question has already been answered.

The search strategy includes the following terms for the publication type: *observational, cohort, case-control, cross-sectional studies, retrospective, prospective studies, non-randomised controlled trial*. Searching major databases, grey literature, proceedings of the relevant conference proceedings, registries, checking cross-references of important publications including reviews, and contacting experts in the field is as important as in any other systematic review.

Having a team of subject experts and methodologists optimises the validity of the results. A transparent and unbiased approach, and use robust methods, and explicit criteria are critical to assure that the review is 'truly' systematic (**T**ransparent, **R**obust, **R**eproducible, **U**nbiased, **E**xplicit).

The Cochrane methodology and MOOSE guidelines (Meta-analysis of Observational Studies in Epidemiology) are commonly followed for conducting and reporting systematic reviews of non-RCTs (Lefebvre et al. 2008; Stroup et al. 2000; Lefebvre et al. 2013).

Data Extraction

Data extraction is done independently by at least two reviewers, using the data collection form designed for the review. For dichotomous outcomes, the number of participants with the event and the number analysed in each intervention group of each study are recorded. Availability of these data helps in creating forest plots of unadjusted ORs.

For continuous outcomes, the mean and standard deviation are entered. Authors of the included studies may need to be contacted to verify the study design and outcomes. The mean and standard deviation could be derived from median and range and from median and interquartile range by using the Hozo and Wan formula respectively (Gueyffier and Cucherat 2019; Hozo et al. 2005; Wan et al. 2014).

Assessment of Risk of Bias in Non-RCTs

The key difference between RCTs vs non-RCTs is the risk of bias due to confounding in the later. Assessment of the risk of bias is hence a critical step in systematic reviews of non-RCTs. The standard tools for this purpose are discussed briefly below.

(1) *The Newcastle Ottawa Scale (NOS)*

The Newcastle-Ottawa Scale (NOS) was developed by a collaboration between the University of Newcastle, Australia, and the University of Ottawa, Canada, to assess the quality of non-randomised studies (http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp).

The NOS scale contains three major domains: a selection of subjects, comparability between groups and outcome measures. The maximum score for each domain is four, two and three points, respectively. Thus, the maximum possible score for each study is 9. A total score ≤ 3 indicates low methodological quality, i.e. high risk of bias.

The NOS is a validated and an easy and convenient tool for assessing the quality of non-RCTs included in a systematic review. It can be used for cohort and case-control studies. A modified version can be used for prevalence studies. The scale has been refined based on the experience of using it in several projects. Because it gives a score between 0 and 9, it is possible to use NOS as a potential moderator in meta-regression analyses (Luchini et al. 2017; Wells et al. 2012; Veronese et al. 2016). The NOS is not without limitations. These include some of the domains that are not univocal, difficulties in adapting it to case-control and cross-sectional studies and the low agreement between two independent reviewers in scoring using NOS (Hartling et al. 2013). Training and expertise are essential for proper use of NOS (Oremus et al. 2012).

(2) *ROBINS-1 tool*

The NOS scale and the Downs-Black checklist are commonly used for assessing the risk of bias in non-RCTs. However, both include items relating to external and internal validity (Downs and Black 1998). Furthermore, lack of comprehensive manuals increases the risk of differences in interpretation by different users (Deeks et al. 2003). The ROBINS-I (“Risk Of Bias In Non-randomised Studies—of

Interventions”), is a new tool for evaluating the risk of bias in non-RCTs (Sterne et al. 2016).

Briefly, the ROBINS-1 tool considers each study as an attempt to mimic a hypothetical pragmatic RCT and covers seven distinct domains through which bias might be introduced. It uses ‘signalling questions’ to help in judging the risk of bias within each domain. The judgements within each domain carry forward to an overall risk of bias judgement across bias domains for the outcome being assessed. For details, the readers are referred to the publication by Sterne et al. (2016).

Data Synthesis

The random effects (REM) model is preferred for meta-analysis assuming heterogeneity. A categorical measure of effect size is expressed as the odds ratio (Mantel Haenszel method). Statistical heterogeneity is assessed by Chi-Squared test, I^2 statistic, and visual inspection of the forest plot (overlap of confidence intervals). The validity of REM results can be crosschecked by comparing them with the fixed-effect model (FEM) meta-analysis. Comparability of results by both models is reassuring.

While conducting meta-analysis of non-RCTs, it is important to pool adjusted and unadjusted effect size estimates separately. Pooled adjusted values must be given more importance to minimise the influence of confounders. It is important to note the type of confounders adjusted for in different studies. When synthesising results, consideration of the risk of bias in included studies is more important than the hierarchy of study design.

Publication bias: This is assessed by a funnel plot unless the number of studies is <10. Statistical tests are used if required, but their limitations need to be taken into account. It is important to note that there is no gold standard against which the funnel plot test results can be compared (Lau et al. 2006). Publication bias is not the only reason for an asymmetrical funnel plot. True heterogeneity also contributes to the small study effect (Lau et al. 2006).

Summary of findings: The data on quality of evidence, the magnitude of intervention effect, and the sum of available data on main outcomes are presented in the ‘Summary of findings table’ as per GRADE (Grading of Recommendations Assessment, Development and Evaluation) guidelines (Guyatt et al. 2013). To start with, the evidence is graded as ‘low’ given the limitations of the design of non-RCTs. It could then be upgraded based on the effect size, dose-response, and effect of all plausible confounding factors.

Important Issues in Presentation and Interpretation of Results

Understanding the properties of odds ratios (McHugh 2009; Szumilas 2010; Bland and Altman 2000; Cummings 2009; Balasubramanian et al. 2015) compared with risk ratios, the significance of unadjusted vs adjusted results, and caveats of different study designs (e.g. cohort vs case-control) is critical in presenting and interpreting the results of systematic reviews and meta-analysis of non-RCTs. It is important to note the type and number of confounders adjusted for in the included studies. Subject expertise is essential in this context. If possible, it is preferable to contact the authors of the included study for individual participant data to conduct analyses controlling for confounders. It is not unusual for the pooled effect estimates to differ based on the design of the non-RCTs. For example, pooled estimates from cohort studies have shown that red cell transfusions were associated with a lower risk of transfusion-associated necrotising enterocolitis (TA-NEC) in preterm infants. In contrast, those from case-control studies showed no association of TA-NEC with red cell transfusions (Saroha et al. 2019).

Evidence from non-RCTs, considering their higher risk of bias, can only be used to generate hypotheses to be tested in RCTs. However, when there are no RCTs in the field of interest, non-RCTs can provide the 'best available' evidence for decision making. The current focus of the Cochrane collaboration on systematic reviews of non-RCTs supports this philosophy (Reeves et al. 2019).

Critical Appraisal of Systematic Reviews of Non-rCTs

AMSTAR 2 is a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions or both (Shea et al. 2017).

In summary, systematic reviews of non-RCTs are an essential part of the evidence in totality, considering RCTs may not always be available or possible for various reasons. Suppose a comprehensive literature search reveals no RCTs. In that case, a systematic review of non-RCTs is justified as long as they directly address the framed question (PICOS), and are well designed, and conducted with minimal risk of bias (Faber et al. 2016). Whether systematic reviews of non-RCTs overestimate or underestimate the effects of the intervention compared to RCTs, continues to be a controversial issue (Abrahama et al. 2010).

References

Abrahama NS, Byrneb 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;63:238–45. https://doi.org/10.1016/j.jclinepi.2009.04.005.

Balasubramanian H, Ananthan A, Rao S, Patole S. Odds ratio vs risk ratio in randomised controlled trials. Postgrad Med. 2015;127(4):359–67.

Bland JM, Altman DG. The odds ratio. BMJ. 27 May 2000; 320: 1468.

Szumilas M. Explaining odds ratios. Can Acad Child Adolesc Psychiatry. 2010; 19(3): 227–229.

Concato J, Shah N, Horwitz RI. Randomised, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887–92.

Cummings P. The relative merits of risk ratios and odds ratios. Arch Pediatr Adolesc Med. 2009;163(5):438–45.

Deeks JJ, Dinnes J, D'Amico R, et al. International Stroke Trial Collaborative Group European Carotid Surgery Trial Collaborative Group. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7:iii–x, 1–173.

Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health. 1998;52:377–84.

Faber T, Ravaud P, Riveros C, Perrodeau E, Dechartres A. Meta-analyses including non-randomized studies of therapeutic interventions: a methodological review. BMC Med Res Methodol. 2016;16:35. https://doi.org/10.1186/s12874-016-0136-0.

Faraoni D, Schaefer ST. Randomised controlled trials vs. observational studies: why not just live together? BMC Anesthesiol. 2016 Oct 21;16(1):102.

Gershon AS, Jafarzadeh SR, Wilson KC, Walkey A. Clinical knowledge from observational studies: everything you wanted to know but were afraid to ask. Am J Respir Crit Care Med. 2018; 198 (7):859–867.

Gerstein HC, McMurray J, Holman RR. Real-world studies no substitute for RCTs in establishing efficacy. Lancet. 2019;393:210–1.

Gilmartin Thomas JFM, Liew D. Observational studies and their utility for practice. Aust Prescr. 2018;41:82–5.

Gueyffier F, Cucherat M. The limitations of observation studies for decision making regarding drugs efficacy and safety. Therapie. 2019;74:181–5.

Guyatt GH, Oxman AD, Santesso N, et al. GRADE guidelines: 12 Preparing summary of findings tables—binary outcomes. J Clin Epidemiol 2013;66:158–172.

Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, Dryden DM. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol. 2013;66:982–93.

Heikinheimo O, Bitzer J, Rodríguez LG. Real-world research and the role of observational data in the field of gynaecology–a practical review. Eur J Contracept Reprod Health Care. 2017;22 (4):250–9.

Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol. 2005;5:13.

Jepsen P, Johnsen SP, Gillman MW, Sorensen HT. Interpretation of observational studies. Heart. 2004;90(8):956–60.

Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 1999;150 (4):327–33.

Lau J, Ioannidis JP, Terrin N, et al. The case of the misleading funnel plot. BMJ. 2006;333:597–600.

Lefebvre C, Manheimer E, Glanville J. Searching for studies. Cochrane handbook for systematic reviews of interventions. New York: Wiley 2008:95–150.

Lefebvre C, Glanville J, Wieland LS, et al. Methodological developments in searching for studies for systematic reviews: past, present and future? Syst Rev. 2013;2:78.

Ligthelm RJ, Borzi V, Gumprecht J, Kawamori R, Wenying Y, Valensi P. Importance of observational studies in clinical practice. Clin Ther. 2007;29 Spec No:1284–92.

Luchini C, Stubbs B, Solmi M, Veronese N. Assessing the quality of studies in meta-analyses: Advantages and limitations of the Newcastle Ottawa Scale. World J Meta-Anal. Aug 26, 2017; 5(4): 80–84. Published online Aug 26, 2017. https://doi.org/10.13105/wjma.v5.i4.80.

Mann CJ. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. Emerg Med J. 2003;20(1):54–60.

Mariani AW, Pego-Fernandes PM. Observational studies: why are they so important? Sao Paulo Med J. 2014;132(1):01–02 https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-31802014000100001&lng=en&tlng=en. Accessed 10 Aug 2020.

McHugh ML. The odds ratio: calculation, usage and interpretation. Biochemic Med. 2009;19 (2):120–126.

Norris S, Atkins D, Bruening W, et al. Selecting observational studies for comparing medical interventions. In: Agency for Healthcare Research and Quality. Methods Guide for Comparative Effectiveness Reviews [posted June 2010]. Rockville, MD. http://www.effectivehealthcare.ahrq.gov/ehc/products/196/454/MethodsGuideNorris_06042010.pdf. Accessed 11 Aug 2020.

Oremus M, Oremus C, Hall GB, McKinnon MC; ECT & Cognition Systematic Review Team. Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa Scales. BMJ Open. 2012;2:e001368.

Reeves BC, Deeks JJ, Higgins JPT, et al. Chapter 24: Including non-randomised studies on intervention effects. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane, 2019. www.training.cochrane.org/handbook. Accessed 10th Aug 2020.

Saroha V, Josephson CD, Patel RM. Epidemiology of necrotising enterocolitis: New considerations regarding the influence of red blood cell transfusions and anemia. Clin Perinatol. 2019;46(1):101–17. https://doi.org/10.1016/j.clp.2018.09.006.

Schillaci G, Battista F, Pucci G. Are observational studies more informative than randomised controlled trials in hypertension? ConSide of the Argument. Hypertension. 2013;62:470–6.

Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358: https://doi.org/10.1136/bmj.j4008 (Published 21/9/2017).

Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA. 2000;283:2008–12.

Veronese N, Carraro S, Bano G, Trevisan C, Solmi M, Luchini C, Manzato E, Caccialanza R, Sergi G, Nicetto D. Hyperuricemia protects against low bone mineral density, osteoporosis and fractures: a systematic review and meta-analysis. Eur J Clin Invest. 2016;46:920–30.

Wan X, Wang W, Liu J, et al. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol. 2014;14:135.

Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality if non-randomised studies in meta-analyses, 2012. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

Individual Participant Data (IPD) Meta-Analysis

Abhijeet Rakshasbhuvankar

Abstract Individual participant data (IPD) meta-analyses are relatively new compared with the traditional aggregate data meta-analyses. IPD meta-analyses involve the collection, checking, and re-analysis of the original data for each participant in each study. IPD meta-analyses have many advantages over traditional meta-analyses using aggregate data such as better assessment of the integrity of studies, and the ability to perform additional analyses at the participant level. Results of IPD meta-analyses can substantially differ from aggregate data meta-analyses. Systematic reviews and meta-analyses based on IPD are considered as the gold standard. However, there are multiple challenges in the way to perform IPD meta-analyses, such as obtaining individual participant data from the studies and a considerable amount of time and financial requirements. This chapter covers the differences in individual participant and aggregate data and the steps, models, advantages and challenges involved in IPD meta-analysis.

Keywords IPD meta-analysis · Individual participant data · Patient-level data · Aggregate data · Data sharing · One-stage model · Two-stage model

Background

Meta-analysis is a statistical combination of results from two or more separate studies (Deeks and Altman 2019). Meta-analyses are commonly based on aggregate data extracted from published results or obtained from investigators. Hence they are also called aggregate data (AD) meta-analyses. Individual participant data (IPD) meta-analyses involve the collection, checking, and re-analysis of the original data for each participant in each study (Tierney and Clarke 2019). Systematic reviews and meta-analyses based on IPD are considered as the gold standard (Riley

A. Rakshasbhuvankar (✉)
School of Medicine, Neonatal Directorate, King Edward Memorial Hospital for Women, University of Western Australia, Perth, WA 6008, Australia
e-mail: Abhijeet.rakshasbhuvankar@health.wa.gov.au

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_14

et al. 2010; Stewart and Parmar 1993). IPD meta-analysis should ideally include data from all studies identified by a thorough systematic literature search as a part of a systematic review. However, IPD may be performed occasionally in non-systematic reviews by including studies in which data is readily available without performing a systematic search (de Weerd et al. 2010) or collaborative reviews where data only from collaborative studies are included (Askie et al. 2018). Apart from interventional studies, IPD meta-analyses can also be performed for diagnostic and prognostic prediction modelling studies (Debray et al. 2015).

IPD meta-analyses are relatively an innovation compared with AD meta-analyses. The first AD meta-analysis in the medical field was performed by Karl Pearson in 1904 to assess the effectiveness of inoculation for enteric fever using data from six observational studies (Report on Certain Enteric Fever Inoculation Statistics 1904; O'Rourke 2007). IPD meta-analyses first appeared in the late 1980s following the development of collaborative trial groups (e.g., Early Breast Cancer Trialists' Collaborative Group, 1987). The number of IPD meta-analyses has increased over the last two decades. However, it remains only a small percentage of all meta-analyses performed (Riley et al. 2010; Simmonds et al. 2015; Huang et al. 2014). Experts have emphasised the need for more IPD meta-analyses (Oxman et al. 1995).

Individual Participant Data (IPD)

IPD refers to the raw data recorded for each participant and includes the patient characteristics and effects of an intervention. On the other hand, AD represents a summary of IPD in the form of average patient characteristics and estimates of the intervention effect (Tierney et al. 2015). IPD is the source of AD (Riley et al. 2010). AD is generally published in the reports as they are easy to interpret and require less space than IPD. As IPD is not published and not peer-reviewed, it may be more liable to error and bias than the published AD, which is peer-reviewed (Chalmers 1987). Hence, checking the data and trial protocol of the studies is critical while dealing with IPD (Stewart and Parmar 1993).

The process of data collection for IPD review can be retrospective or prospective (Kawahara et al. 2018). In the retrospective process, the already collected IPD data is sought either by contacting investigators of the studies or through a repository (Tudur Smith et al. 2014). Such data collection needs collaboration between reviewers and study investigators. Inviting study investigators as co-authors for meta-analysis may help the process. In the prospective data collection process, the data is collected prospectively through a collaborative research group established by the researchers in the field (e.g., Early Breast Cancer Trialists' Collaborative Group) (Riley et al. 2010).

Limitations of AD Meta-Analyses

The availability of only summary statistics (AD) limits the possible analyses and may reduce the power in AD meta-analysis. Moreover, the availability and quality of such data may vary across studies, and reporting of data is often influenced by publication and reporting bias (Riley et al. 2010), affecting the reliability of AD meta-analysis (Tierney et al. 2015). AD meta-analyses are prone to ecological bias, which results when observed across-study relationships do not accurately reflect the individual-level relationship within a trial (Hua et al. 2017). AD meta-analyses are inadequate to study interactions between covariates and treatment effects as they investigate across-studies interactions between aggregated treatment effects and covariates at study levels without due consideration of within-study interactions at the individual level (Hua et al. 2017).

Advantages of IPD Meta-Analysis

IPD meta-analyses have many advantages over AD meta-analyses. Some advantages are inherent to IPD analysis, while others are a byproduct of the time and effort devoted to the collaboration between researchers required for IPD meta-analysis.

1. Less reporting bias: IPD meta-analysis does not rely on published information but includes all available trial data (Stewart and Parmar 1993). Besides, the availability of the raw data helps in analysing outcomes that are not reported in the published articles because of space limitations, perceived less relevance, or statistical insignificance.
2. Better assessment of the integrity of trials: For example, adequacy of randomisation can be assessed by comparing randomisation protocol with the order of recruited patients in IPD. The availability of the raw data enables thorough data checks to identify any inaccuracies and errors and ensures the appropriateness of analyses.
3. Improved consistency across trials: IPD enables the use of standard definitions for patient characteristics (e.g., age groups, eligibility criteria, comorbidities), interventions (e.g., the specific dose from a range of drug doses) or outcomes (e.g., cut-off points, criteria for positivity) (Lyman and Kuderer 2005).
4. Enables additional analyses: IPD helps in addressing questions that are not addressed in the original publication, studying interactions between covariates and treatment effects, adjusting for the same variables across studies, exploring heterogeneity at the patient level, subgroup analyses of patient-level data, and survival and other time-to-event analyses (Davey Smith et al. 1997).
5. Enables updating outcome-related data: For example, follow-up data or time to event data to the latest one if it has been collected. Therefore, IPD meta-analyses are very useful in reviews addressing long-term outcomes.

6. Encourages collaboration between researchers: The cooperation of multiple researchers in IPD review may help in complete identification of relevant trials, better compliance with providing missing data, more balanced interpretation of the results, broader endorsement and dissemination of the results, better clarification of the further research, and collaboration on further research (Oxman et al. 1995; Tierney et al. 2015).

Methods

Table 1 shows the important steps while undertaking an IPD systematic review and meta-analysis (Tierney and Clarke 2019). The statistical methods for IPD analysis can be complicated, require advanced statistical software, and are less well developed as compared with "conventional" AD meta-analysis.

It is important to note that IPD from different studies cannot be pooled together as if it was derived from a single large RCT. We must account for the clustering resulting from the data derived from different studies using either One-stage or Two-stage model (Simmonds et al. 2005).

1. **The two-stage model**: It resembles AD meta-analysis, and involves the pooling of data within an individual study to derive aggregate data in the first stage followed by meta-analysis to derive effect estimate in the second stage. In the two-stage model, while investigating the association between covariates and treatment effect, ecological or aggregation bias should be avoided by estimating within-trial association and then pooling the association estimates across trials using conventional meta-analysis (Stewart et al. 2012). The two-stage model is less complicated, easier to interpret, and enables the generation of forest plots. However, it is less efficient for studying interactions between covariates and treatment effects, especially in the presence of small trials and clinical heterogeneity. It may be adequate for IPD meta-analysis involving large and homogenous trials.
2. **The one-stage model**: It is a regression analysis stratified for studies to estimate the intervention effect. It improves power to detect interactions between covariates and treatment effects (Lambert et al. 2002; Simmonds and Higgins 2007). It is flexible as it allows the inclusion of multiple covariates in a single model and avoids ecological bias. However, it is more complex, requires a higher degree of statistical expertise, increases the potential for data dredging, does not generate forest plots, and challenging to interpret (Stewart et al. 2012; Turner et al. 2000; Higgins et al. 2001). It is essential when an IPD meta-analysis contains small trials and clinically heterogeneous populations (Stewart et al. 2012).

Table 1 Checklist when undertaking a new IPD review (Tierney and Clarke 2019)

Step	**Details**
Decide if IPD review is appropriate for the topic.	• Aggregate data does not permit good quality review • The aim is to explore subpopulation • The aim is to optimise the analysis of time-to-event outcomes
Assess if IPD review is possible.	• Sufficient IPD available • Sufficient time, resources, skills and expertise available
Collect IPD	• Contact and establish rapport with authors, encourage them to join as co-authors • Generic data sharing platforms: e.g. Clinical Study Data Request • International collaborative clinical trials • Data from a topic-based repository, e.g. Early Breast Cancer Trialists' Group • Ethics approval is generally not required if reviewers are addressing the same question as to the original investigators • Obtain sufficient data
Data management	• Data may need to be redefined or recoded to make it homogenous to allow pooling • Check for completeness and integrity of data • Check for risk of bias in included studies • Check data for appropriateness of randomisation, allocation sequence concealment, and attrition
Analysis	• Important to account for clustering of participants in an IPD (a) Two-stage model (b) One-stage model (c) Combination of One and Two-stage models • Exploring the effect of trial and participant characteristics: This must be limited in numbers, prespecified in the protocol and based on biological plausibility (a) Subgroup analysis (b) Meta-regression
Reporting	• PRISMA-IPD guideline

Reporting

Optimum reporting of systematic reviews is essential to make it easier to understand, and critique, and for implementing the findings. Standardised reporting checklist for AD meta-analyses was first developed in 1996 (QUOROM: QUality Of Reporting Of Meta-analyses) (Moher et al. 1999). It was revised in 2009 and included systematic reviews (PRISMA: Prefered Reporting Items for Systematic Reviews and Meta-Analyses) (Moher et al. 2009). Subsequently, the PRISMA guidelines were modified for systematic reviews and meta-analyses of IPD to allow the reviewers to address important IPD-specific issues: e.g., whether eligibility criteria were applied at study level or individual level (item 6), how IPD were requested and collected, what information was sought and what could or could not be obtained and from how many studies (item 10), details of data integration (item A1) and risk of bias checking (item 12), data synthesis model used (item 14), methods used to study participant-level characteristics and whether these were prespecified (item A2), results with inclusion and exclusion of studies for which IPD were not available (item 23) (Stewart et al. 2015).

Results of IPD Versus AD Meta-Analysis

IPD and AD meta-analyses results correlate in the majority of cases (Tudur Smith et al. 2016; Huang et al. 2016). However, they may also defer substantially. The Cochrane review comparing meta-analyses of RCTs based on IPD versus AD showed disagreement in the statistical significance in 20% of the comparisons. IPD comparisons were more likely to yield significant results than those based on AD because of more number of participants and longer length of follow-up in IPD meta-analyses (Tudur Smith et al. 2016). The disagreement may be at the level of significance or the direction of effect. For example, in meta-analyses comparing laparoscopic versus open hernia repair, AD meta-analysis showed persistent pain to be more common in the laparoscopic group (OR: 2.03; 95% CI: 1.03 to 4.01, 3 trials) while IPD meta-analysis showed persistent pain to be less common in the laparoscopic group (OR: 0.54, 95% CI: 0.46 to 0.64, 20 trials) (Collaboration 2000; McCormack et al. 2004, 2003). The disagreement between IPD and AD meta-analyses may happen even when analyses are based on identical trials and participants (Tudur Smith et al. 2016). In a comparison of AD and IPD meta-analyses the effect of paternal cell immunisation for preventing recurrent miscarriages leading to live birth was significant in AD meta-analysis (RR: 1.29; 95% CI 1.03 to 1.60) but insignificant in IPD meta-analysis of the same studies (RR: 1.17; 95% CI: 0.97 to1.37) (Jeng et al. 1995).

Limitations and Challenges

Obtaining IPD from eligible studies is one of the most challenging steps in the process of IPD meta-analysis. IPD may not be obtained because of various reasons, including difficulties in contacting authors, or their unwillingness to share data or loss of the data (Clarke 2005). Unavailability of a small proportion of IPD (probably < 5–10%) may not need any additional analyses (Rogozińska et al. 2017). However, if a significant proportion of IPD is not available, AD may be included in the meta-analysis, and sensitivity analysis excluding AD should be performed to test the robustness of the results. In a systematic review of 760 published IPD meta-analyses, only 25% of the IPD meta-analyses could retrieve 100% IPD from the eligible studies (Nevitt et al. 2017). To improve access to IPD, the International

Committee of Medical Journal Editors (ICMJE) mandated all clinical trials that start recruitment on or after 1st January 2019 to include a data sharing plan in trial's registration as a condition for consideration for publication of the trial's report in member journals (Taichman et al. 2017).

IPD allows exploring differences in the treatment effects based on the subgroup of patients. However, caution is required in interpreting the results of subgroup analyses if they were not prespecified, many subgroups were tested, the difference was suggested by comparisons *between* rather than *within* the studies, the difference was not consistent across studies, and no indirect evidence that supports the hypothesised difference (Oxman and Guyatt 1992; Yusuf et al. 1991). Such subgroup analyses may lead to erroneous conclusions because of bias (systematic error) and the play of chance (random error).

IPD meta-analysis requires a considerable amount of time, personnel, financial resources, and international cooperation of all individuals and groups who have conducted relevant original research (Oxman et al. 1995).

Summary

Results of a meta-analysis using AD and IPD correlate in the majority of cases (Tudur Smith et al. 2016; Huang et al. 2016). However, IPD meta-analyses can produce critical results that might not have been obtainable in any other way (Clarke and Stewart 1998). IPD offers the advantage of a more thorough analysis and investigation of subgroup differences to go "beyond the grand mean" and help in choosing the treatment most suitable for an individual patient, i.e., "individualised medicine" (Davey Smith et al. 1997). The main obstacles in performing IPD meta-analyses include difficulties in procuring data, and the need for statistical expertise, financial resources, and considerable time.

References

Askie LM, Darlow BA, Finer N, et al. Association between oxygen saturation targeting and death or disability in extremely preterm infants in the neonatal oxygenation prospective meta-analysis collaboration. JAMA. 2018;319(21):2190–201.

Chalmers TC. Meta-analysis. Lancet. 1987;1(8548):1492.

Clarke MJ. Individual patient data meta-analyses. Best Pract Res Clin Obstet Gynaecol. 2005;19 (1):47–55.

Clarke M, Stewart L. Re: "Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies". Am J Epidemiol. 1998;148(1):102–3.

Davey Smith G, Egger M, Phillips AN. Meta-analysis. Beyond the grand mean? BMJ. 1997;315 (7122):1610–1614.

de Weerd M, Greving JP, Hedblad B, et al. Prevalence of asymptomatic carotid artery stenosis in the general population: an individual participant data meta-analysis. Stroke. 2010;41(6):1294–7.

Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med. 2015;12(10):e1001886.

Deeks JJ HJ, Altman DG (editors). Analysing data and understanding meta-anlyses. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editor. Cochrane handbook for systematic reviews of interventions, 2nd ed. Chichester (UK): John Wiley and Sons; 2019. p. 241–284.

Higgins JP, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Stat Med. 2001;20(15):2219–41.

Hua H, Burke DL, Crowther MJ, Ensor J, Tudur Smith C, Riley RD. One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information. Stat Med. 2017;36(5):772–89.

Huang Y, Mao C, Yuan J, et al. Distribution and epidemiological characteristics of published individual patient data meta-analyses. PLoS ONE. 2014;9(6):e100151.

Huang Y, Tang J, Tam WW, et al. Comparing the overall result and interaction in aggregate data meta-analysis and individual patient data meta-analysis. Medicine (Baltimore). 2016;95(14): e3312.

Jeng GT, Scott JR, Burmeister LF. A comparison of meta-analytic results using literature vs individual patient data. Paternal cell immunisation for recurrent miscarriage. JAMA. 1995;274 (10):830–836.

Kawahara T, Fukuda M, Oba K, Sakamoto J, Buyse M. Meta-analysis of randomised clinical trials in the era of individual patient data sharing. Int J Clin Oncol. 2018;23(3):403–9.

Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. J Clin Epidemiol. 2002;55(1):86–94.

Collaboration EH. Laparoscopic compared with open methods of groin hernia repair: systematic review of randomised controlled trials. Br J Surg. 2000;87(7):860–867.

Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. BMC Med Res Methodol. 2005;5:14.

McCormack K, Scott NW, Go PM, Ross S, Grant AM. Laparoscopic techniques versus open techniques for inguinal hernia repair. Cochrane Database Syst Rev. 2003(1):Cd001785.

McCormack K, Grant A, Scott N. Value of updating a systematic review in surgery using individual patient data. Br J Surg. 2004;91(4):495–9.

Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264–269, w264.

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 Quality of Reporting of Meta-analyses. Lancet. 1999;354(9193):1896–900.

Nevitt SJ, Marson AG, Davie B, Reynolds S, Williams L, Smith CT. Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review. BMJ. 2017;357:j1390.

O'Rourke K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. J R Soc Med. 2007;100(12):579–82.

Oxman AD, Clarke MJ, Stewart LA. From science to practice: meta-analyses using individual patient data are needed. JAMA. 1995;274(10):845–846.

Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med. 1992;116 (1):78–84.

Report on certain enteric fever inoculation statistics. Br Med J. 1904;2(2288):1243–6.

Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:c221.

Rogozińska E, Marlin N, Thangaratinam S, Khan KS, Zamora J. Meta-analysis using individual participant data from randomised trials: opportunities and limitations created by access to raw data. Evid Based Med. 2017;22(5):157–62.

Simmonds MC, Higgins JP. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. Stat Med. 2007;26(15):2982–99.

Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomised trials: a review of methods used in practice. Clin Trials. 2005;2(3):209–17.

Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: a review of current practice. Contemp Clin Trials. 2015;45(Pt A):76–83.

Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet. 1993;341(8842):418–22.

Stewart GB, Altman DG, Askie LM, Duley L, Simmonds MC, Stewart LA. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. PLoS ONE. 2012;7(10):e46042.

Stewart LA, Clarke M, Rovers M, et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA. 2015;313 (16):1657–65.

Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. Lancet. 2017;389(10086):e12–4.

Tierney JFSL, Clarke M. Individual participant data. In: Higgins JPTTJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions. Chichester (UK): John Wiley and Sons; 2019. p. 643–58.

Tierney JF, Vale C, Riley R, et al. Individual Participant Data (IPD) meta-analyses of randomised controlled trials: guidance on their use. PLoS Med. 2015a;12(7):e1001855.

Tierney JF, Pignon JP, Gueffyier F, et al. How individual participant data meta-analyses have influenced trial design, conduct, and analysis. J Clin Epidemiol. 2015;68(11):1325–35.

Tudur Smith C, Dwan K, Altman DG, Clarke M, Riley R, Williamson PR. Sharing individual participant data from clinical trials: an opinion survey regarding the establishment of a central repository. PLoS ONE. 2014;9(5):e97886.

Tudur Smith C, Marcucci M, Nolan SJ, et al. Individual participant data meta-analyses compared with meta-analyses based on aggregate data. Cochrane Database Syst Rev. 2016;9(9): Mr000007.

Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med. 2000;19(24):3417–32.

Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials. JAMA. 1991;266(1):93–8.

Systematic Reviews of Diagnostic Test Accuracy

Mohan Pammi and Yemisi Takwoingi

Abstract Systematic reviews of diagnostic test accuracy (DTA) are increasingly being published. Diagnostic accuracy is the ability of a test to discriminate between those who have or do not have a target condition. The accuracy of a test is determined by assessing the results of an index test against a reference standard, sometimes known as the 'gold' standard. The reference standard is the best available way to verify the presence or absence of the target condition. DTA systematic reviews summarise evidence on the accuracy of a single index test or compare the accuracy of two or more tests, including an investigation of the reasons for heterogeneity. Heterogeneity in DTA systematic reviews may be due to characteristics of the population, index test and reference standard, as well as the design and conduct characteristics of the studies. Systematic reviews of DTA present greater challenges than those of randomised controlled trials of interventions. This chapter briefly covers the principles and practice of systematic reviews of DTA.

Keywords Diagnostic accuracy · Likelihood ratio · Predictive value · Receiver operating characteristic (ROC) plot · Systematic review · Sensitivity · Specificity

Introduction

Medical literature has exploded in the past few decades, and it has become essential to synthesise evidence into summaries and synopses. Properly conducted systematic reviews of primary research studies use a scientific process, which limits bias in the identification and selection of studies and enables critical appraisal and synthesis of

M. Pammi (✉)
Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA
e-mail: mohanv@bcm.edu

Y. Takwoingi
Institute of Applied Health Research, Public Health Building, University of Birmingham, Birmingham B15 2TT, UK
e-mail: y.takwoingi@bham.ac.uk

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_15

relevant studies that address a specific clinical question (Cook et al. 1997). Systematic reviews can help establish whether findings are consistent and generalisable, or whether findings vary between studies (i.e. heterogeneity). A systematic review may include at least one meta-analysis. Meta-analysis is a scientific technique for pooling the results of multiple studies to obtain more precise estimates and to quantify the extent of heterogeneity.

Systematic reviews of diagnostic test accuracy (DTA) are increasingly being published. Diagnostic accuracy is the ability of a test to discriminate between those who have or do not have a target condition. The accuracy of a test is determined by assessing the results of an index test (a new or existing test of interest) against a reference standard, sometimes known as a 'gold' standard (Table 1). The reference standard is the best available way to verify the presence or absence of the target condition. DTA systematic reviews summarise evidence on the accuracy of a single index test or compare the accuracy of two or more tests, including an investigation of reasons for heterogeneity (Leeflang 2014). Heterogeneity is common in DTA systematic reviews and may be due to characteristics of the population, index test and reference standard, as well as features related to the design and conduct of studies (Macaskill 2013).

The methods used for systematic reviews have an impact on their validity. Several stages in the conduct of DTA systematic reviews present greater challenges than those of systematic reviews of randomised controlled trials (RCTs) of interventions. Recognising the complexity of DTA reviews, the Cochrane Collaboration, the world's largest producer of systematic reviews, delayed introducing this review type into the Cochrane Library until there were sufficient development and understanding of methodology to support their implementation and production. The first Cochrane DTA review was published in 2008, 12 years after the formal registration of the Cochrane Screening and Diagnostic Test Methods Group (2). Many Cochrane DTA reviews have since been published (https://www.cochranelibrary.com).

Table 1 Classification of index test results against reference standard results (2 × 2 table)

	Reference standard + ve	Reference standard-ve	Total
Index test + ve	a (true positives)	b (false positives)	a + b (test positive)
Index test -ve	c (false negatives)	d (true negative)	c + d (test negative)
Total	a + c (disease positive)	b + d (disease negative)	a + b+c + d total analyzed

Adapted from Takwoingi et al. (2015)

Introduction to Diagnostic Accuracy Studies

The ideal study design to assess the clinical performance of a test is a study of a consecutive series of patients. These patients should be prospectively recruited from the target population in whom the test will be applied in practice. For example, neonates clinically suspected of having sepsis or infection should be recruited from a neonatal intensive care unit where an index test, a molecular assay, for the diagnosis of sepsis will be applied.

Most test evaluations focus on the accuracy of a single test without making comparisons with alternative tests that can be used at the same point in the diagnostic pathway (Takwoingi et al. 2013). However, for clinical decision making, evaluations of a single test are of limited value when alternative tests are available. Well-designed comparative (head-to-head) studies of two or more tests enable evaluation of new tests against existing testing pathways and guide test selection, thereby facilitating decision making.

The common measures used to describe the accuracy of a test are sensitivity, specificity, positive and negative predictive values, and positive and negative likelihood ratios (Table 2). The results of a diagnostic test can be a binary outcome (e.g. positive or negative blood culture), continuous outcome (e.g. BNP for diagnosis of hemodynamically significant patent ductus arteriosus (Kulkarni et al. 2015)) or an outcome with an ordered set of categories (e.g. intraventricular hemorrhage from grade I to grade IV).

If the test result is a continuous or an ordinal outcome, plotting sensitivity and specificity at different thresholds for defining the positivity of the test result is useful for exploring the relationship between sensitivity and specificity across thresholds. The visual graphical representation of the relationship between sensitivity and specificity is known as a receiver operating characteristic (ROC) plot (Akobeng 1992). Traditionally, the ROC plot is a plot of sensitivity against 1-specificity. The line joining the points of sensitivity and 1-specificity of a test at different thresholds is called the ROC curve. The diagonal line on the ROC plot joining the lower left-hand corner (0, 0) and the upper right-hand corner (1, 1) depicts the characteristics of a test which is not useful in clinical practice. This line indicates that the test detects an equal proportion of true and false positives and cannot discriminate between those with disease and those without the disease. The position of the ROC curve depends on the accuracy of the test, with a more accurate test having a curve that is closer to the upper left corner of the ROC plot. The ROC curve is useful: to determine the threshold, which optimises either sensitivity or specificity or both, assess the diagnostic accuracy of a test and to compare two or more diagnostic tests. Measures such as the diagnostic odds ratio (DOR) and the area under the curve (AUC) can be used to describe the ROC curve as they summarise the accuracy of a test across all possible thresholds.

Table 2 Measures of test accuracy

Measure	Estimation	Definition
Sensitivity (sens)	a/a + c	The proportion of those *with* the target condition correctly identified as having the condition
Specificity (spec)	d/b + d	The proportion of those *without* the target condition correctly identified by the test as not having the condition
Positive predictive value PPV	a/a + b	The proportion of those *with* the target condition out of the test positives
Negative predictive value NPV	d/c + d	The proportion of those *without* the target condition out of the test negatives
Positive likelihood ratio LR+	a/(a + c)/b/(b + d) or sens/1-spec	The ratio of the proportion who tested positive out of those *with* the target condition to the proportion who tested positive out of those *without* the target condition
Negative likelihood ratio LR-	c/(a + c)/d/(b + d) or 1-sens/spec	The ratio of the proportion who tested negative out of those *with* the target condition to the proportion who tested negative out of those *without* the target condition
Diagnostic odds ratio	ad/bc or LR +/LR-	The ratio of the odds of positivity in those *with* the target condition relative to the odds of positivity in those *without* the condition

Adapted from Takwoingi et al. (2015)

Roadmap for Performing a DTA Systematic Review

Step 1: Define the review question

Like any systematic review, the first and crucial step is to formulate and refine the question to be answered by the systematic review. The main items of a DTA review question are.

(i) **P**atients or population in whom the test will be used
(ii) **I**ndex test: the new test or test of interest.
(iii) **C**omparator test: applies to reviews which compare the accuracy of two or more tests (comparative accuracy reviews). The comparator may be an existing test or current practice, or other index tests.
(iv) **T**arget condition or the disease that is defined by the reference standard
(v) **R**eference standard or the gold standard that defines the target condition

The review question determines the search strategy and eligibility criteria for selecting studies for inclusion in the review, and the interpretation of the review findings (Leeflang 2014). It is good practice to spend time defining and refining the review question to ensure the appropriate clinical question is addressed. The other major aspect to consider in the review question is where the test fits in the diagnostic pathway. New diagnostic tests may replace an existing test, be used as a triage test before an existing test, or as an 'add on' to an existing test in the pathway (Bossuyt et al. 2006).

Step 2: Search for relevant literature

The search should be comprehensive to avoid missing relevant studies. As a minimum, at least two bibliographic databases such as MEDLINE and EMBASE should be searched (Preston et al. 2015). Conference proceedings using BIOSIS or other topic relevant conference abstracts are also useful data sources. Search terms should include terms related to the key elements of the review question; these should be used as keywords or medical subject headings (MESH). Methodological search filters based on terms such as sensitivity, specificity or diagnostic accuracy may not decrease the number of irrelevant articles retrieved and may miss relevant ones. Therefore, such filters are not recommended (Beynon et al. 2013). It is highly recommended that a librarian assists with the development of the search strategy. For transparency and replication purposes, the search strategy and databases used should be reported in the systematic review.

Step 3: Study selection and data extraction

Ideally, two authors should independently assess studies for inclusion and discuss any discrepancies. The titles and abstracts of the articles identified are screened for relevance, followed by the screening of the full text of all potentially relevant titles. Studies that meet the eligibility criteria are selected for inclusion in the review. For the studies excluded, reasons for exclusion should be documented. The flow of studies through the screening and selection process should be illustrated using a PRISMA flow diagram (http://prisma-statement.org/PRISMAStatement/FlowDiagram.aspx). Characteristics of the population, index test, reference standard, and target condition, study design features and test accuracy data should be extracted using a piloted data extraction form. The data should be extracted by two authors independently and any discrepancies resolved by discussion.

Step 4: Assessment of methodological quality

Assessment of the methodological quality of the included studies is essential. The QUADAS-2 tool (available at www.quadas.org) is the checklist recommended by Cochrane (Macaskill 2013; Whiting et al. 2011). The tool comprises four domains, namely patient selection, index test, reference standard, and flow and timing. The domains are assessed in terms of risk of bias (i.e. internal validity) and applicability concern (i.e. external validity), except for the flow and timing domain, which only addresses the risk of bias. Each domain has signalling questions, which aid the assessor in making an overall risk of bias judgement (high risk, low risk or unclear risk) for the domain. The results of the QUADAS-2 assessment are typically summarised graphically (Fig. 1).

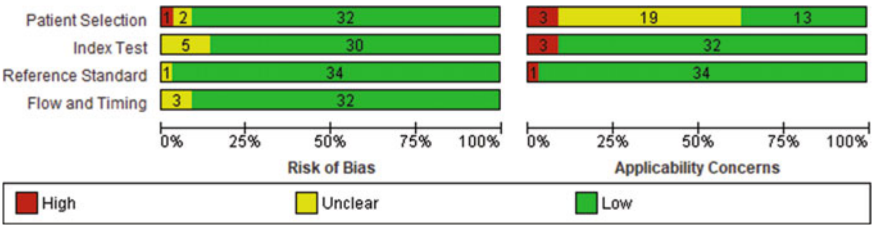
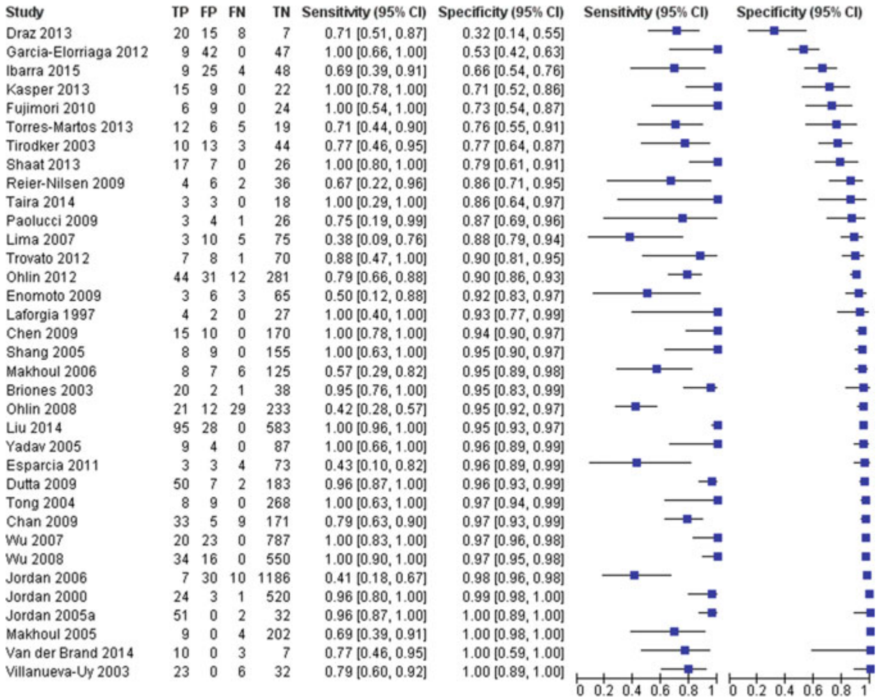


Fig. 1 Assessment of methodological quality of included studies. (Adapted from (Pammi et al. 2017))

Step 5: Data analysis and synthesis

Sensitivities and specificities from the included studies should be plotted on forest plots (e.g. Fig. 2) and in ROC space for preliminary investigation of the data. If there is adequate data, meta-analysis is used for data synthesis. Hierarchical (mixed) models are recommended for meta-analysis of DTA studies. These models account for both within- and between-study variability, as well as the correlation between sensitivity and specificity across studies (Takwoingi et al. 2015; Macaskill et al. 2010). The two most commonly used hierarchical models are the bivariate model (Reitsma et al. 2005; Chu and Cole 2006) and the hierarchical summary ROC (HSROC) model (Rutter and Gatsonis 2001). The bivariate model focuses on estimating a summary point (summary sensitivity and specificity) while the HSROC model focuses on estimating a summary curve. Both models are mathematically equivalent when covariates for test comparisons or investigations of heterogeneity are not included (Harbord et al. 2007). A **covariate** is a variable that may affect test performance, e.g. type of assay kit in measurements of BNP in a neonate with patent ductus arteriosus.

The summary sensitivity and specificity of a test should only be estimated if the test has a binary outcome (e.g. abnormal versus normal), or if studies report accuracy at the same threshold for an ordinal or continuous outcome. Confidence and prediction regions can be drawn around this summary point on an SROC plot to illustrate uncertainty around the summary estimates and the extent of heterogeneity, respectively (Fig. 3). The 95% confidence region can be regarded as a two-dimensional 95% confidence interval around the summary point that also reflects the correlation between sensitivity and specificity. A 95% confidence region denotes an area, based on the available data, within which we would expect the 'real value' to be 95% of the time. The 95% prediction region around the summary point indicates the region where we would expect the results from a new study in the future to lie 95% of the time and is, therefore, wider than the confidence region as it goes beyond the uncertainty in the available data. If studies report different thresholds, then the estimation of a summary curve is more appropriate (Fig. 4). However, if a study reports 2 × 2 data at multiple thresholds, a threshold needs to be selected as only one 2 × 2 table for a test can be included per the study in an analysis using the HSROC model. Methods that extend hierarchical models to



Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Draz 2013	20	15	8	7	0.71 [0.51, 0.87]	0.32 [0.14, 0.55]
Garcia-Elorriaga 2012	9	42	0	47	1.00 [0.66, 1.00]	0.53 [0.42, 0.63]
Ibarra 2015	9	25	4	48	0.69 [0.39, 0.91]	0.66 [0.54, 0.76]
Kasper 2013	15	9	0	22	1.00 [0.78, 1.00]	0.71 [0.52, 0.86]
Fujimori 2010	6	9	0	24	1.00 [0.54, 1.00]	0.73 [0.54, 0.87]
Torres-Martos 2013	12	6	5	19	0.71 [0.44, 0.90]	0.76 [0.55, 0.91]
Tirodker 2003	10	13	3	44	0.77 [0.46, 0.95]	0.77 [0.64, 0.87]
Shaat 2013	17	7	0	26	1.00 [0.80, 1.00]	0.79 [0.61, 0.91]
Reier-Nilsen 2009	4	6	2	36	0.67 [0.22, 0.96]	0.86 [0.71, 0.95]
Taira 2014	3	3	0	18	1.00 [0.29, 1.00]	0.86 [0.64, 0.97]
Paolucci 2009	3	4	1	26	0.75 [0.19, 0.99]	0.87 [0.69, 0.96]
Lima 2007	3	10	5	75	0.38 [0.09, 0.76]	0.88 [0.79, 0.94]
Trovato 2012	7	8	1	70	0.88 [0.47, 1.00]	0.90 [0.81, 0.95]
Ohlin 2012	44	31	12	281	0.79 [0.66, 0.88]	0.90 [0.86, 0.93]
Enomoto 2009	3	6	3	65	0.50 [0.12, 0.88]	0.92 [0.83, 0.97]
Laforgia 1997	4	2	0	27	1.00 [0.40, 1.00]	0.93 [0.77, 0.99]
Chen 2009	15	10	0	170	1.00 [0.78, 1.00]	0.94 [0.90, 0.97]
Shang 2005	8	9	0	155	1.00 [0.63, 1.00]	0.95 [0.90, 0.97]
Makhoul 2006	8	7	6	125	0.57 [0.29, 0.82]	0.95 [0.89, 0.98]
Briones 2003	20	2	1	38	0.95 [0.76, 1.00]	0.95 [0.83, 0.99]
Ohlin 2008	21	12	29	233	0.42 [0.28, 0.57]	0.95 [0.92, 0.97]
Liu 2014	95	28	0	583	1.00 [0.96, 1.00]	0.95 [0.93, 0.97]
Yadav 2005	9	4	0	87	1.00 [0.66, 1.00]	0.96 [0.89, 0.99]
Esparcia 2011	3	3	4	73	0.43 [0.10, 0.82]	0.96 [0.89, 0.99]
Dutta 2009	50	7	2	183	0.96 [0.87, 1.00]	0.96 [0.93, 0.99]
Tong 2004	8	9	0	268	1.00 [0.63, 1.00]	0.97 [0.94, 0.99]
Chan 2009	33	5	9	171	0.79 [0.63, 0.90]	0.97 [0.93, 0.99]
Wu 2007	20	23	0	787	1.00 [0.83, 1.00]	0.97 [0.96, 0.98]
Wu 2008	34	16	0	550	1.00 [0.90, 1.00]	0.97 [0.95, 0.98]
Jordan 2006	7	30	10	1186	0.41 [0.18, 0.67]	0.98 [0.96, 0.98]
Jordan 2000	24	3	1	520	0.96 [0.80, 1.00]	0.99 [0.98, 1.00]
Jordan 2005a	51	0	2	32	0.96 [0.87, 1.00]	1.00 [0.89, 1.00]
Makhoul 2005	9	0	4	202	0.69 [0.39, 0.91]	1.00 [0.98, 1.00]
Van der Brand 2014	10	0	3	7	0.77 [0.46, 0.95]	1.00 [0.59, 1.00]
Villanueva-Uy 2003	23	0	6	32	0.79 [0.60, 0.92]	1.00 [0.89, 1.00]

Fig. 2 Forest plot of molecular tests for diagnosis of neonatal sepsis. FN: false negative; FP: false positive; TN: true negative; TP: true positive. The forest plot shows estimates of sensitivity and specificity with 95% confidence intervals (CIs) for each included study. The studies are sorted on the plot by specificity. **Cochrane Database of Systematic Reviews**25 Feb 2017 https://doi.org/10.1002/14651858.cd011926.pub2 (Adapted from (Pammi et al. 2017))

allow for the inclusion of multiple 2 × 2 tables from each included study exist and two of the methods are particularly promising (Steinhauser et al. 2016; Jones et al. 2019).

Risk of bias and applicability concerns graph: Review authors' judjements about each domain are presented as percentages across included studies. The numbers shown on each bar represent the number of studies that were scored as high, unclear or low in terms of risk of bias or applicability concern for that domain.

To assess the relative accuracy of two or more tests, two strategies are typically used; use all available studies that have evaluated at least of one the tests (indirect test comparison, Fig. 4) or restrict the analysis to only studies that have compared the tests head-to-head (direct test comparison) (Macaskill et al. 2010; Leeflang et al. 2008). Direct comparisons are less prone to bias due to confounding because the tests have been compared in the same study population. However, the availability of such comparative studies is often limited (Takwoingi et al. 2013). If an indirect comparison is a primary analysis, a direct comparison should also be performed as a secondary analysis if comparative studies are available. A meta-regression approach

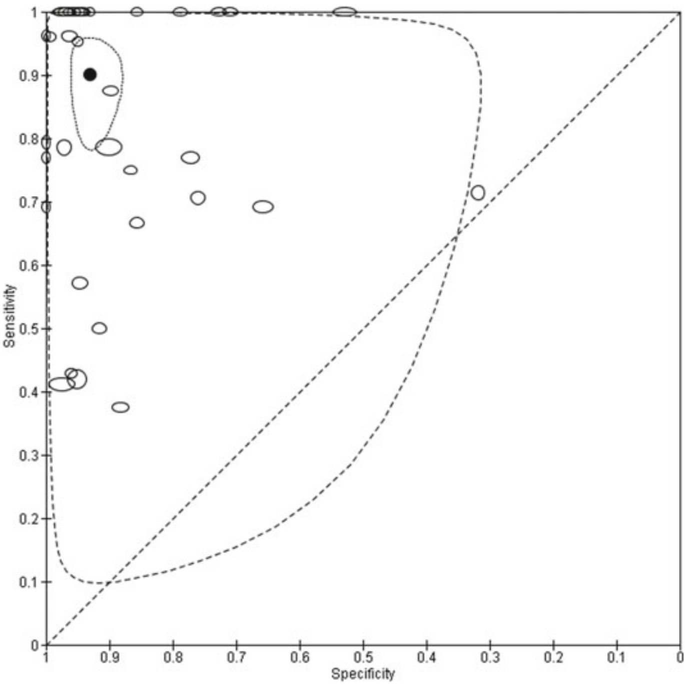


Fig. 3 Studies reporting accuracy of molecular assays in neonatal sepsis are plotted in ROC space as clear circles. The summary estimate is depicted by the black filled circle and is surrounded by 95% confidence and 95% prediction regions. **Cochrane Database of Systematic Reviews** 25 Feb 2017 https://doi.org/10.1002/14651858.cd011926.pub2. (Adapted from (Pammi et al. 2017))

is typically used for comparative meta-analysis by adding a covariate for test type to a hierarchical model. The choice of model (bivariate or HSROC model) depends on whether summary points or curves are appropriate given the research question and the available data.

Visual inspection of forest and SROC plots are useful for exploring variability between studies. For the molecular assay example, the forest plot (Fig. 2) show that sensitivity estimates are more variable between studies than specificity. This may be due to the small number of cases in many studies as well as other factors. The SROC plot (Fig. 3) also show considerable scatter of the studies in ROC space. To formally investigate heterogeneity, meta-regression can be performed by adding a potential source of heterogeneity as a covariate to a hierarchical model. The selection of covariates should be justified and pre-specified in the review protocol. Further details on methods for meta-analysis and examples are available in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Macaskill et al. 2010). Software programs, tutorials and online learning modules are available on the Cochrane Screening and Diagnostic Tests Methods Group website (https://methods.cochrane.org/sdt/).

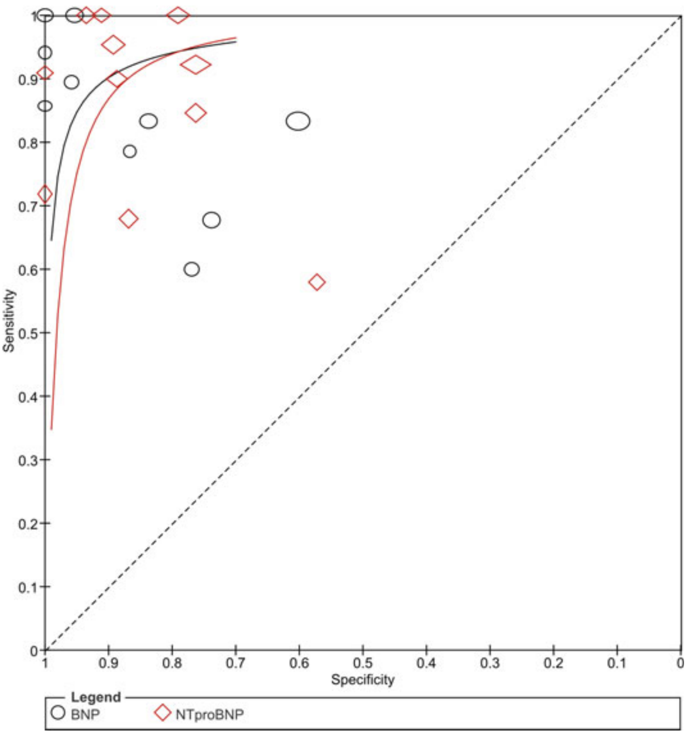


Fig. 4 Summary curves for BNP (black line) and NT-proBNP (red line) for diagnosis of hemodynamically significant PDA in preterm neonates (Adapted from (Kulkarni et al. 2015))

Step 6: Interpretation and drawing conclusions

The results of the meta-analysis should be interpreted within the context of the systematic review question. In addition, the implications of the results in clinical practice and the consequences of a false positive or a false negative test result should be explained. Expressing summary sensitivity and specificity using natural frequencies will aid understanding of the review findings and the consequences of inaccurate test results. The meta-analysis of molecular assays for neonatal sepsis gave a mean sensitivity of 0.90 (95% CI 0.82 to 0.95) and a mean specificity of 0.93 (95% CI 0.89 to 0.96) (Pammi et al. 2017). If we apply the summary estimates of this review to a theoretical cohort of 1000 very low birth weight neonates screened for late-onset sepsis (sepsis after the first 72 h of life, prevalence 10%), ten culture-positive cases will be missed, and 63 neonates without sepsis will be treated unnecessarily. Ideally, we do not want to miss any case of neonatal sepsis as the consequences are severe and overtreatment is not a huge issue. Therefore, the review concluded that currently available molecular assays do not have sufficient diagnostic accuracy to replace microbial cultures. The limitations of a review, including issues related to heterogeneity, risk of bias and applicability concerns should be considered when drawing conclusions.

Table 3 Key differences between systematic reviews of diagnostic accuracy studies and systematic reviews of *RCTs of interventions

Process	Systematic review of diagnostic accuracy studies	Systematic review of RCTs
Question formulation	PICT (population, index test, comparator and target condition)	PICO (population, intervention, comparison, and outcome)
Study identification	Studies poorly tagged in electronic databases and therefore difficult to find	RCTs better tagged and easier to find
Study selection	Cross-sectional and cohort studies	Parallel RCTs, cluster-randomised and cross-over trials
Methodological quality assessment	QUADAS-2 tool	Cochrane risk of bias tool
Meta-analysis	Hierarchical models (bivariate and HSROC models)	Traditional univariate random-effects and fixed-effect models

*RCTs: Randomised controlled trials

Summary and Conclusions

Systematic reviews of diagnostic accuracy are challenging and should be undertaken by a review team with adequate clinical and methodological expertise. Table 3 summarises key differences between systematic reviews of diagnostic accuracy and those of RCTs of interventions.

References

Akobeng AK. Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatr (Oslo, Norway: 1992). 2007;96(5):644–647.

Beynon R, Leeflang MM, McDonald S, et al. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. Cochrane Database Syst Rev. 2013(9):MR000022.

Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089–92.

Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalised linear mixed model approach. J Clin Epidemiol. 2006;59(12):1331–1332; author reply 1332-1333.

Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med. 1997;126(5):376–80.

Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239–51.

Jones HE, Gatsonsis CA, Trikalinos TA, Welton NJ, Ades AE. Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. Stat Med. 2019;38(24):4789–803.

Kulkarni M, Gokulakrishnan G, Price J, Fernandes CJ, Leeflang M, Pammi M. Diagnosing significant PDA using natriuretic peptides in preterm neonates: a systematic review. Pediatrics. 2015;135(2):e510–25.

Leeflang MM. Systematic reviews and meta-analyses of diagnostic test accuracy. Clin Microbiol Infect. 2014;20(2):105–13.

Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149(12):889–97.

Leeflang MM, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. Syst Rev. 2013;2:82.

Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane handbook for systematic reviews of diagnostic test accuracy, version 1.0: The Cochrane Collaboration; 2010. http://srdta.cochrane.org/.

Pammi M, Flores A, Versalovic J, Leeflang MM. Molecular assays for the diagnosis of sepsis in neonates. Cochrane Database Syst Rev. 2017;2:CD011926.

Preston L, Carroll C, Gardois P, Paisley S, Kaltenthaler E. Improving search efficiency for systematic reviews of diagnostic test accuracy: an exploratory study to assess the viability of limiting to MEDLINE, EMBASE and reference checking. Syst Rev. 2015;4:82.

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982–90.

Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20(19):2865–84.

Steinhauser S, Schumacher M, Rucker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC Med Res Methodol 2016;16(1):97.

Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med. 2013;158(7):544–54.

Takwoingi Y, Riley RD, Deeks JJ. Meta-analysis of diagnostic accuracy studies in mental health. Evid Based Mental Health. 2015;18(4):103–9.

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.

Network Meta-Analysis

Sanjay Patole

Abstract Clinicians are often faced with results of randomised controlled trials comparing different interventions for a condition. However, selecting from a range of interventions is difficult when no head-to-head trials are comparing their safety and efficacy. Network meta-analysis (NMA) extends the principles of conventional meta-analysis to allow assessment of multiple treatments in a single analysis. Hence it is also called as multiple treatment meta-analysis or mixed treatment comparisons. Importantly NMA can provide evidence on 'relative ranking' of multiple interventions. The ability to synthesise indirect evidence and evaluate multiple interventions with a common comparator in one analysis separates NMA from conventional pairwise meta-analyses. NMA is vital for evidence-based decision-making because it allows assessment of direct as well as indirect evidence. Given its complexity compared with conventional meta-analyses, the involvement of both subject experts and experienced biostatistician is necessary when planning an NMA. This is particularly important because crucial judgements and assumptions are involved. This chapter briefly covers the principles of NMA.

Keywords Network · Meta-analysis · Transitivity · Coherence · Equivalence · League table · Rankogram

Introduction

Network meta-analysis (NMA) extends the principles of conventional meta-analysis to allow assessment of multiple treatments in a single analysis (Dias and Caldwell 2019; Rouse et al. 2017; Dias et al. 2018; ter Veer et al. 2019; Dobler

S. Patole (✉)
School of Medicine, University of Western Australia, Perth, WA 6009, Australia
e-mail: sanjay.patole@health.wa.gov.au

S. Patole
Neonatal Directorate, King Edward Memorial Hospital for Women,
Perth, WA 6008, Australia

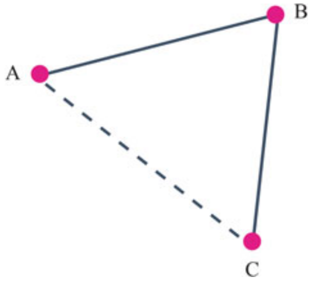
S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_16

et al. 2018; Tonin et al. 2017). Hence it is also called as multiple treatment meta-analysis or mixed treatment comparisons. Importantly, NMA can provide evidence on ‘relative ranking’ of multiple interventions.

Considering its complexity compared with conventional meta-analyses, the involvement of both subject experts and experienced biostatistician is necessary when planning an NMA (Cipriani et al. 2013). This is particularly important because crucial judgements and assumptions are involved (Mills et al. 2012; Faltinsen et al. 2018). The principles of NMA are covered briefly considering that a detailed discussion on its methodology and interpretation is beyond the scope of this chapter.

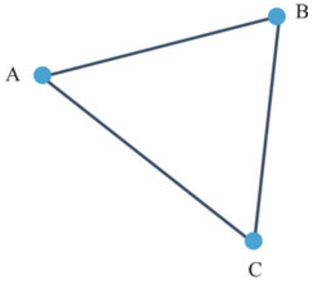
Clinicians are often faced with results of randomised controlled trials (RCT) comparing different interventions for a condition. However, selecting from a range of interventions is difficult when no head-to-head trials are comparing their safety and efficacy. NMA is vital for evidence-based decision-making because it allows assessment of both, direct as well as indirect evidence (Quan et al. 2017; Chaimani et al. 2019).

The ability to synthesise indirect evidence and evaluate multiple interventions with a common comparator in one analysis separates NMA from conventional pairwise meta-analyses (Kiefer et al. 2015; Hoaglin et al. 2011; Jansen et al. 2011). The direct evidence is the estimate of relative effects of the interventions provided by RCTs. In contrast, the indirect evidence is inferred by observing the results of direct comparisons. Indirect evidence, being observational, is subject to bias due to confounders. In its simplest form, the principle of deriving indirect evidence could be explained as follows: In a head-to-head RCT, intervention A is better than B. In another head-to-head RCT, intervention B is equal to C, assuming everything else is similar between the studies, the conclusion derived is that intervention A is better than C, when in fact there has been no RCT directly comparing A against C. Intervention C is the common comparator in this case. This is an example of an



*Head to head comparisons (*represented by solid lines*) provide *‘Direct’* evidence on the effect of intervention A vs. B, and B vs. C. There is no trial directly comparing A vs. C. However, everything else being similar, the effect of such a comparison can be *’indirectly’* derived (*represented by dotted lines)* by effect estimates of A vs. B, and B vs. C.

Fig. 1 Network of evidence (Open-loop) from clinical trials*

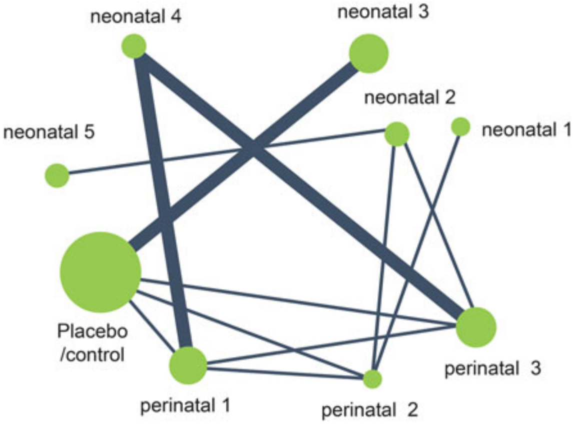


*Three-arm clinical trial providing *'direct'* evidence for head to head comparison of interventions A, B, and C.

Fig. 2 Network of evidence (Closed loop) from a clinical trial*

'open loop' or 'network' (A-B-C) of evidence (Fig. 1 (Dias and Caldwell 2019; Rouse et al. 2017; Chaimani et al. 2019). An example of a 'closed' loop of (direct) evidence can be provided by an RCT comparing three interventions (A, B, C) against each other (Fig. 2). A typical network diagram in a network meta-analysis of RCTs comparing various interventions for a condition is shown in Fig. 3.

An important benefit of NMA is that it preserves within-trial randomisation by pooling the relative treatment effects estimated across RCTs. As long as the interventions assessed to form a connected network of comparisons, the relative



*Hypothetical example: Head to head comparisons of trials of probiotic supplementation in the perinatal or neonatal period to prevent allergy in childhood. Interventions compared against each other or a placebo/control.

Note: The size of the circle (*"Node"*) is proportional to the sample size, and the width of the connecting line is proportional to the number of trials involved in the comparison.

Fig. 3 Network diagram in a meta-analysis of clinical trials*

effects and 95% confidence intervals of each intervention compared with every other can be obtained (Dias and Caldwell 2019; Rouse et al. 2017; Chaimani et al. 2019).

Similar to a conventional systematic review, development of a clear focussed question and the PICO framework (patient, intervention, comparator, and outcome) are important initial steps in NMA (Dias and Caldwell 2019; ter Veer et al. 2019; Dobler et al. 2018; Tonin et al. 2017; Chaimani et al. 2019). It is important to consider whether the selected outcomes are clinically important, and the surrogate outcomes are valid for interpretation of the results of NMA. Defining the 'treatment network' is critical (Sturtz and Bender 2012). This includes decisions on the size of the network, nature of interventions, and their clinical relevance. Literature search must be comprehensive, considering the much broader context of an NMA compared with conventional systematic reviews. An effect modifier is a clinical or methodological characteristic of the included trials that has the potential to modify the effect. Subject expertise is important in this context as effect modifiers can be a source of significant heterogeneity influencing the entire network. Pre-stating effect modifiers as well as an assessment of their presence and distribution between studies, is thus important (Dias and Caldwell 2019; Rouse et al. 2017; Dobler et al. 2018; Tonin et al. 2017; Chaimani et al. 2019).

Key Assumptions and Concepts in NMA

The findings of an NMA are valid only if the assumption that except for the interventions being evaluated, there are no systematic differences between the trials included in the analysis, is correct (Dias and Caldwell 2019; Rouse et al. 2017; Dobler et al. 2018; Tonin et al. 2017; Chaimani et al. 2019). Transitivity (also called as similarity) refers to this key assumption in NMA (Fig. 4). It reflects an equal probability that any patient in the network could have received any of the interventions included in the network. Transitivity concerns the validity of making indirect comparisons by assuming a balanced distribution of clinical and

- *Transitivity (similarity):* All trials included in the NMA are comparable in terms of potential effect modifiers.
- *Homogeneity:* There is no significant heterogeneity between trial results in pairwise comparisons.
- *Consistency:* There is no significant discrepancy or inconsistency between direct and indirect evidence.

Fig. 4 Basic assumptions for indirect comparisons (Kiefer et al. 2015)

methodological characteristics of the included trials with direct comparisons. Intransitivity occurs when effect modifiers are not balanced between comparisons. Transitivity relates to the statistical term *'coherence'* (Chaimani et al. 2019). It requires that intervention B is similar when it is used in trials comparing B vs. A, and B vs. C with respect to effect modifiers. Coherence equations provide mathematical links between effects of the interventions assessed so that some effects can be estimated from others provided the assumption of transitivity is correct (Dias and Caldwell 2019; Rouse et al. 2017; Dobler et al. 2018; Tonin et al. 2017; Chaimani et al. 2019).

Homogeneity refers to the comparability of trials within each pairwise comparison in the network (Dias and Caldwell 2019; Rouse et al. 2017; Dobler et al. 2018; Tonin et al. 2017; Chaimani et al. 2019). The degree of heterogeneity for each comparison can be assessed qualitatively (e.g. participant and trial design characteristics) and quantitatively (e.g. I-squared statistic) (Donegan et al. 2013). Unlike homogeneity, transitivity cannot be evaluated quantitatively. Transitivity must be evaluated by careful review of the characteristics of the trials (Chaimani et al. 2019; Donegan et al. 2013).

Consistency refers to the statistical agreement (equivalence) between direct and indirect evidence (Higgins et al. 2012; Dias et al. 2013, 2010; Krahn et al. 2014). Similar to the assessment of heterogeneity, inconsistency can be assessed both qualitatively and quantitatively. Transitivity requires that all interventions included in an NMA should be jointly randomisable. Transitivity must be considered at each step of NMA. Intransitivity (inconsistency) indicates substantial variation in the distribution of effect modifiers between studies included in the network. A quantitative analysis is not appropriate in the presence of intransitivity. Only a qualitative synthesis is justified under such circumstances (Dias and Caldwell 2019; Rouse et al. 2017; Dobler et al. 2018; Tonin et al. 2017; Chaimani et al. 2019).

Decision versus supplementary set: Decision set is the set of interventions in a network that clinicians would be willing to choose for the desired outcome. Supplementary set (e.g. placebo) includes interventions included in the network to improve inference among decision set interventions. Synthesis set includes interventions in the decision as well as the supplementary set (Chaimani et al. 2019).

Important aspects of analysis: Selection of a 'reference' treatment against which all interventions will be compared, and pre-stating the approach to heterogeneity and consistency are important steps (Rouse et al. 2017; Chaimani et al. 2019). The reference treatment can be a placebo, no treatment, or a commonly used treatment. It is important to note that heterogeneity can be comparison-specific or common across comparisons. The approach to consistency can be global (assessed in the entire network) or local (only where the problem is evident). Assessment of effect modifiers in the loops with inconsistency is important in this context once errors in data extraction are ruled out (Rouse et al. 2017). Just as the approach to significant unexplained heterogeneity in a conventional pairwise meta-analysis, it is not appropriate to conduct NMA if there is unexplained significant inconsistency (Rouse et al. 2017; Chaimani et al. 2019).

The commonly used NMA models include the multivariate model or hierarchical model (Rouse et al. 2017). The fixed-effect model or random-effects model can be used for NMA. The random-effects model assumes that the between-study heterogeneity is the same across all comparisons, i.e. a single measure of heterogeneity is calculated across the whole network (Rouse et al. 2017). However, it is possible to fit models allowing for different heterogeneity for each comparison (Lu and Ades 2009). Meta-regression or sensitivity analyses could be used to explore the reasons for significant heterogeneity (Cooper et al. 2009). The results of such analyses are beneficial in guiding further research rather than a clinical practice because at best, they can only generate hypotheses.

As a statistical model, NMA can be fitted using a frequentist or Bayesian approach (Spiegelhalter 2004; Spiegelhalter et al. 2004). Bayesian NMAs are commonly used as they provide ranking and probability outputs for decision-making and allow for greater flexibility in the fitted models (Spiegelhalter 2004; Spiegelhalter et al. 2004). For a detailed discussion on these issues, the reader is referred to the recommended publications for further reading (Dias et al. 2018; Senn et al. 2013).

Presenting results and rankings: A league table is used to show the comparisons of relative effects between pair of interventions for up to two outcomes (Dias and Caldwell 2019; Chaimani et al. 2019; Salanti et al. 2011). The probability of each intervention taking a particular rank is also presented. It is recommended to present rankings using the mean rank, or the cumulative ranking probabilities given the uncertainty in relative effect estimates and relative ranking. The 'rankogram' or the surface under the cumulative ranking curve (SUCRA) values, which take into account the estimated effect sizes and their accompanying uncertainty are used for this purpose (Dias and Caldwell 2019; Chaimani et al. 2019; Salanti et al. 2011). It is important to consider the effect of treatment while interpreting its rank. A high rank alone does not guarantee significant benefits or for that matter, any benefit for a given patient (Dias and Caldwell 2019; Chaimani et al. 2019; Salanti et al. 2011).

Quality of evidence from NMA: The quality of evidence from an NMA is evaluated by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach based on six domains. There are two approaches for applying GRADE to NMA. Both start with evaluating each domain for each direct comparison. For details, the reader is referred to the publications by Salanti et al. (2014), and Puhan et al. (2014).

Reporting: An extension of the PRISMA statement is recommended for reporting NMA (Hutton et al. 2015).

In summary, NMA is a relatively newer and complex technique that allows the synthesis of direct as well as indirect evidence from RCTs that have compared more than two interventions (Dias and Caldwell 2019; Quan et al. 2017; Chaimani et al. 2019; Salanti 2012; Li et al. 2011). Considering that it provides the evidence in totality, NMA improves the efficiency of decision making and the precision of estimates) (Dias and Caldwell 2019; Chaimani et al. 2019). Furthermore, NMA results are more robust as multiple sources of evidence are used (Quan et al. 2017). Critical assessment of the plausibility of the assumption of consistency is critical for assessing the validity and optimal interpretation of NMA (Dias and Caldwell 2019; Chaimani et al. 2019). The risk of bias is higher in NMA compared with conventional pairwise meta-analyses as it combines studies with a higher degree of variability. The influence of factors such as the number of trials in the network, number of trials with more than two comparisons, heterogeneity (*variability* between direct and indirect comparisons), inconsistency (*discrepancy* between direct and indirect comparisons), and bias must be explored adequately for optimal interpretation of NMA (Chaimani et al. 2019; Jansen et al. 2011; Li et al. 2011; Mills et al. 2013).

References

Chaimani A, Caldwell DM, Li T, Higgins JPT, Salanti G. Chapter 11: Undertaking network meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane; 2019. www.training.cochrane.org/handbook.

Cipriani A, Higgins JP, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. Ann Intern Med. 2013;159:130–7.

Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. Stat Med. 2009;28:1861–81.

Dias S, Caldwell DM. Network meta-analysis explained. Arch Dis Child Fetal Neonatal Ed January 2019; 104, 1: F8-F12.

Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. Network meta-analysis for decision making. First published: 12 January 2018 Print ISBN: 9781118647509|Online ISBN: 9781118951651| https://doi.org/10.1002/9781118951651 © 2018 John Wiley & Sons Ltd.

Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010;29:932–44.

Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomised controlled trials. Med Decis Making. 2013;33:641–56.

Dobler CC, Wilson ME, Murad MH. A pulmonologist's guide to understanding network meta-analysis. Eur Respir J. 2018;52:1800525. https://doi.org/10.1183/13993003.00525-2018.

Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. Assessing key assumptions of network meta-analysis: A review of methods. Res Syn Methods. 2013;4:291–323.

Faltinsen EG, Storebø OJ, Jakobsen JC, Boesen K, Lange T, Gluud C. Network meta-analysis: the highest level of medical evidence? BMJ Evid Based Med. 2018;23(2):56–9.

Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: Concepts and models for multi-arm studies. Res Syn Methods. 2012;3:98–110.

Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: Report of the ISPOR task force on indirect treatment comparisons good research practices: Part 2. Value Health. 2011;14:429–37.

Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med. 2015;162:777–84.

Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: Report of the ISPOR task force on indirect treatment comparisons good research practices: part 1. Value Health. 2011;14:417–28.

Kiefer C, Sturtz S, Bender R. Indirect comparisons and network meta-Analyses. Dtsch Arztebl Int. 2015;112(47):803–8. https://doi.org/10.3238/arztebl.2015.0803.

Krahn U, Binder H, König J. Visualizing inconsistency in network meta-analysis by independent path decomposition. BMC Med Res Methodol. 2014;14:131. https://doi.org/10.1186/1471-2288-14-131.

Li T, Puhan MA, Vedula SS, Singh S, Dickersin K, The Ad Hoc Network Meta-analysis Methods Meeting Working Group. Network meta-analysis-highly attractive but more methodological research is needed. BMC Med. 2011; 9: 79. https://doi.org/10.1186/1741-7015-9-79.

Lu G, Ades A. Modelling between-trial variance structure in mixed treatment comparisons. Biostatistics. 2009;10:792–805.

Mills EJ, Ioannidis JP, Thorlund K, Schünemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. JAMA. 2012;308:1246–53.

Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis. BMJ. 2013;346:

Puhan MA, Schünemann HJ, Murad MH, et al for the GRADE Working Group. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014; 349:g5630. https://doi.org/10.1136/bmj.g5630 (Published 24 September 2014).

Quan H, Zhang B, Chuang-Stein C, Jones B & On behalf of the EFSPI integrated data analysis efficacy working group. integrated data analysis for assessing treatment effect through combining information from all sources. Stat Biopharm Res. 2017; 9:1, 52–64.

Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. Intern Emerg Med. 2017;12(1):103–11.

Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: Many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Syn Methods. 2012;3:80–97.

Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JP. Evaluating the quality of evidence from a network meta-analysis. PLoS One. 2014;9(7):e99682. Published 2014 Jul 3. https://doi.org/10.1371/journal.pone.0099682.

Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol. 2011;64(2):163–71.

Senn S, Gavini F, Magrez D, Scheen A. Issues in performing a network meta-analysis. Stat Methods Med Res. 2013;22(2):169–89.

Spiegelhalter DJ. Incorporating Bayesian ideas into health-care evaluation. Stat Sci. 2004;19:156–74.

Spiegelhalter DJ, Abrams KR, Myles J. Bayesian approaches to clinical trials and health-care evaluation. New York: Wiley; 2004.

Sturtz S, Bender R. Unsolved issues of mixed treatment comparison meta-analysis: Network size and inconsistency. Res Syn Methods. 2012;3:300–11.

ter Veer E, van Oijen MGH, van Laarhoven HWM. The use of (network) meta-analysis in clinical oncology. Front Oncol. 27 August 2019| https://doi.org/10.3389/fonc.2019.00822.

Tonin FS, Rotta I, Mendes AM, Pontarolo R. Network meta-analysis: a technique to gather evidence from direct and indirect comparisons. Pharm Prac 2017 Jan-Mar;15(1):943: 1–11. https://doi.org/10.18549/PharmPract.2017.01.943.

Systematic Reviews of Animal Studies

Gayatri Athalye-Jape

Abstract Systematic reviews (SRs) and meta-analysis of clinical studies are well established as the highest level of evidence-based medicine. The concept of summarising evidence from preclinical or animal studies has evolved over the past decade. This process is important for providing animal researchers with a unique framework to study, collate, plan, design and report preclinical research, thereby adding to its translational potential. Furthermore, the concept of preclinical SRs is important to consolidate a humane and cost-effective approach to animal experiments. This chapter highlights the evolution, establishment of preclinical systematic review centre (SYRCLE), importance of the 3Rs, limitations of animal research, benefits of preclinical SRs, method of conducting and reporting the SR, interpreting results and evaluating the level of evidence using GRADE and finally optimising the 'laboratory benchtop' research to reach its highest translational potential at the 'patient's bedside.'

Keywords Animal studies · ARRIVE guidelines · Funding · Systematic reviews · SYRCLE · GRADE · 3Rs · Reproducibility

Introduction

The use of systematic reviews (SR) and meta-analysis is well established in the field of medicine mainly due to the Cochrane Collaboration, which has established a unique framework to assist healthcare providers and policymakers in making evidence-based decisions. However, SRs of preclinical or animal studies is a relatively newer and novel concept (Korevaar et al. 2011; Peters et al. 2006; van Luijk et al. 2014). Millions of animals are used annually for the scientific and educational

G. Athalye-Jape (✉)
Neonatal Directorate, King Edward Memorial Hospital, Perth, WA 6008, Australia
e-mail: gayatri.jape@health.wa.gov.au

G. Athalye-Jape
School of Medicine, University of Western Australia, Perth, Australia

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_17

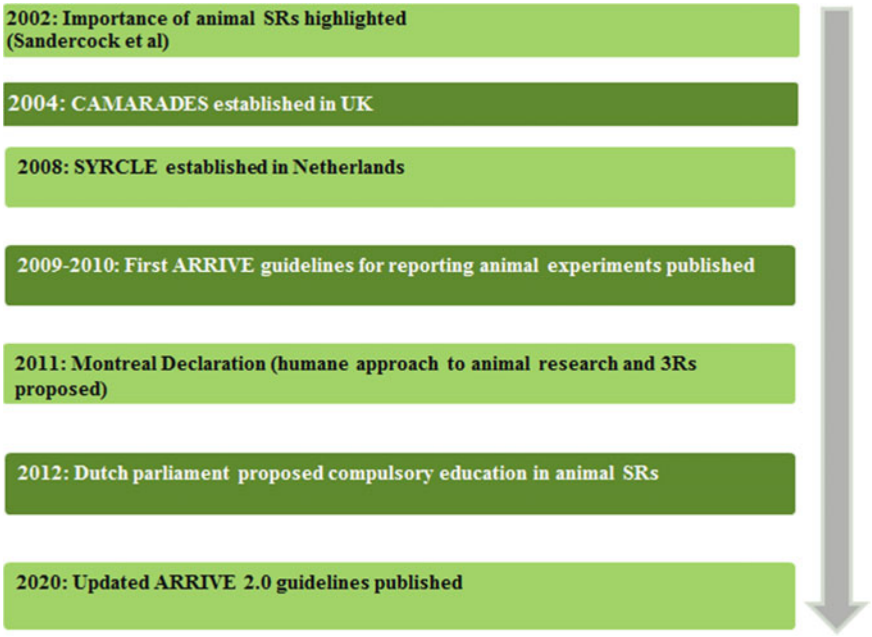
purpose all over the world including in Australia, and several millions of dollars are spent on these experiments (Humane Society International data; www.hsi.org, accessed August 2020). In March 2019, the Australian government passed legislation to end animal testing for new cosmetic ingredients given the fact that more than 20,000 chemical ingredients met the safety and availability standards for use in the cosmetic industry. However, healthcare professionals and the common public strongly believe that animal experiments have significantly contributed to the understanding of human diseases, although this belief is not supported by adequate evidence. Estimating the real contribution of animal studies to healthcare for humans is difficult, given the limitations of the current methods for adequate evaluation of the clinical and translational relevance of such studies. Furthermore, it has recently been argued that completing animal studies before clinical studies may be challenging and not always feasible or achievable (Pound and Ritskes-Hoitinga 2020).

The Need for Systematic Reviews of Animal Research

An emphatic statement on the Animal Experimentation Fact Sheet on the "Animals Australia" website highlights the limitations of animal research: "*Such research continues with little broad and independent evaluation because funding bodies and research institutions are reluctant to embrace the possibility that existing animal models and methods have largely failed. To do so would ruin careers, break the ever-present promises to health charities and the community, and knock out existing 'high tech' animal breeding (and facilities) supply businesses.*" (https://www.animalsaustralia.org/factsheets/animal_experimentation.php).

Pound et al. have emphasised the importance of conducting SRs of animal studies (2004). SRs of animal studies help assess the validity of preclinical evidence, raising awareness of poor study design and encouraging improvements in scientific reporting whilst providing transparency and preventing unnecessary duplication of studies (Pound and Ritskes-Hoitinga 2020) De Vries et al. have illustrated the optimum design, conduct and analysis of animal and human experiments through systematic reviews to generate reliable results and address moral concerns around animal research (de Vries et al. 2014). Van Luijk et al. have conducted a review that included 91 SRs of animal studies. Their results showed that while systematic reviews were worthwhile, there was scope for improvement in their internal validity (2014) (Fig. 1).

Sandercock and Roberts were the first to indicate the importance of SRs of animal experiments as a prerequisite for designing clinical trials (2002). The Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES) collaboration was established in 2004 at the University of Edinburgh, UK. It provided a framework for SRs and meta-analysis of experimental animal data reflecting on the translational failure of 'stroke' related animal experiments (http://www.dcn.ed.ac.uk/camarades/). Sy-RF (http://syrf.org.uk/) is the CAMARADES in vivo systematic reviews and meta-analysis facility. It



2002: Importance of animal SRs highlighted (Sandercock et al)

2004: CAMARADES established in UK

2008: SYRCLE established in Netherlands

2009-2010: First ARRIVE guidelines for reporting animal experiments published

2011: Montreal Declaration (humane approach to animal research and 3Rs proposed)

2012: Dutch parliament proposed compulsory education in animal SRs

2020: Updated ARRIVE 2.0 guidelines published

Fig. 1 Timeline summarising evolution of evidence based animal research

provides an online easily accessible free resource of methodological support, mentoring, guidance, educational materials and practical assistance to preclinical researchers. CAMARADES now has five global, national coordinating centres: University of Edinburgh, Florey Institute of Neuroscience and Mental health, Radboud University Nijmegen Medical Centre, University of California San Francisco and Ottawa Hospital Research Institute.

The Concept of 3Rs: Replacement, Reduction and Refinement

In 2005, the Nuffield Council on Bioethics urged funders of animal research to support SRs (http://nuffieldbioethics.org/wp-content/uploads/The-ethics-of-research-involving-animals-full-report.pdf.). The American Council on Science and Health urged the US government agencies to adopt non-animal-based research methods to test carcinogens. Following this, an improvised version of the concept of the 3R's (**R**eplacement, **R**eduction and **R**efinement of animal use) was used to implement principles of humane science in animal research (Russell and Burch 1959). This was the principal theme of the Montréal declaration (25th August 2011) at the Eighth World Congress on Alternatives and Animal Use in the Life Sciences, which aimed to change the culture of planning, executing, reporting, reviewing and translating animal research. It reinforced the importance of SR of animal studies to produce a scientifically sound and transparent summary of all available evidence (https://3rs.ccac.ca/documents/en/WC8_Declaration_of_Montreal_FINAL.pdf 2020).

(a) **Replacement**: This is defined as accelerating the development and use of models and tools based on the latest science and technologies, for addressing important scientific questions without the use of animals.

Types of replacement: (i) *Full*: Use of tissues and cells, cell lines, mathematical and computer models and human volunteers; (ii) *Partial:* Use of animals not considered capable of suffering such as invertebrates (Drosophila, nematode worms, and social amoebae)

(b) **Refinement**: Methods which minimise animal suffering and improve welfare such as the use of anaesthesia and analgesia, use of humane endpoints when death is an expected outcome, controlling the size and growth of tumours, provision of an enriched spacious environment encouraging normal behaviour, withdrawal of water and food for restricted periods, and housing social animals such as mice and rats with other animals.

(c) **Reduction**: Methods which minimise the number of animals per experiment. The reduction also involves robust, reproducible experiments which are appropriately designed and analysed and add to the knowledge.

Cumulative Meta-Analysis for Achieving 'Reduction' Component of the 3Rs

A cumulative meta-analysis may be considered to achieve the 'Reduction' component of the 3Rs (Lau et al. 1992). It consists of a series of meta-analyses where each successive meta-analysis incorporates one additional study. A chronological placement of the meta-analyses displays current evidence and the shift of conclusions over a specified time. Sena et al. conducted a cumulative meta-analysis on experimental studies on stroke and effects of recombinant tissue plasminogen activator (rtPA). Their results showed that the estimate of efficacy was already stable in 2001 (n = 1500 animals). However, another 1888 animals were unnecessarily used after 2001 to establish the effects of rtPA for stroke (Sena et al. 2010a). Using a cumulative meta-analysis graph, a stable treatment effect can be observed at a pre-defined cut off using sufficient data indicating that further animal studies may not be needed.

Establishment of SYRCLE, ARRIVE Guidelines, and the GSPC

The **SY**stematic **R**eview **C**entre for **L**aboratory animal **E**xperimentation (**SYRCLE**) was established in 2008 in Nijmegen, The Netherlands, for evidence-based research in animals as per the guidelines of the Cochrane Collaboration. In 2012, the Dutch parliament recommended compulsory education and training in SRs of animal studies for animal researchers. Ritskes-Hoitinga et al. suggested a collaboration between BSc and MSc curricula designers in Europe and the Cochrane review groups and the routine use of SRs (2014). The 2009 survey (Kilkenny et al. 2009) of reporting of the quality, design and statistical analysis in animal studies led to the publication of the **A**nimal **R**esearch: **R**eporting **I**n **V**ivo **E**xperiments (**ARRIVE**) guidelines, (Kilkenny et al. 2010) which have recently been updated as ARRIVE guidelines 2.0 (Percie du Sert et al. 2020). These guidelines are a checklist of information included in preclinical research publications. They are aimed at researchers, reviewers, journals publishing preclinical research, ethical review boards, funders and institutions. They provide a framework for planning, and conducting animal studies and also for writing and reviewing manuscripts on animal research. The CONSORT guidelines for reporting clinical trials prompted the publication of the **Gold Standard Publication Checklist** (GSPC) for improved reporting of animal research (Hooijmans et al. 2010).

Reducing the Number of Animals Needed for a Study

3Rs-REDUCTION is an online educational program to improve the design of animal research (http://www.3rs-reduction.co.uk/). **Fund for the Replacement of Animals in Medical Experiments (FRAME)** conducts workshops complementary to the 3Rs-REDUCTION program to enhance the use of a minimum number of animals for achieving the primary outcome of a study. The **Reproducibility Initiative** by organisations such as the Science Exchange, PLOS ONE, Figshare and Mendeley helps in identifying and acknowledging good quality experimental studies which have good reproducibility. (https://www.scienceexchange.com/reproducibility) (Hooijmans and Ritskes-Hoitinga 2013).

Funding agencies such as the Dutch ZonMw provide funding for scientists to publish negative or neutral results whereas initiatives such as **REACH** (The European Community regulation on chemicals and their safe use) facilitate data sharing and adherence to the moral principle of **'*More knowledge with fewer animals*.'** Figshare, an online digital repository, enables researchers to preserve and share research output in various forms including figures, datasets, images and videos. The launch of *F1000Research,* an open-access scientific journal in 2012 offers an opportunity to publish, receive peer review and share datasets with other peer groups (Hooijmans and Ritskes-Hoitinga 2013). **SYRCLE** has a registry for systematic reviews of animal studies and has reported a 35% reduction in the number of animals used since commencing SRs of animal studies (Cochrane 2017).

Current Issues with Animal Studies

Without prior SRs for guidance, selecting the most appropriate animal model is challenging. Practical issues (e.g. cost, ease of handling) often override the decision making in determining the optimum model. Rigorous assessment of animal species and the process of disease induction is often not given due consideration. Other issues include the suboptimal and non-standardised methodology including poor internal validity, inherent differences between the designs of clinical and animal studies, insufficient reporting of animal experiments and publication bias (Hooijmans and Ritskes-Hoitinga 2013; Sena et al. 2010b, 2014).

Benefits of Systematic Reviews in Animal Models

SR and meta-analyses of animal studies increase the sample size by pooling the data from included studies, improving the precision and power for assessing the outcomes evaluated. They help in designing future animal studies, selecting animal models, and most importantly, in designing clinical trials. For example, the systematic review of animal studies by van Drongelen et al. showed that the 'Sprague Dawley' rat was the most suitable model to assess changes in the mesenteric arteries in pregnancy-induced hypertension (2012). As discussed earlier, SRs of animal studies improve the quality of future studies by enforcing clear and transparent reporting, and data sharing to optimise the reproducibility of results. Furthermore, they assist in the implementation of the 3Rs (Pound and Ritskes-Hoitinga 2020; Hooijmans and Ritskes-Hoitinga 2013; Sena et al. 2014).

Methods

The essential steps involved in a systematic review of animal studies are shown in Fig. 2.

Table 1 shows the differences in the systematic reviews of clinical versus animal studies.

1. Formulating a focussed research question
2. Preparing and registering a protocol (SYRCLE)
3. Defining exclusion and inclusion criteria
4. Systematic search for original publications across at least two databases
5. Selecting relevant citations
6. Assessing quality/ validity of included studies (if individual studies reported as per ARRIVE guidelines or GSPC adherence) and use SYRCLE ROB tool
7. Extracting data
8. Meta-analysis (data synthesis)
9. Interpretation of results and evaluating level of evidence using GRADE (Grading of Recommendations Assessment, Development and Evaluation)

Fig. 2 Steps of a systematic review of animal studies

Table 1 Comparison of the process of a systematic review

	Clinical studies	Animal studies
Registration of protocol	PROSPERO	SYRCLE, PROSPERO
Literature Search	MEDLINE, Embase, Cochrane, etc.	SYRCLE's step by step guide; Databases: MEDLINE, Embase, ISI Web of Science, Google Scholar, Grey literature
Study characteristics, data extraction	PICOS/PICOT	SYRCLE's method
ROB assessment	Cochrane ROB.2 tool	SYRCLE ROB tool
Publication bias	Funnel plot, Egger's test	Funnel plot
Data synthesis and meta-analysis	Revman-5	Revman-5
Quality of Evidence 40	GRADE	GRADE
Reporting the review	PRISMA	PRISMA

Abbreviations PROSPERO: The International Prospective Register of Systematic Reviews (www.crd.york.ac.uk); SYRCLE: SYstematic Review Centre for Laboratory animal Experimentation; PICOS (T): Participant, Intervention, Comparison, Outcome, Study Design, T (time); ROB: Risk of Bias; GRADE: Grading of Recommendation, Assessment, Development and Evaluation; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Accurate Sample Size Calculation

Calculating an accurate sample size is crucial as a small sample can miss the real effect, and an exceedingly large sample can interfere with 3Rs by wasting resources and animals. 'Power analysis' is the recommended method for calculating sample size for animal experiments (Charan and Kantharia 2013).

Reporting

The ARRIVE guidelines 2.0 are a checklist of information to include in publications describing animal research and ensure transparent and thorough reporting (Kilkenny et al. 2010; Percie du Sert et al. 2020). Since their inception in 2010, the ARRIVE guidelines and associated resources have been used at several steps during the course of a study including the following: (i) ***Study planning***: the guidelines and accompanying Explanation and Elaboration document provide advice on experimental design, minimisation of bias, sample size and statistical analyses, helping researchers to design rigorous and reliable in vivo experiments, (ii) ***Study***

Conduct: this allows researchers to record important information about study methods for manuscript preparation, (iii) ***Manuscript writing***: guidelines used as a memory-aid to ensure inclusion of all relevant information, and (iv) ***Manuscript review:*** to ensure transparency for evaluation and reproducibility of the research (Percie du Sert et al. 2020).

Assessing the Quality of Evidence Using GRADE

Wei et al. explored the use of GRADE (Guyatt et al. 2011) in preclinical SRs and concluded that it was suitable (2016). The principles of considerations on imprecision (95% CI narrow or wide and show minimal or no overlap, optimal information size), inconsistency (point estimates, 95% CI, I2 statistic, p-value) and publication bias (comprehensive assessment) are similar to those in SRs of clinical interventions. Indirectness should be judged by assessing the differences between the **PICO** (**P**articipant, **I**ntervention, **C**omparison, **O**utcome) characteristics and the question of interest. Furthermore, the translational potential of preclinical SRs on clinical trial design or decisions or health policy-related impact should be considered while assessing indirectness (Wei et al. 2016).

A subgroup for GRADE in preclinical SRs was set up by the GRADE working group (https://www.gradeworkinggroup.org/*)* during the 23rd Cochrane Colloquium (October 2015).

Overcoming Limitations of Preclinical Studies

Leenaars et al. have reported that despite their increased number and quality, the agreement between animal and clinical studies was anywhere between 0 to 100% (2019). Following factors make it difficult to translate research from 'bench to bedside':

Unavoidable factors: There are fundamental differences between humans and other species, and within animal species, strains and cell lines. Biological differences should be taken into consideration to improve study design, generate reliable outcomes, reduce expenditure, and implement the 3Rs. For example, antioxidants were shown to be beneficial in animal models of acute ischemic stroke but were harmful in a clinical trial (Macleod et al. 2008). This was attributed to multiple baseline differences in the biological set up of animals and humans and the absence of co-morbidities associated with stroke (e.g. hypertension and diabetes) in animals (O'Collins et al. 2006). Similar contradictory findings have been reported for various conditions (Lucas et al. 2002; Kalra et al. 2002; Lee et al. 2003; Roberts et al. 2002; Mapstone et al. 2003; Tameris et al. 2013; Kashangura et al. 2015).

Avoidable factors: These include poor methodology (neglect of randomisation and blinding, flawed statistical methods, minimal use of sample size calculation), publication bias (overestimation of intervention effects as negative or neutral results are often not published), (Pound and Ritskes-Hoitinga 2020, 2018; Perel et al. 2007) and differences between animal and clinical trial designs (animal studies are usually conducted as a phase one or two projects and use different protocols) (Hyman 2012).

In summary, ensuring robust design, rigorous methodology, transparent reporting, and humane approach is critical for improving the contribution of animal studies to clinical research (Ferreira et al. 2020). SRs of animal studies thus have an important role in designing clinical research (van Luijk et al. 2014; Ritskes-Hoitinga et al. 2014; Symonds and Budge 2018; Bahadoran et al. 2020).

Key Messages

Optimising animal studies
• Systematic review before undertaking an animal study • Selecting evidence-based and most appropriate animal model • Protocol finalisation including sample size and power • Protocol registration (SYRCLE, PROSPERO) • Conduct-Reporting (GSPC and ARRIVE guidelines) • Result interpretation using GRADE guidelines • Think of 'Cumulative meta-analyses'

References

Bahadoran Z, Mirmiran P, Kashfi K, et al. Importance of systematic reviews and meta-analyses of animal studies: challenges for animal-to-human translation. J Am Assoc Lab Anim Sci. 2020. https://doi.org/10.30802/aalas-jaalas-19-000139.

Charan J, Kantharia ND. How to calculate sample size in animal studies? J Pharmacol Pharmacother. 2013;4:303–6.

Cochrane. Cochrane-REWARD prizes for reducing waste: 2017 winners. http://www.cochrane.org/news/cochrane-reward-prizes-reducing-waste-2017-winners. Accessed 11 Aug 2020.

de Vries RB, Wever KE, Avey MT, et al. The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. ILAR J. 2014;55:427–37.

Ferreira GS, Veening-Griffioen DH, Boon WPC, et al. Levelling the translational gap for animal to human efficacy data. Animals (Basel). 2020;10:1199.

Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011;64:380–2.

Hooijmans CR, Ritskes-Hoitinga M. Progress in using systematic reviews of animal studies to improve translational research. PLoS Med. 2013;10.

Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. Altern Lab Anim. 2010;38:167–82.

http://nuffieldbioethics.org/wp-content/uploads/The-ethics-of-research-involving-animals-full-report.pdf. Accessed 14 Aug 2020.

https://3rs.ccac.ca/documents/en/WC8_Declaration_of_Montreal_FINAL.pdf. Accessed 18th July 2020.

Hyman SE. Revolution stalled. Sci Transl Med. 2012;4: 155–11.

Kalra PR, Moon JC, Coats AJ. Do results of the ENABLE (Endothelin Antagonist Bosentan for Lowering Cardiac Events in Heart Failure) study spell the end for non-selective endothelin antagonism in heart failure? Int J Cardiol. 2002;85:195–7.

Kashangura R, Sena ES, Young T, Garner P. Effects of MVA85A vaccine on tuberculosis challenge in animals: SR. Int J Epidemiol. 2015;44:1970–81.

Kilkenny C, Browne W, Cuthill IC, et al. NC3Rs Reporting Guidelines Working Group. Animal research: reporting in vivo experiments: the ARRIVE guidelines. Br J Pharmacol. 2010; 160:1577–9.

Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS ONE. 2009;4:e7824.

Korevaar DA, Hooft L, ter Riet G. Systematic reviews and meta-analyses of preclinical studies: publication bias in laboratory animal experiments. Lab Anim. 2011;45:225–30.

Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med. 1992;327:248–54.

Lee DS, Nguyen QT, Lapointe N, et al. Meta-analysis of the effects of endothelin receptor blockade on survival in experimental heart failure. J Cardiac Fail. 2003;9:368–74.

Leenaars CH, Kouwenaar C, Stafleu FR, et al. Animal to human translation: a systematic scoping review of reported concordance rates. J Transl Med. 2019;17:223.

Lucas C, Criens-Poublon LJ, Cockrell CT, et al. Wound healing in cell studies and animal model experiments by low level laser therapy; were clinical studies justified? ASR Lasers Med Sci. 2002;17:110–34.

Macleod MR, van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke. 2008;39:2824–9.

Mapstone J, Roberts I, Evans P. Fluid resuscitation strategies: a SR of animal trials. J Trauma Acute Care Surg. 2003;55:571–89.

O'Collins VE, Macleod MR, Donnan GA, et al. 1,026 experimental treatments in acute stroke. Ann Neurol. 2006;59:467–77.

Percie du Sert N, Hurst V, Ahluwalia A, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. PLoS Biol. 2020.

Perel P, Roberts I, Sena E, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ. 2007;334:197.

Peters JL, Sutton AJ, Jones DR, et al. A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. J Environ Sci Health B. 2006;41:1245–58.

Pound P, Ebrahim S, Sandercock P, et al. Reviewing animal trials systematically (RATS) Group: Where is the evidence that animal research benefits humans? BMJ. 2004; **328**:514–7.

Pound P, Ritskes-Hoitinga M. Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. J Transl Med. 2018;16:304.

Pound P, Ritskes-Hoitinga M. Can prospective systematic reviews of animal studies improve clinical translation? J Transl Med. 2020. https://doi.org/10.1186/s12967-019-02205-x.

Ritskes-Hoitinga M, Leenaars M, Avey M et al. Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. Cochrane Database Syst Rev. 2014; 3 Art. No.: ED000078. https://doi.org/10.1002/14651858.ed000078.

Roberts I, Kwan I, Evans P, et al. Does animal experimentation inform human healthcare? Observations from a SR of international animal experiments on fluid resuscitation. BMJ. 2002;324:474–6.

Russell WMS, Burch RL. The principles of humane experimental technique. London, UK: Methuen; 1959. p. 238.

Sandercock P, Roberts I. Systematic reviews of animal experiments. Lancet. 2002;360:586.

Sena ES, Briscoe CL, Howells DW, et al. Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: systematic review and meta-analysis. J Cereb Blood Flow Metab. 2010;30:1905–13.

Sena ES, van der Worp HB, Bath PM, et al. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol. 2010b;8.

Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. J Cereb Blood Flow Metab. 2014;34:737–42.

Symonds ME, Budge H. Comprehensive literature search for animal studies may have saved STRIDER trial. BMJ. 2018;362:k4007.

Tameris MD, Hatherill M, Landry BS, et al. Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: a randomised, placebo-controlled phase 2b trial. Lancet. 2013;381:1021–8.

van Drongelen J, Hooijmans CR, Lotgering FK, et al. Adaptive changes of mesenteric arteries in pregnancy: a meta-analysis. Am J Physiol Heart Circ Physiol. 2012;303:H639–57.

van Luijk J, Bakker B, Rovers MM, et al. Systematic reviews of animal studies; missing link in translational research? PLoS ONE. 2014;9:e8998.

Wei D, Tang K, Wang Q, et al. The use of GRADE approach in systematic reviews of animal studies. J Evid Based Med. 2016;9:98–104.