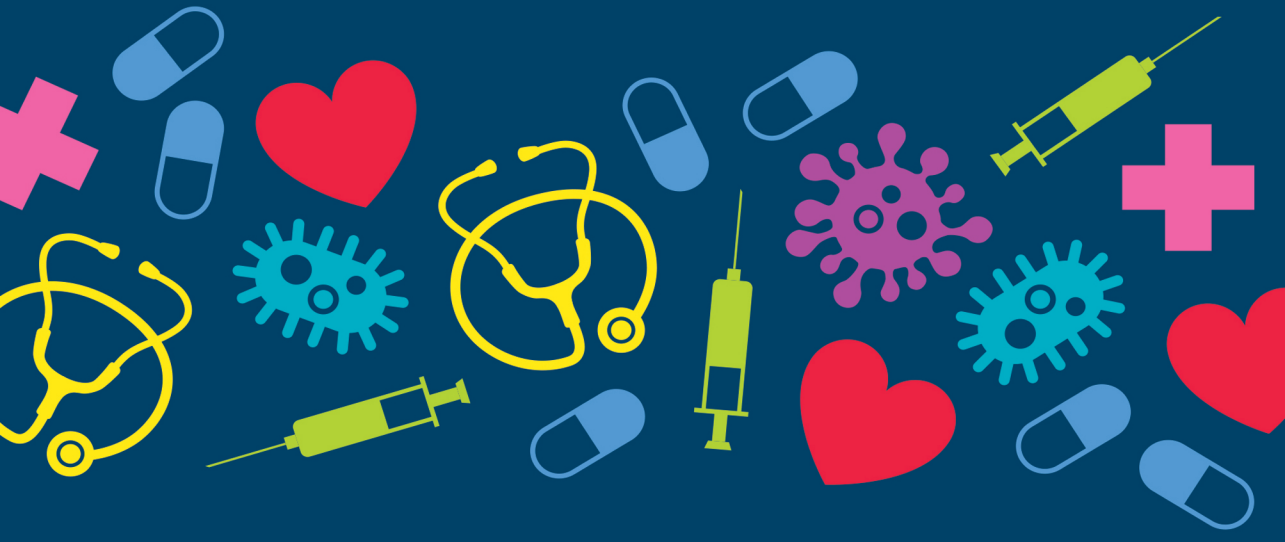
A GUIDE TO PERFORMING SYSTEMATIC REVIEWS OF HEALTH AND DISEASE



KURINCHI GURUSAMY

UCLPRESS

A GUIDE TO PERFORMING SYSTEMATIC REVIEWS OF HEALTH AND DISEASE

A GUIDE TO PERFORMING SYSTEMATIC REVIEWS OF HEALTH AND DISEASE

KURINCHI GURUSAMY

UCL PRESS

UCL Press, the UK's first fully open access university press, has established a new open access textbook programme. Publishing for courses and modules both at UCL and for use at universities globally, it reflects our commitment to widening the use of open educational resources. The rising cost of textbooks, combined with issues of availability, particularly digitally, are a challenge for universities and a potential barrier to student learning. The textbooks in this programme provide benefits to students by removing those barriers and have been commissioned and developed with student use in mind. Open access publishing provides a means to make textbooks accessible to a wider audience, while also reducing the burden on library budgets.

If you are using this textbook to teach a course at your institution, please notify us at uclpresspublishing@ucl.ac.uk so that we can track adoptions and build metrics that evidence how open access textbooks are used and the value they deliver. We welcome your feedback on this book and the programme. Do get in touch to tell us your thoughts.

First published in 2025 by
UCL Press
University College London
Gower Street
London WC1E 6BT

Available to download free: www.uclpress.co.uk

A CIP catalogue record for this book is available from The British Library.

ISBN: 978-1-80008-778-1 (Hbk.)
ISBN: 978-1-80008-659-3 (Pbk.)
ISBN: 978-1-80008-658-6 (PDF)
ISBN: 978-1-80008-660-9 (epub)
DOI: https://doi.org/10.14324/111.9781800086586

Dedicated to all the researchers who worked, are working,
or want to work hard to improve society.

Contents

List of figures

List of tables

List of abbreviations

CI	confidence intervals
COMET	Core Outcome Measures in Effectiveness Trials
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
GCP	good clinical practice
GLMM	Generalised Linear Mixed Model
GRADE	Grading of Recommendations, Assessment, Development and Evaluations
HRQoL	health-related quality of life
ICEMAN	Instrument to assess the Credibility of Effect Modification Analyses
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICJME	International Committee Journal of Medical Editors
ICTRP	International Clinical Trials Registry Platform
IPD	individual participant data meta-analysis
IV	inverse variance method
ln	natural logarithm
MeSH	Medical Subject Headings
MH	Mantel-Haenszel method
MID	minimal important difference
NMA	network meta-analysis
NRSI	non-randomised studies of intervention
OMI	outcome measurement instruments
PICO	population (or participant), intervention, comparator, outcomes
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PRISMA-A	PRISMA extension for acupuncture
PRISMA-E 2012	PRISMA-Equity extension
PRISMA-S	PRISMA literature search extension
PRISMA-ScR	PRISMA extension for scoping reviews
PROMIS®	Patient-Reported Outcomes Measurement Information System
RCT	randomised controlled trial
ROB 2	Risk of Bias in Randomised Trials version 2
ROBINS-I	Risk of Bias in Non-randomised Studies – of Interventions

ROBIS	Risk of Bias in Systematic Reviews
ROB-ME	Risk of Bias due to Missing Evidence in a Meta-analysis tool
SUTVA	stable unit treatment value assumption
SWiM	Synthesis Without Meta-analysis
TACIT	Tool for Addressing Conflicts of Interest in Trials
TIDieR	Template for Intervention Description and Replication
UMLS	Unified Medical Language System
VAS	visual analogue scale
WHO	World Health Organization

Acknowledgements

I am forever indebted to my daughter Neela whose birth changed me from a self-centred person to someone who wants to contribute to society. She unleashed the innovator in me: I learned my innovative thinking by observing how she arrived at ingenious solutions during her early childhood. I look forward to the days when she finishes her education and can focus on her ambitions.

I am very thankful to my parents, who love me despite my neurodivergent behaviour. While it is understandable that my likely neurodivergent father accepted my behaviour most times, it must have been very challenging for my neurotypical mother. But, almost always, she accepted me as I am.

I thank my friends and counsellors, my sister Deepa and Anand. Deepa has put up with my neurodivergent behaviour from a young age. Anand provided me career advice and logical reasoning for my personality profile, which was instrumental in me acknowledging my difficulties and working around them when necessary.

One might expect close family and friends to be patient with neurodivergent behaviour and be well-wishers, but the help I received from many of my work colleagues has been exceptional.

I could not have achieved what I have without the mentorship of Professor Brian Davidson, University College London, who was my PhD supervisor and is currently my boss. It is one thing to tolerate my neurodivergent behaviour after my impactful research, but it is another to tolerate it when I had not achieved much. He also acted as my advocate to my work colleagues and put me on a firm footing to achieve what I have done. He continues to help and support me in fulfilling my potential.

I thank Professor Christian Gluud and managing editor Dimitrinka Nikolova from Copenhagen Trial Unit, Denmark, Mr Martyn Parker, Peterborough and Stamford Hospitals NHS foundation trust, and Professor Steve Morris, University of Cambridge, for their contributions to my achievements.

I also thank my divisional leads Professor Vivek Mudera, Natalie Berry and Professor Marilena Loizidou for having the confidence to appoint me in leadership positions, and all my work colleagues who have put up with my neurodivergent behaviour. At least these days I warn them that I am neurodivergent and society is used to professors being forthright with their comments and challenging the norms (is that not how they become professors?), but a lot of my work colleagues tolerated this behaviour before I knew that I was neurodivergent or was confident enough to reveal it to others. Thank you for your patience and support.

Finally, I would like to thank Katrina Halliday, Andrew Ware and Dhara Snowden from UCL Press for their help, support and advice in making this textbook a reality, and Julie Willis from Westchester Publishing Services UK, project manager for the publication. I also thank Rachel Carter, freelance copy editor, for editing the text and identifying some major typographical errors: these were corrected and prevented major errors in readers' understanding – phew, close shave!

In fond memory of my grandparents, whose love was unconditional and contributed considerably to my happy childhood in addition to being inspirational, and of Professor T. C. Mahendran, Madras Medical College, an inspirational surgeon, and Professor Andy Burroughs, Royal Free Hospital, an inspirational researcher.

1 Introduction to study designs and randomised controlled trials

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- *Explain the main purpose of clinical research from an evidence-based healthcare perspective*
- *List the different study designs*
- *Describe the salient features of different study designs that distinguish them from each other*

OVERVIEW

A clinical study is a research study using human volunteers. The main intention of clinical studies is to add to medical knowledge. The evidence from previous clinical research is a major aspect of informed decision making and may determine the patient's preference for a diagnostic test or treatment, another fundamental element of evidence-based healthcare. Several aspects about a disease can be studied in clinical research and the study design will depend upon the aspect of the disease that needs to be investigated.

Quantitative research can be further divided into observational studies, quasi-experimental studies and experimental studies. In observational studies, the researcher takes a passive role in making measurements on the study subjects. Major types of observational analytical studies are cross-sectional studies, cohort studies and case-control studies.

In experimental studies, the researcher applies an intervention and examines its effect. Randomised controlled trial (RCT) is the main type of experimental study design. The major variations to the standard two-armed

parallel RCT include three-armed parallel RCT, factorial RCT, cluster RCT and cross-over RCT.

Quasi-experimental studies share some features of observational studies, in that the researcher does not apply an intervention, but have some advantages over the observational studies as they consider some errors in interpretation of data inherent in observational data and adjust for these in analysis. The two common quasi-experimental studies to find whether an intervention works are the interrupted time series design and difference-in-differences design.

CLINICAL STUDIES

What are clinical studies?

According to the National Institute of Health, a clinical study is a research study using human volunteers. [1] The main intention of clinical studies is to add to medical knowledge.

Why do we need clinical studies?

Using the best available evidence from systematic research is one of the fundamental elements of evidence-based healthcare. [2] The evidence from previous clinical research is a major aspect of informed decision making and may determine the patient's preference for a diagnostic test or treatment, another fundamental element of evidence-based healthcare. In addition, it is important to understand more about a disease so that government can make appropriate funding decisions in terms of resources available to implement a clinical policy. Even in non-state funded healthcare systems, such as insurance company funded healthcare systems or patient (or patient carer) funded healthcare systems, the results from clinical research may have a major impact on the resources available to perform a treatment. Thus, clinical studies have an impact on all the fundamental elements of evidence-based healthcare. In addition, existing clinical research is often one of the key factors taken into consideration in grant-awarding bodies' funding decisions about future clinical research, irrespective of whether these are government-based grant funding bodies, charity-based funding bodies, local funding bodies or industry-based funding bodies.

STUDY DESIGN

Reason for different study designs

Several aspects of a disease can be studied in clinical research and the study design will depend upon the aspect of the disease that needs to be investigated.

Understanding how frequently a disease occurs will help in the calculation of personal, social and economic impact, that is, the social and economic burden of the disease. Understanding the risk factors for a disease and underlying processes which lead to the disease will help us understand the pathophysiology of the disease and prevent or limit the disease process. Understanding the natural history or natural course of the disease will help us understand the prognosis in terms of whether the disease affects the longevity or quality of life or both. This will help in determining whether there is a necessity to detect the disease early and whether there is any necessity to intervene so that the natural history of the disease can be altered. This will also help us understand the social and economic burden of the disease.

Understanding how to identify a disease is necessary before any attempts are made to alter the natural history of the disease. This might involve understanding the implications of making an early diagnosis or of missing a diagnosis altogether. Comparing the different interventions available for the prevention or treatment of the disease in terms of how safe the intervention is, how effective the intervention is in terms of improving the outcomes that are important to people at high risk of developing the disease or people who have already developed the disease, and the resources required to carry out the intervention will help in making informed decisions.

In this context, there is a subtle difference between the terms 'intervention' and 'treatment'. The term 'intervention' refers to an intervention in people with or without the disease, while the term 'treatment' refers to an intervention in people with disease.

Finally, understanding the prognostic factors will allow the intervention to be tailored to individual patients' needs, an aspect that some people call 'personalised medicine'.

Types of studies

Classification

There are several ways in which research studies can be classified. One way is to classify them as quantitative or qualitative research. In this book, we are interested in quantitative research. Therefore, we will only describe quantitative research further.

Quantitative research indicates research based on information that can be shown in numbers. Quantitative research can be further divided into observational studies, quasi-experimental studies and experimental studies.

In observational studies, the researcher takes a passive role in making measurements on the study subjects. Observational studies can be descriptive studies which examine the distributions of predictors and outcomes in a

population, and analytical studies that find the association between the studies. The major categories of analytical observational studies in the context of clinical (and preclinical) research are cross-sectional studies, cohort studies and case-control studies. [3]

In experimental studies, the researcher applies an intervention and examines its effect. [3] RCT is the main type of experimental study design. [3]

Quasi-experimental studies are another category of studies and include five major study designs: instrumental variables, regression discontinuity, interrupted time series, difference-in-differences, and fixed effects designs. [4] They share some features of observational studies, in that the researcher does not apply an intervention, but have some advantages over the observational studies as they consider some errors in interpretation of data inherent in observational data and adjust for these in analysis. [5]

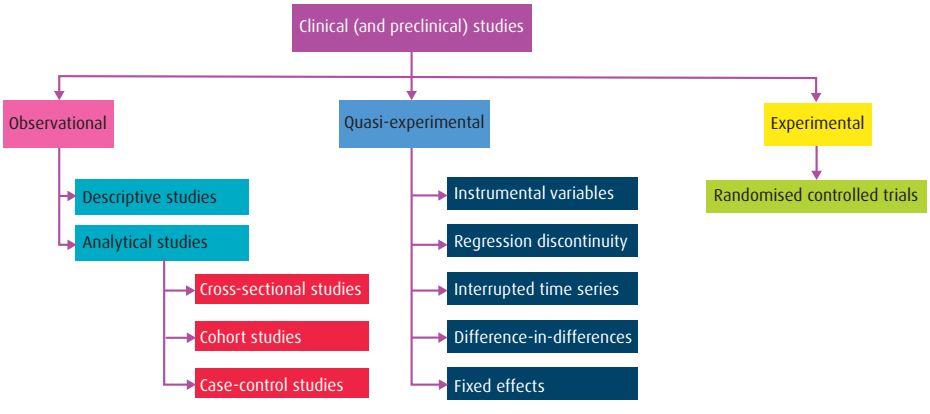
As the main purpose of quasi-experimental and experimental studies is to compare different interventions, experimental studies are all analytical studies. A graphical representation of the types of quantitative research is shown in Figure 1.1.

Among the different study designs, quasi-experimental studies and some cohort and case-control studies which investigate the association between an intervention and outcome are called non-randomised studies of intervention (NRSI). RCTs and NRSI are the focus of this book.

Observational studies

As indicated earlier, in observational studies, the researcher takes a passive role in making measurements on the study subjects. The major types of observational analytical studies are cross-sectional studies, cohort studies and case-control studies. Measurement of both exposure and outcomes is performed. In the

Figure 1.1 Graphical representation of the types of quantitative research



context of investigating the association between an intervention and outcome, exposure indicates 'intervention'. At least two groups differing by exposure are compared in terms of outcomes.

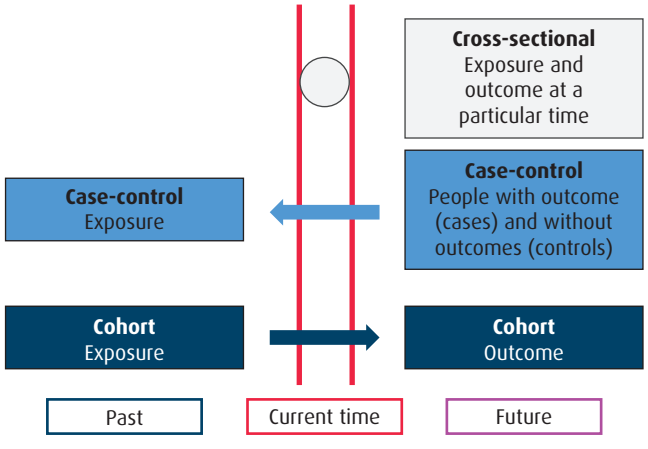
In cross-sectional studies, the exposure and outcomes are measured at a certain point or period (of time). In case-control studies, we start with cases ('*people with outcomes*') and controls ('*people without outcomes*') and measure the exposure in cases and controls. In cohort studies, we start with people without outcomes, measure their exposure, and follow them over time to find out if they develop the outcome.

It is useful to learn some terms in the context of observational studies before learning more about the studies themselves. 'Prospective' indicates collection of information that happens now and in the future. 'Retrospective' indicates collection of already existing information, for example, review of records. A graphical representation of differences in the analytical observational studies is shown in Figure 1.2.

The cohort study design shown in Figure 1.2 is a combined retrospective-prospective cohort study design. In a prospective cohort study design, the cohort exposure happens in the current time or even in the future and subjects are followed in the future. In other words, the exposure and outcome happen further along the timelines than shown in the figure. In retrospective cohort studies, the outcomes in the cohort have already been measured. In other words, the exposure and outcome happen ahead of the timelines shown in the figure.

As only cohort studies and case-control studies are of relevance to intervention reviews, we describe the major features of these two study designs only.

Figure 1.2 Graphical representation of differences in timing of measurements in analytical observational studies



Cohort studies

Cohort means a group of soldiers that marched together (in ancient Roman terms). In the context of clinical research, a cohort is a group of subjects, specified at the outset and followed over time. [3]

A cohort study in which the researcher starts collecting the data now and follows participants over a period (of time) is called a prospective cohort study. On the other hand, in a retrospective cohort study, the data have already been collected. In one sense, you can consider a retrospective cohort study as travelling in a time capsule to a past time to identify a cohort and following them up until current time. However, unlike time travel fiction, you are not able to alter the events; you can only observe. In other words, you cannot influence what data were collected on the individuals and how they were collected. Sometimes, you can get a mixed retrospective-prospective study. This is like a retrospective study except that we do not stop the follow-up of patients at current time; we will follow them in the future too.

Case-control studies

In a case-control study, the researcher begins by choosing one sample of people with the outcome (the 'cases') and another sample of people without that outcome (the 'controls'); they then compare the levels of predictor variables in the two samples to see which predictors are associated with the outcome. [3]

Experimental studies

RCT is the only true experimental study. In experimental studies, the researcher applies an intervention and examines its effect. In the classic RCT design, the researcher begins by assigning a participant to intervention or control randomly (random allocation or random assignment). This is shown in Figure 1.3. The control group could be another intervention or no active treatment. Participants are then followed to see whether they develop the outcome.

There are many variations to the standard design of RCT.

Multi-armed parallel RCT

In a multi-armed parallel RCT, the eligible population is randomised to three or more arms rather than to intervention and control. An example of a three-armed parallel RCT is shown in Figure 1.4. In the figure, the eligible population is randomised to three groups, namely, adjuvant radiotherapy, adjuvant chemotherapy or to no adjuvant treatment, rather than to just two arms. The three groups are then followed to measure the outcomes. Apart from the number of arms, the multi-armed parallel RCTs are conducted in the same way as two-armed parallel RCTs. It is also possible to use a multi-armed parallel RCT to determine whether the effect is greater with increasing dose of the treatment.

Figure 1.3 Standard two-armed parallel RCT

The figure shows that individual participants are randomised to intervention and control groups.

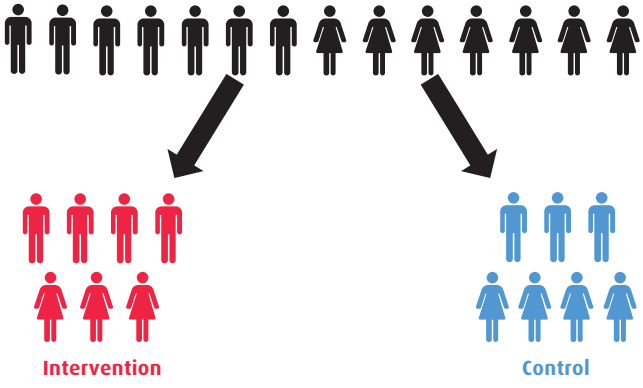
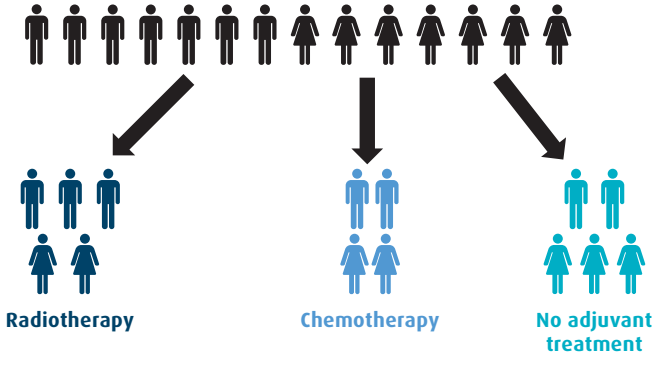


Figure 1.4 Three-armed parallel RCT

The figure shows that the eligible population is randomised to three groups, namely, adjuvant radiotherapy, adjuvant chemotherapy and no adjuvant treatment, rather than to just two arms.

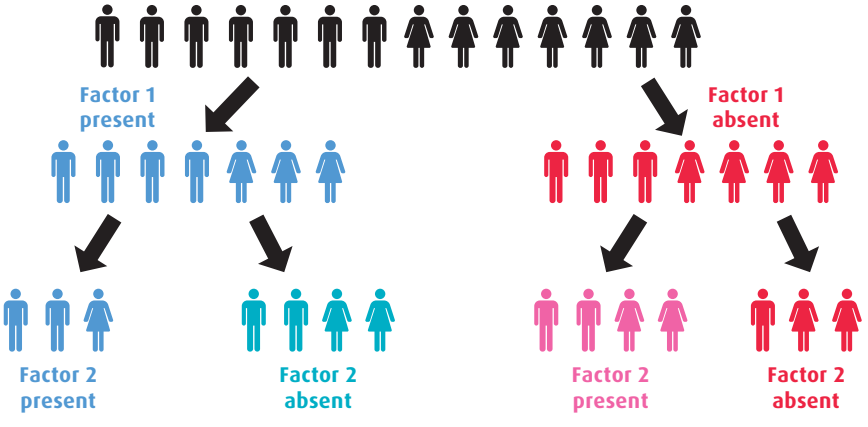


Factorial RCT

In standard two-armed parallel RCTs, only one intervention is compared to control. In multi-armed RCTs, multiple interventions can be compared but we do not know the additive value of the multiple interventions. In the example given previously, we can calculate the effect of adjuvant chemotherapy versus no adjuvant therapy and radiotherapy versus no adjuvant therapy, but we cannot know the effect of the combination of adjuvant radiotherapy plus chemotherapy versus no adjuvant treatment. Now, we could include another arm to the three-armed trial, that is, radiotherapy plus chemotherapy. However, this will increase the sample size even further. An alternative is to conduct a factorial RCT.

Figure 1.5 Factorial RCT

This figure shows that the participants are divided into four groups. One group has both the interventions (or factors) present. The second group has the first factor present but the second factor absent. The third group has the second factor present but the first factor absent. In the fourth group, neither factor is present.



The simplest form of factorial randomised controlled design is shown in Figure 1.5. Here the participants are divided into four groups. One group has both the interventions (or factors) present. The second group has the first factor present but the second factor absent. The third group has the second factor present but the first factor absent. In the fourth group, neither factor is present. This type of trial design is efficient in that it is possible to determine the effect of two factors on an outcome even though the sample size required is the same as that of a simple two-armed parallel trial. [6]

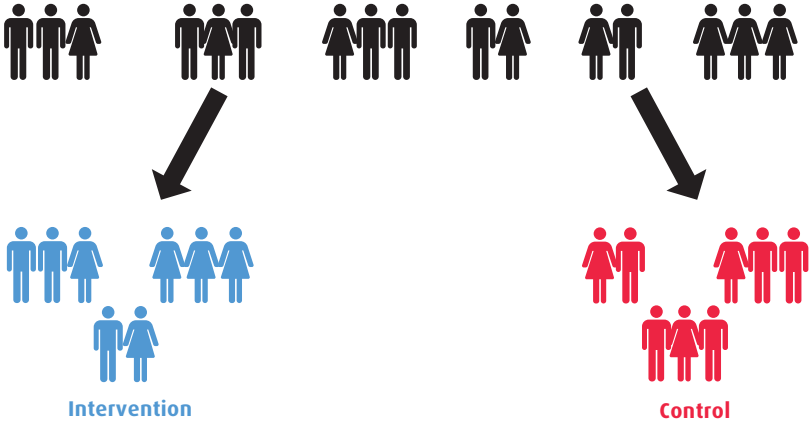
Cluster RCT

In a cluster RCT, rather than randomising individual participants, clusters of participants are randomised (Figure 1.6). First the population is split into clusters. This is usually based on geographical location such as a town, a hospital or a General Practice. Other form of clusters could be family clusters. Each cluster is randomised to receive intervention or control. The outcomes are measured in each member of the cluster.

The reason for cluster RCT is that some interventions can only be studied by a cluster RCT design. For example, nutritional supplementation (fortification) in school diet to prevent iron deficiency anaemia or thyroid hormone deficiency can only be evaluated by cluster RCT as fortification with iron or iodine will be applied to all cooked meals in a school. It is not practical to have two kitchens in the school and randomise individual school children. In this situation, each school will be randomised to fortification or no fortification, and all the children

Figure 1.6 Cluster RCT

In a cluster RCT, rather than randomising individual participants, clusters of participants are randomised.



in a school will receive or not receive the fortified diet depending upon whether the school has been randomised to fortification or no fortification.

Another example is that of an infection control policy, which can only be applied at ward level (for example, different disinfectants are used in different wards to decrease hospital-acquired infections) or even hospital level (for example, routine screening for nasal colonisation of Methicillin-resistant *Staphylococcus aureus* (MRSA) versus no routine screening to decrease MRSA transmission in the hospital).

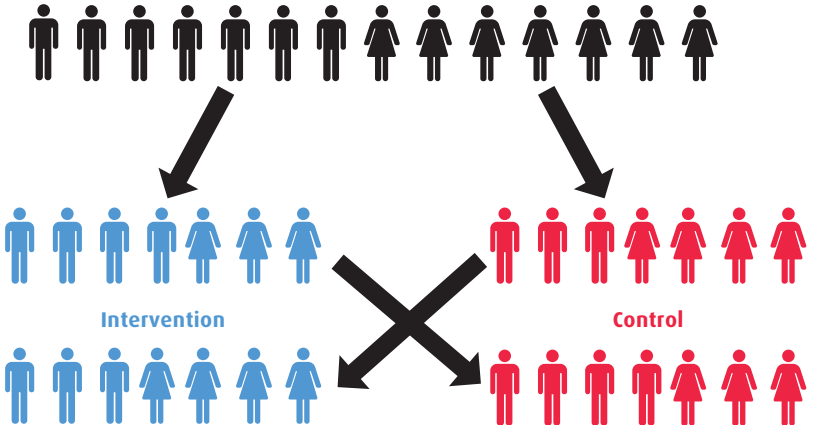
Cross-over RCT

In cross-over RCT, all the participants in the trial receive both the intervention and control. However, the order in which the participants receive the intervention and control is chosen at random as shown in Figure 1.7. Cross-over RCTs are suitable for stable chronic conditions which do not get cured by the treatments. For example, people with diabetes may undergo a cross-over RCT evaluating two different drug-therapy regimens to identify which of these two regimens offers better diabetic control. There must be a wash-out period between cross-over to ensure that there is no residual effect of the first treatment. This wash-out period depends upon the half-life of the drug.

Cross-over RCT is not a suitable study design for diseases which progress or resolve over time, for example, common cold. They are also not suitable for conditions which are cured by the intervention. This is because when the groups cross over, the disease is no longer present, and the effect estimates will be biased. They are also not suitable to study long-term outcomes because it is not possible to determine whether the long-term effects were because of the intervention or control.

Figure 1.7 Cross-over RCT

The figure shows a cross-over RCT, in which all the participants in the trial receive both the intervention and control. However, the order in which the participants receive the intervention and control is chosen at random.



Quasi-experimental studies

The causal inference (that is, the observed difference in the outcome is because of the intervention) for all the types of quasi-experimental studies is based on the assumption that the stable unit treatment value assumption (SUTVA) is met. In other words, the value of an outcome for a unit exposed to the treatment of interest is the same irrespective of the mechanism used to assign the treatment and independent of the treatments that other units receive. [4] This means that the participants should not be assigned to receive intervention or control based on the likelihood of developing the outcome. For example, while evaluating the effect of major surgery versus palliative treatment on survival, assigning a participant with lower cancer burden and fewer comorbidities to surgery and higher cancer burden or more comorbidities to palliative treatment will violate the SUTVA assumption, as the choice of surgery versus palliative treatment is based on the likelihood of survival, as higher cancer burden and more comorbidities are associated with poor survival.

In addition, the 'unconfoundedness' assumption (meaning, there is no confounding) must be met in observational studies to establish causal inference. In quasi-randomised studies, the unconfoundedness assumption is replaced by other (often weaker) assumptions. [4]

However, these assumptions are difficult to test in most types of quasi-experimental studies. The two common quasi-experimental studies to find whether an intervention works are the interrupted time series design and the difference-in-differences design.

Interrupted time series design

In an interrupted time series design, the trend in the outcome before the intervention and after the intervention is compared. For example, you can record the trend in hospital-acquired infections in a hospital by recording the proportion of people admitted to that hospital who developed infections every three months (for two or three years) before introducing an infection control policy. Then you can record the trend in hospital-acquired infections in a hospital by recording the proportion of people admitted to that hospital who developed infections every three months (for two or three years) after introducing the infection control policy. Provided that there has been no change in any other policy, the difference in trends before and after the introduction of infection control policy can be attributed to the new policy. This is depicted in Figure 1.8.

Figure 1.8 Interrupted time series design

The figure shows an interrupted time series design, in which the trend in the outcome before the intervention and after the intervention is compared.

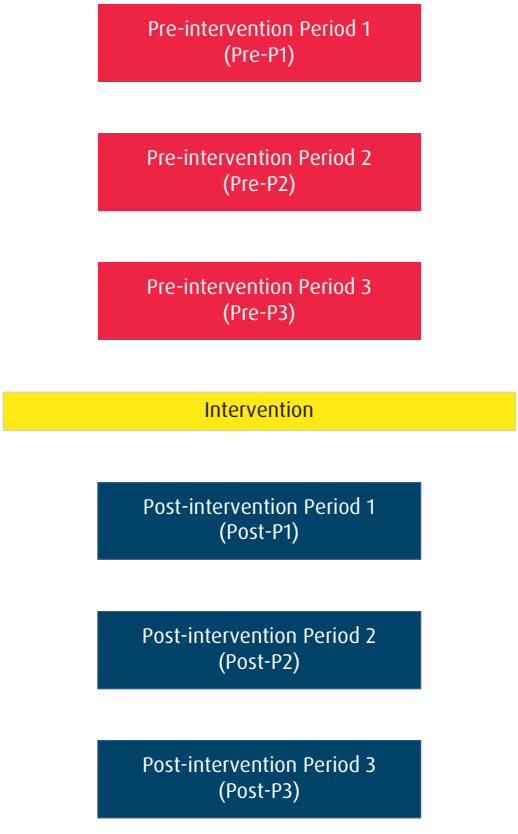
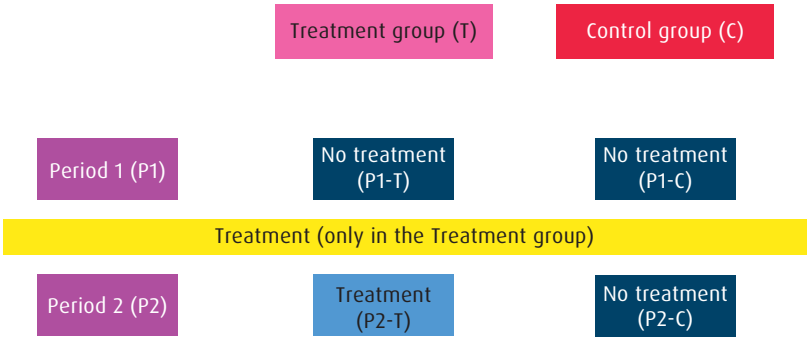


Figure 1.9 Difference-in-differences design

This figure shows what happens in a difference-in-differences study design. In the figure, 'intervention' is called 'treatment'. In the figure, the first difference (D1) is P2-T – P1-T, while the second difference (D2) is P2-C – P1-C. The difference between the first and second differences gives the difference-in-differences, that is, difference-in-differences = D1 – D2.



Difference-in-differences design

Other names for difference-in-differences design include controlled before and after studies, untreated control group design with independent pre-test and post-test samples, and control group design with pre-test and post-test. The logic behind this design is that if an intervention works, we should observe that the outcome improves more in individuals receiving the intervention than in individuals not receiving the intervention (control group). Therefore, two differences are calculated: the first difference is before and after an intervention in the same group of individuals (or an organisation) and the second difference is in the control group between two time periods, the period after which the intervention group received the intervention and period before which the intervention group did not receive the intervention. This is shown in Figure 1.9.

CHAPTER SUMMARY

In this chapter, we learnt about the major interventional study designs that are relevant for the remainder of the book.

PRACTICE QUESTIONS

State True or False for the following statements.

1. A case-control study is an experimental study.
2. In a cohort study, the researcher takes a passive role in making measurements on the study subjects.
3. An RCT is an experimental study.
4. Cohort studies can be prospective or retrospective.
5. RCTs are always prospective.

Answers to the practice questions can be found in the Appendix.

REFERENCES

1. National Institute of Health. Learn About Clinical Studies. https://clinicaltrials.gov/ct2/about-studies/learn (accessed on 21 March 2022). 2019.
2. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996; 312(7023):71–2.
3. Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB. Designing clinical research. Lippincott Williams and Wilkins. 2013.
4. Bärnighausen T, Oldenburg C, Tugwell P, Bommer C, Ebert C, Barreto M, et al. Quasi-experimental study designs series—paper 7: assessing the assumptions. J Clin Epidemiol. 2017; 89:53–66.
5. Bärnighausen T, Røttingen JA, Rockers P, Shemilt I, Tugwell P. Quasi-experimental study designs series—paper 1: introduction: two historical lineages. J Clin Epidemiol. 2017; 89:4–11.
6. Gurusamy KS, Gluud C, Nikolova D, Davidson BR. Design of surgical randomized controlled trials involving multiple interventions. J Surg Res. 2011; 165(1):118–27.

2 Introduction to systematic reviews

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- *Understand what a systematic review is, how it differs from a general review of a topic, and the distinction between systematic review and meta-analysis*
- *Know the types of systematic reviews*
- *Know the general steps in performing a systematic review*
- *Know the general reporting structure of systematic reviews*
- *Understand the advantages of systematic review, including the hierarchy of systematic reviews in the evidence pyramid for finding whether an intervention works*

OVERVIEW

A systematic review attempts to collate all empirical evidence that fits prespecified eligibility criteria to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimising bias, thus providing reliable findings from which conclusions can be drawn and decisions made.

Meta-analysis is the use of statistical techniques to combine the results of included studies. Meta-analysis is not mandatory in systematic reviews and is not appropriate in certain types of systematic reviews. When appropriate, by combining information from all relevant studies, meta-analysis can provide more precise estimates of the effects of healthcare than those derived from the individual studies included within a review.

Systematic reviews can be classified in many ways. One way to classify systematic reviews of health and disease is based on the types of studies included in the systematic review. The types of systematic reviews include incidence

and/or prevalence reviews, aetiology and/or risk factor reviews, diagnostic test accuracy reviews, effectiveness or intervention reviews, prognostic reviews, experiential (qualitative) reviews, expert opinion/policy reviews, costs/economic evaluation reviews, psychometric reviews, methodological reviews and overview of reviews.

Regardless of the type of systematic review, the general steps in a systematic review include **d**etermining the research question; **i**dentifying methods for selecting studies, data extraction, analysis and reporting, and completing the protocol registration; **s**electing the studies; **e**xtracting the data; **a**nalysing the data; **i**nterpreting the results; and **r**eporting the findings (mnemonic: DISEAIR).

Systematic reviews of high-quality randomised trials and n-of-1 trials provide the best evidence to find out whether an intervention works.

WHAT IS A SYSTEMATIC REVIEW?

A systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria to answer a specific research question. [1] This is in comparison with a review, which may restrict the studies included to support the specific viewpoints of the author of the review. Systematic review uses explicit, systematic methods that are selected with a view to minimising bias, thus providing reliable findings from which conclusions can be drawn and decisions made. [1]

The key characteristics of a systematic review include the following. [1]

1. A clearly stated set of objectives with an explicit, reproducible methodology.
2. A systematic search that attempts to identify all studies that would meet the eligibility criteria.
3. An assessment of the validity of the findings of the included studies.
4. Systematic presentation, and synthesis, of the characteristics and findings of the included studies.

How does a general review of a topic differ from systematic reviews?

In contrast to systematic reviews, general reviews of a topic are not objective or reproducible. This results from various factors. There is no definitive search strategy, objective way of inclusion or exclusion of study, or formal way of assessing the validity of included studies. Usually, there is no formal way of combining data. As a result of these factors, a general review of a topic usually reflects the beliefs of the author which are not based on evidence.

META-ANALYSIS

Meta-analysis is the use of statistical techniques to combine the results of included studies. Meta-analysis is not mandatory in systematic reviews. By combining information from all relevant studies, meta-analysis can provide more precise estimates of the effects of healthcare than those derived from the individual studies included within a review. [1]

Can meta-analysis be performed without a systematic review? Yes, it can be performed. There is no difficulty in combining a biased selection of studies to provide a biased result. However, it is inappropriate to do this, as you are highly likely to get biased results by performing a meta-analysis without systematic review.

TYPES OF SYSTEMATIC REVIEWS

Systematic reviews can be classified in many ways. One method of classification of systematic reviews of health and disease is shown in Table 2.1. This is based on a classification system proposed by Munn et al. [2] with some modifications and simply relates to the types of studies included in the systematic review.

A brief description and some examples of the distinct types of systematic reviews are provided below.

Table 2.1 Types of systematic reviews in health and disease

Incidence and/or prevalence reviews
Aetiology and/or risk factor reviews
Diagnostic test accuracy reviews
Effectiveness or intervention reviews
Prognostic reviews
Experiential (qualitative) reviews
Expert opinion/policy reviews
Costs/economic evaluation reviews
Psychometric reviews
Methodological reviews
Overview of reviews

Incidence and/or prevalence reviews

As the name indicates, incidence and/or prevalence reviews are systematic reviews of incidence or prevalence of a health condition or disease. Often these are combined with risk factor reviews. Meta-analysis is usually possible in incidence and prevalence reviews, provided that similar types of people are studied, and similar definitions or methods of diagnosis of the health condition or disease are used in the different studies. Some examples of incidence and prevalence reviews are those by Derman et al. [3] and Petersdorf et al. [4]

Aetiology or risk factor reviews

In aetiology or risk factor reviews, the association between aetiology or risk factors and a health condition or disease is studied. Meta-analysis is usually possible in risk factor reviews, provided that similar types of people are studied, and that similar methods of measuring the risk factors and similar definitions of the health condition or disease are used in the different studies. Some examples of risk factor reviews are those by Ardura-Garcia et al. [5] and Petersdorf et al. [4]

Diagnostic test accuracy reviews

In diagnostic test accuracy reviews, the accuracy of a test in diagnosing a health condition or disease is studied. Meta-analysis is usually possible in diagnostic test accuracy reviews, provided that similar types of people are studied, and similar definitions or methods of diagnosis of the health condition or disease are used in the different studies. Some examples of diagnostic test accuracy reviews are those by Best et al. [6] and Wijedoru et al. [7]

Effectiveness or intervention reviews

Effectiveness or intervention reviews are reviews that evaluate whether an intervention works. Meta-analysis is usually possible in effectiveness or intervention reviews, provided that similar types of people are studied, and similar definitions of outcomes are used in the different studies. In this context, outcome refers to some measure of success, for example, decreased deaths or increased health-related quality of life (HRQoL).

The standard intervention reviews compare only two interventions at a time. There are special types of reviews called network meta-analysis where multiple interventions can be compared simultaneously.

Some examples of intervention reviews are by Kalafateli et al., [8] which is a standard intervention review that compares only two interventions at a time, and by Komolafe et al., [9] which is a network meta-analysis.

The standard intervention reviews include only aggregate data from studies and information provided by contacting the study authors. Individual

participant data review involves re-analysing and combining the data provided by primary study authors. [10] Some examples of individual participant data meta-analysis are those by Lee et al. [11] and Han et al. [12]

Prognostic reviews

In prognostic reviews, the accuracy of a prognostic factor or a prognostic index in predicting the outcome of a health condition, disease or intervention is studied. Meta-analysis is usually possible in prognostic reviews, provided that similar types of people are studied, similar definitions of prognostic factors or indices are applied, and similar definitions of the health condition or disease are used in the different studies. Some examples of prognostic reviews are those by Aalaei-Andabili et al. [13] and Zhou et al. [14]

Experiential (qualitative) reviews

In experiential or qualitative reviews, one or more of beliefs, perspectives, attitudes, barriers, facilitators, experiences, or acceptability of interventions and tests are studied. The list is not exhaustive. Meta-analysis is usually not possible in qualitative reviews. Some examples of qualitative reviews are those by Lake et al. [15] and Li et al. [16]

Expert opinion/policy reviews

In expert opinion or policy reviews, expert opinions or policies are systematically reviewed. Meta-analysis is usually not possible in policy reviews. Some examples of policy reviews are those by Colbert et al. [17] and Just et al. [18]

Costs/economic evaluation reviews

In costs or economic evaluation reviews, costs or economic evaluations are systematically reviewed. Meta-analysis is possible when similar types of participants, interventions, methods to collect costs and utilities (a measure of health) are used in the different studies, but not usually performed because of the variability in these aspects. Some examples of costs and economic evaluation reviews are those by Rocha-Filho et al. [19] and Jorge et al. [20]

Psychometric reviews

In psychometric reviews, measurement properties of an instrument or scale used for measuring a health outcome, say depression or mobility, are studied. Meta-analysis is usually possible when similar types of participants, instruments and measurement properties are used in the different studies. Some examples of psychometric reviews are those by Wiitavaara et al. [21] and Lee et al. [22]

Methodological reviews

In methodological systematic reviews, methods targeted at improving the research methodologies are studied. Meta-analysis is usually possible in the presence of more than one study evaluating the same research methodology. Some examples of methodological reviews are those by Treweek et al. [23] and Lundh et al. [24]

Overview of reviews

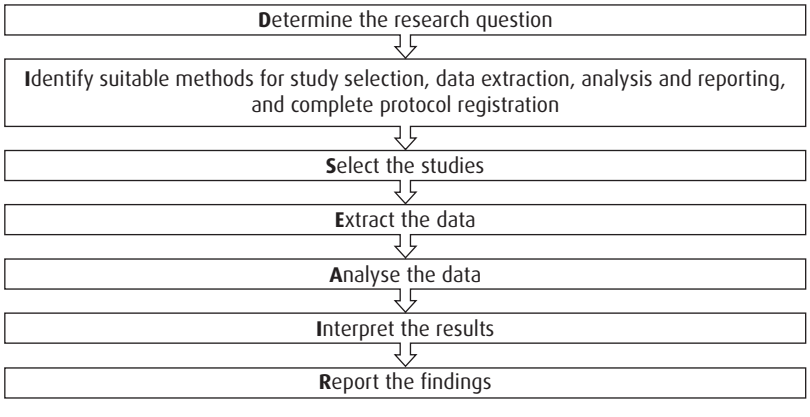
In overview of reviews, existing reviews and systematic reviews are studied. Meta-analysis is usually not appropriate in overview of reviews, as the same participants are likely to be included in multiple systematic reviews of the same topic or the overview includes related but different comparisons. Whichever the case may be, it would be inappropriate to perform the meta-analysis. Some examples of overview of reviews are those by Momsen et al. [25] and Garzón-Orjuela et al. [26]

STEPS IN PERFORMING SYSTEMATIC REVIEWS

Regardless of the type of systematic review, the general steps in performing the review include determining the research question; identifying methods for selecting studies, data extraction, analysis and reporting, and completing the protocol registration; selecting the studies; extracting the data; analysing the data; interpreting the results; and reporting the findings. This is shown in Figure 2.1.

Figure 2.1 General steps in systematic reviews

This figure shows the general steps in performing systematic reviews. The steps can be remembered by the mnemonic 'DISEAIR'.



The steps can be remembered by the mnemonic 'DISEAIR'. A brief description of the steps follows below.

Step 1: Determine the research question

Why do we need to determine the research question as the first step in a systematic review? The research question determines the type and focus of the systematic review. It also helps in other aspects of the conduct and reporting of systematic reviews. The research question format for different types of systematic reviews is available from Munn et al. [2]

Step 2: Identify suitable methods for study selection, data extraction, analysis and reporting, and complete protocol registration

The methods for study selection, data extraction, analysis and reporting depend upon the type of review performed. Appropriate guidance is usually available for some specific types of reviews. If none is available, you must use the guidance for the closest type of reviews. The protocol should be registered before the review is started.

Step 3: Select the studies

A formal search strategy is used to identify studies that meet the eligibility criteria. The formal search strategy includes free text terms and controlled vocabulary terms. Controlled vocabulary terms are 'keywords' to which different terms are mapped. The aim is to retrieve all the eligible studies in most types of systematic reviews. However, in many types of reviews, there is a trade-off between retrieving all the eligible studies versus the feasibility of retrieving them because of the number of references that the searches retrieve from the databases searched.

Study selection should be performed by at least two people independently.

Step 4: Extract the data

Data extraction should be performed by at least two people independently. The data extracted depend upon the type of review but should include data related to the source of information, risk of bias, applicability, the measures that we are interested in, and conflicts of interests in the study.

Step 5: Analyse the data

The data analysis methods depend upon the type of review, but the plan should be made before data are extracted. The analysis should include main analysis, any subgroup analysis, sensitivity analysis and reporting bias.

Subgroup analysis involves comparing results from one subset of studies to another, usually based on clinical or methodological differences. Sensitivity

analysis involves changing the data, for example, excluding some studies based on some feature, or changing the analysis methods to find out if the results are robust to the changes. Reporting bias involves exploring whether some studies have not been reported based on their results.

An example of data extraction and analysis in an intervention review is shown in Figure 2.2.

Step 6: Interpret the results

The interpretation of results depends upon the type of review. Broadly, in systematic reviews of clinical studies, focus should be on clinical impact.

Step 7: Report the findings

General reporting structure of the systematic review

Reporting the findings not only involves reporting the results, but reporting the rationale and methods followed. The general reporting structure of systematic reviews is shown in Table 2.2. This would be broadly similar irrespective of the types of studies included in the systematic review.

Checklist for reporting systematic reviews

The checklist for reporting systematic reviews is the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The current version is PRISMA 2020. [27,28] The latest information about PRISMA is available from www.prisma-statement.org/. In terms of intervention reviews, additional reporting guidance should be followed for network meta-analysis [29] and when no meta-analysis is performed in intervention reviews. [30] Various other extensions are available, which you can use depending on the focus of the review. [31–36]

ADVANTAGES OF SYSTEMATIC REVIEWS

Now that you have been introduced to systematic reviews, it is useful to know their advantages.

One of the major advantages of systematic reviews is that they decrease the risk of random errors. They reduce the chance of false positive and false negative results by combining the data. Systematic reviews are objective and include all studies related to a topic. This allows us to assess the consistency of evidence. By including all the studies related to a topic, systematic reviews provide a balanced view on the topic.

Figure 2.2 Example of data extraction and analysis in an intervention review

The figure shows a forest plot which is usually used for visual representation of meta-analysis results (the results are based on hypothetical data).

Kidney failure

Study	Novel Events	Novel Total	Standard Events	Standard Total	Weight (fixed)	Weight (random)	Odds Ratio MH, Fixed + Random, 95% CI
Avgerinos 1994	5	269	5	298	13.5%	14.7%	1.11 [0.32, 3.88]
Choi 2012a	4	189	2	208	5.4%	7.9%	2.23 [0.40, 12.30]
Choi 2012b	2	188	2	132	6.8%	5.9%	0.70 [0.10, 5.03]
Durand 2017	3	209	4	198	11.8%	10.1%	0.71 [0.16, 3.20]
Garcia-Tsao 2003	5	348	3	352	8.5%	11.1%	1.70 [0.40, 7.15]
Inamdar 2016	4	383	8	381	23.1%	15.7%	0.49 [0.15, 1.65]
Manns 2016	6	327	7	360	19.0%	19.0%	0.94 [0.31, 2.83]
Mullen 2002	8	397	4	361	11.9%	15.7%	1.84 [0.55, 6.15]
Total (fixed effect, 95% CI)		**2310**		**2290**	**100.0%**	**—**	**1.06 [0.66, 1.69]**
Total (random effect, 95% CI)					**—**	**100.0%**	**1.05 [0.59, 1.88]**
Prediction interval							**[0.67, 1.66]**

Heterogeneity: $Tau^2 = 0$; $Chi^2 = 3.97$, df = 7 (P = .78); $I^2 = 0\%$
Test for overall effect (fixed effect): Z = 0.24 (P = .81)
Test for overall effect (random effects): $t_7 = 0.22$ (P = .83)

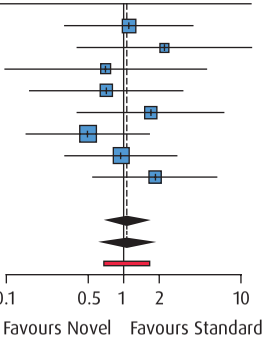


Table 2.2 General reporting structure of systematic reviews

Background
Objectives
Methods • Selection criteria (for including and/or excluding studies) • Identification of studies • Risk of bias/applicability concerns • Data extraction • Data analysis • Reporting bias
Results • Search results • Characteristics of included studies • Bias risk in the studies/applicability concerns in the studies • Meta-analysis results or narrative summary • Investigation of heterogeneity • Sensitivity analysis • Reporting bias
Discussion • Summary of main results • Reliability of the results • Applicability of results • How the findings of this study compare to those of other studies/recommendations/policies on the topic
Conclusions • Guidance for practice • Guidance for research

HIERARCHY OF SYSTEMATIC REVIEWS IN EVIDENCE-BASED HEALTHCARE IN THE EVIDENCE PYRAMID

As this book focuses on intervention reviews, we refer to the hierarchy of intervention systematic reviews in the evidence pyramid based on the Oxford Centre for Evidence-Based Medicine levels of evidence [37] (Figure 2.3). Systematic reviews of RCTs or n-of-1 trials are the highest level of evidence that is currently available. [37] In this context, n-of-1 trials indicate studies where an individual with chronic, stable or slowly progressive disease undergoes different interventions, usually using the cross-over randomised trial design to find what works for them. [38]

Systematic reviews cannot exist without RCTs. So, systematic reviews are not substitutes for RCTs but depend upon high-quality RCTs.

Steps in intervention systematic reviews linked to the book chapters

This book is about intervention reviews. The chapters which cover each step in intervention reviews are shown in Table 2.3.

Figure 2.3 Evidence pyramid

This figure shows the updated Oxford Centre for Evidence-Based Medicine levels of evidence for finding out if an intervention works.

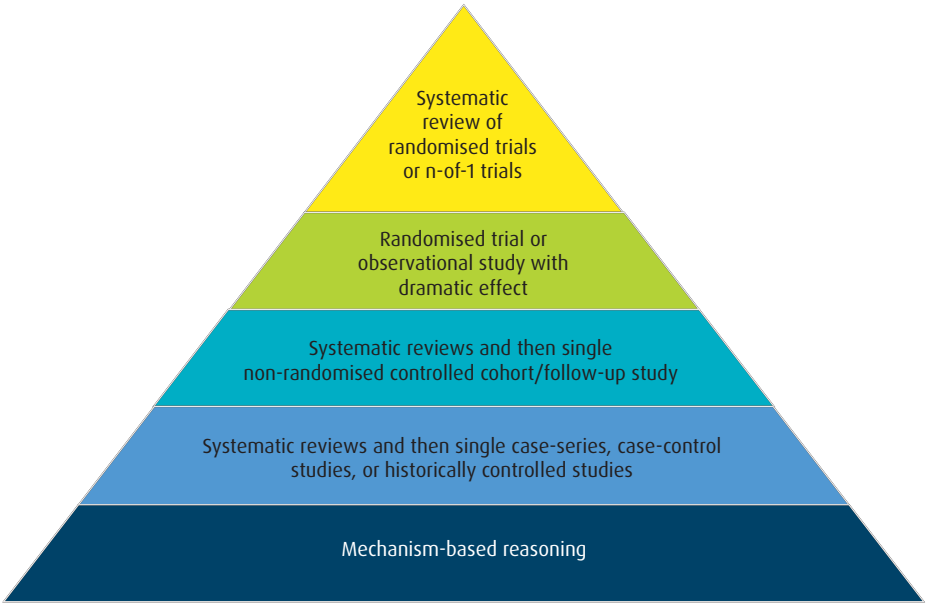


Table 2.3 Steps in intervention systematic reviews linked to the book chapters

Step	Chapter
Determine the research question	Chapter 3
Identify suitable methods for study selection, data extraction, analysis and reporting, and complete protocol registration	These are covered under the specific steps covering these aspects. Protocol registration is covered in Chapter 8 as this would require knowledge of the methods.
Select the studies	Chapter 4
Extract the data	Chapter 5
Analyse the data	Chapter 6
Interpret the results	Chapter 7
Report the findings	Chapter 8

SUMMARY

In this chapter, we have given a brief introduction to systematic reviews and their importance in evidence-based healthcare.

PRACTICE QUESTIONS

State True or False for the following statements.

1. Qualitative systematic reviews usually contain meta-analysis.
2. It is appropriate to perform a meta-analysis without performing a systematic review.

Answers to the practice questions can be found in the Appendix.

REFERENCES

1. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, 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(7): e1000100.
2. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Method. 2018; 18(1):5.
3. Derman W, Badenhorst M, Eken MM, Ezeiza-Gomez J, Fitzpatrick J, Gleeson M, et al. Incidence of acute respiratory illnesses in athletes: a systematic review and meta-analysis by a subgroup of the IOC consensus on 'acute respiratory illness in the athlete'. Br J Sports Med. 2022; 56(11):630–8.
4. Petersdorf K, Groettrup-Wolfers E, Overton PM, Seitz C, Schulze-Rath R. Endometrial hyperplasia in pre-menopausal women: a systematic review of incidence, prevalence, and risk factors. Eur J Obstet Gynecol Reprod Biol. 2022; 271:158–71.
5. Ardura-Garcia C, Garner P, Cooper PJ. Is childhood wheeze and asthma in Latin America associated with poor hygiene and infection? A systematic review. BMJ Open Respir Res. 2018; 5(1): e000249.

6. Best LM, Takwoingi Y, Siddique S, Selladurai A, Gandhi A, Low B, et al. Non-invasive diagnostic tests for Helicobacter pylori infection. Cochrane Database Syst Rev. 2018; 3(3): Cd012080.
7. Wijedoru L, Mallett S, Parry CM. Rapid diagnostic tests for typhoid and paratyphoid (enteric) fever. Cochrane Database Syst Rev. 2017; 5(5): Cd008892.
8. Kalafateli M, Buzzetti E, Thorburn D, Davidson BR, Tsochatzis E, Gurusamy KS. Pharmacological interventions for acute hepatitis C infection. Cochrane Database Syst Rev. 2018; 12(12): Cd011644.
9. Komolafe O, Buzzetti E, Linden A, Best LM, Madden AM, Roberts D, et al. Nutritional supplementation for nonalcohol-related fatty liver disease: a network meta-analysis. Cochrane Database Syst Rev. 2021; 7(7): Cd013157.
10. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010; 340: c221.
11. Lee B, Turner S, Borland M, Csonka P, Grigg J, Guilbert TW, et al. Efficacy of oral corticosteroids for acute preschool wheeze: a systematic review and individual participant data meta-analysis of randomised clinical trials. Lancet Respir Med. 2024; 12(6):444–56.
12. Han B, Chen Y, Liang W, Yang Y, Ding Z, Yin P, et al. Is the interspinous process device safe and effective in elderly patients with lumbar degeneration? A systematic review and meta-analysis of randomized controlled trials. Eur Spine J. 2024; 33(3):881–91.
13. Aalaei-Andabili SH, Alavian SM. Regulatory T cells are the most important determinant factor of hepatitis B infection prognosis: a systematic review and meta-analysis. Vaccine. 2012; 30(38):5595–602.
14. Zhou Y, Dai M, Zhang Z. Prognostic significance of the Systemic Immune-Inflammation Index (SII) in patients with small cell lung cancer: a meta-analysis. Front Oncol. 2022; 12:814727.
15. Lake AJ, Neven ACH, Williams A, O'Reilly SL, Hendrieckx C, Morrison M, et al. Barriers to and enablers of type 2 diabetes screening among women with prior gestational diabetes: a systematic review update and qualitative synthesis applying the Theoretical Domains Framework. Diabet Med. 2022; 39(4): e14772.
16. Li H, Glecia A, Arisman K, Peternelj-Taylor C, Holtslander L, Leidl D. Mental health and addiction related emergency department visits: a systematic review of qualitative studies. Community Ment Health J. 2022; 58(3):553–77.
17. Colbert S, Wilkinson C, Thornton L, Feng X, Richmond R. Online alcohol sales and home delivery: an international policy review and systematic literature review. Health Policy. 2021; 125(9):1222–37.

18. Just D, Tai S, Palmier-Claus J. A systematic review of policy and clinical guidelines on positive risk management. Ment Health. 2021:1–12.

19. Rocha-Filho CR, Martins JWL, Lucchetta RC, Ramalho GS, Trevisani GFM, da Rocha AP, et al. Hospitalization costs of coronaviruses diseases in upper-middle-income countries: a systematic review. PLoS One. 2022; 17(3): e0265003.

20. Jorge SP, David NV, Alba GC, Germán VR, Antonio CJ. Health economic evaluation of exercise interventions in people over 60 years old: a systematic review. Exp Gerontol. 2022; 161:111713.

21. Wiitavaara B, Heiden M. Content and psychometric evaluations of questionnaires for assessing physical function in people with low back disorders: a systematic review of the literature. Disabil Rehabil. 2020; 42(2):163–72.

22. Lee J, Lee EH, Moon SH. Systematic review of the measurement properties of the Depression Anxiety Stress Scales-21 by applying updated COSMIN methodology. Qual Life Res. 2019; 28(9):2325–39.

23. Treweek S, Pitkethly M, Cook J, Fraser C, Mitchell E, Sullivan F, et al. Strategies to improve recruitment to randomised trials. Cochrane Database Syst Rev. 2018; 2(2): Mr000013.

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

25. Momsen AH, Ørtenblad L, Maribo T. Effective rehabilitation interventions and participation among people with multiple sclerosis: an overview of reviews. Ann Phys Rehabil Med. 2022; 65(1):101529.

26. Garzón-Orjuela N, Samacá-Samacá DF, Luque Angulo SC, Mendes Abdala CV, Reveiz L, Eslava-Schmalbach J. An overview of reviews on strategies to reduce health inequalities. Int J Equity Health. 2020; 19(1):192.

27. 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: n71.

28. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021; 372: n160.

29. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, 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(11):777–84.

30. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020; 368: l6890.

31. Welch V, Petticrew M, Petkovic J, Moher D, Waters E, White H, et al. Extending the PRISMA statement to equity-focused systematic reviews (PRISMA-E 2012): explanation and elaboration. J Clin Epidemiol. 2016; 70:68–89.

32. Welch V, Petticrew M, Tugwell P, Moher D, O'Neill J, Waters E, et al. PRISMA-Equity 2012 extension: reporting guidelines for systematic reviews with a focus on health equity. PLoS Med. 2012; 9(10): e1001333.

33. Zorzela L, Loke YK, Ioannidis JP, Golder S, Santaguida P, Altman DG, et al. PRISMA harms checklist: improving harms reporting in systematic reviews. BMJ. 2016; 352: i157.

34. Wang X, Chen Y, Liu Y, Yao L, Estill J, Bian Z, et al. Reporting items for systematic reviews and meta-analyses of acupuncture: the PRISMA for acupuncture checklist. BMC Complement Altern Med. 2019; 19(1):208.

35. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018; 169(7):467–73.

36. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Individual Participant Data: the PRISMA-IPD statement. JAMA. 2015; 313(16):1657–65.

37. OCEBM Levels of Evidence Working Group (Oxford Centre for Evidence-Based Medicine). The Oxford Levels of Evidence 2. www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence (accessed on 21 March 2022). 2011.

38. Kravitz R, Duan N (eds), DEcIDE Methods Center N-of-1 Guidance Panel. Design and Implementation of n-of-1 Trials: A User's Guide. https://effectivehealthcare.ahrq.gov/products/n-1-trials/research-2014-5 (accessed on 22 March 2022). 2014.

3 Determine the research question

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- *Justify the need to determine the research question as the first step in a systematic review*
- *Determine a suitable research question that is important to stakeholders and formulate it in such a way that allows further steps in a systematic review of intervention*
- *Develop the eligibility criteria for the systematic review*
- *Choose appropriate outcomes*

OVERVIEW

Determining the research question is the first step in the systematic review. The research question determines the type and focus of the review. It also helps in other aspects of the conduct and reporting of systematic reviews.

While determining the research question, you should choose a research topic which is important to stakeholders (patients, clinicians, healthcare funders), meets an unmet need, is valuable (that is, adds value to existing knowledge), and is feasible with the available resources.

Some ways of identifying topics that are important to stakeholders include choosing research topics identified as research priorities by patients and healthcare providers in the James Lind Alliance Priority Setting Partnerships, journal publications, research recommendations from recent clinical practice guidelines, a call by funders to address a specific research question, group discussions involving patients and healthcare providers, or debate in a journal. Whichever method you choose, we recommend involving patients in identifying priorities rather than basing this solely on an expert panel.

The research question determines the eligibility criteria in a systematic review. The suggested format for eligibility criteria for an intervention review is the population, intervention, comparator and outcomes (PICO) format. However, we recommend some mechanism for identifying the studies that meet all eligibility criteria other than 'outcomes' to allow the exploration of reporting biases. Study design is also commonly used as a criterion for systematic reviews.

While choosing the population, unless there is a strong reason, restriction should not be based on gender or ethnicity. When choosing outcomes, choose those that are important to stakeholders. The Core Outcome Measures in Effectiveness Trials (COMET) initiative can help with the choice of outcomes, while the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative can help with choosing the best way of measuring outcomes. Patient-Reported Outcomes Measurement Information System (PROMIS®) health measures are validated outcome measurement instruments. Use of surrogate outcomes and composite outcomes requires adequate justification.

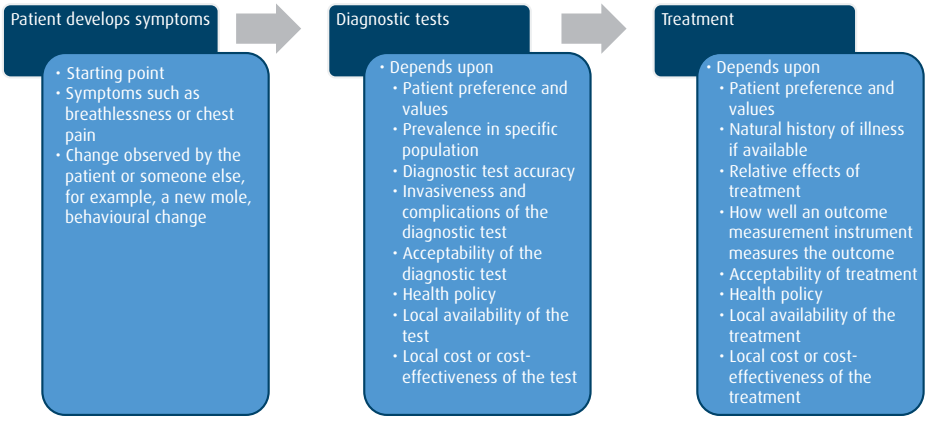
WHY DO WE NEED TO DETERMINE THE RESEARCH QUESTION AS THE FIRST STEP IN A SYSTEMATIC REVIEW?

You may recollect from Chapter 2 that determining the research question is the first step in a systematic review. We briefly mentioned in Chapter 2 that determining the research question determines the type and focus of the systematic review and that it helps in other aspects of the conduct and reporting of systematic reviews. Here we provide more information about how this is achieved.

Typical clinical pathway

To understand how the research question determines the type and focus of the systematic review, it is important to understand the typical clinical pathway that a patient undergoes when they approach a healthcare provider. A typical clinical pathway in the context of illness is shown in Figure 3.1. Briefly, a patient develops symptoms such as breathlessness or chest pain or a change which could be observed by the patient or someone else, for example, a new mole, or behavioural change. Diagnostic tests are then performed. In this context, symptoms and simple clinical examination can be considered diagnostic tests. Additional diagnostic tests may be performed, which depends upon various factors such as patient preference and values, prevalence in specific population (that is, in the subset of population similar to the patient in terms of demographic features, symptoms and signs), diagnostic test accuracy, the invasiveness, complications and acceptability of the diagnostic test, health policy, local availability of the test, and the local cost of the test or cost-effectiveness of the test. If the patient is found to have the health

Figure 3.1 Typical clinical pathway in the context of illness



condition or illness, treatment may be performed. The treatment carried out depends upon various factors that include patient preference and values; natural history of illness if available (what happens without treatment of illness or symptoms); relative effects of treatment (which may vary with different patient-related characteristics such as age or birth sex, disease-related prognostic factors such as the symptoms, signs, results of diagnostic tests including severity of the disease or extent of cancer, and healthcare provider-related prognostic factors such as the setting – for example, community-based care, primary care, secondary care, specialty centres – and the outcomes of previous similar patients treated by the healthcare provider); how well an outcome measurement instrument measures the outcome; acceptability of treatment; health policy; local availability of the treatment; and local cost or cost-effectiveness of the treatment.

The contexts in population health aimed at preventing an illness or complications of an illness and in screening the population at high risk of developing an illness are slightly different: people may not have symptoms, but the remaining aspects such as diagnosis and interventions (which is a more acceptable term than treatment when the context is preventing rather than treating an illness), and the factors that determine these, are broadly along the same principles as in the context of illness.

LINKING THE TYPE OF SYSTEMATIC REVIEW TO THE TYPICAL CLINICAL PATHWAY

It is clear from the clinical pathway in the context of illness that several types of information are required to allow the development of an evidence-based clinical

Table 3.1 Types of systematic reviews for different information

Information required	Types of studies that provide this information	Type of systematic review
Prevalence in specific population	Cross-sectional study	Prevalence review
Natural history of illness if available	Cohort study	Incidence review
Risk factors for development of illness	Case-control study Cohort study	Risk factor review
Risk factors for poor prognosis	Case-control study Cohort study	Prognostic review
Diagnostic test accuracy	Diagnostic test accuracy study	Diagnostic test accuracy review
Invasiveness and complications of diagnostic test	Invasiveness, complications: Diagnostic test accuracy study	Part of diagnostic test accuracy review
Relative effects of treatment	RCT Non-randomised study of intervention	Intervention review
How well an outcome measurement instrument measures the outcome	Psychometric properties study	Psychometric review
Patient preference and values, and acceptability of diagnostic test and treatment	Qualitative research study Surveys	Qualitative review for qualitative research studies Systematic review and meta-analysis of proportions (similar methods as prevalence reviews) for surveys
Costs and cost-effectiveness of tests and treatments	Costs study Cost-effectiveness study	Costs/economic evaluation review
Comparison of health policies with other regions or countries	Policy	Policy reviews

pathway. The types of systematic reviews for distinct types of information required are different, as shown in Table 3.1.

Other aspects of the conduct and reporting of systematic reviews

The methods used for distinct types of reviews, including the search strategies, types of studies included, data extracted, methods used for analysis and interpretation, and reporting differ in different types of systematic reviews. In this book, we will be focusing on these aspects in systematic reviews of interventions.

CHOOSING A RESEARCH QUESTION OF IMPORTANCE TO STAKEHOLDERS

Why is it important to choose a research question of importance to stakeholders?

An estimated 85% of biomedical research is waste. [1] A major reason for this is choosing the wrong research questions, that is, questions that are not considered important by patients and clinicians. [1] Therefore, to avoid research waste any research question should be important to stakeholders (patients, clinicians, healthcare funders), meet an unmet need, be valuable (add value to existing knowledge) and be feasible with the available resources.

Identifying research questions important to stakeholders

Some ways of identifying topics that are important to stakeholders include choosing research topics identified as research priorities by patients and healthcare providers in the James Lind Alliance Priority Setting Partnerships; [2] journal publications, for example, public research priorities for fungal diseases [3] or research priorities in prehabilitation for patients undergoing cancer surgery; [4] research recommendations from recent clinical practice guidelines; [5] a call by funders to address a specific research question; group discussions involving patients and healthcare providers; or debate in a journal. Whichever method you choose, we recommend involving patients in identifying priorities rather than basing this solely on an expert panel.

Establishing the unmet need

Once you have identified some research topics of importance to stakeholders, you can perform a scoping search. This is an informal search to check for any preexisting systematic reviews and potential articles addressing the issue. You can search for any ongoing systematic reviews by searching the PROSPERO database in addition to searching a medical journals database such as MEDLINE. If you find multiple studies that potentially meet the inclusion criteria but no systematic reviews, then consider performing a systematic review. If you find multiple studies that potentially meet the inclusion criteria but there is an existing systematic review, you can perform a systematic review if you find new studies or use different (improved) methods that have the potential to alter the findings of the existing systematic review. [6] An issue related to the improvement in the methods of systematic reviews is to find out if the existing systematic review has been performed without bias. This assessment about the bias in the existing systematic reviews can be done in an objective manner using the 'ROBIS' ('Risk of Bias in Systematic Reviews') tool. [7]

In the absence of multiple studies, consider performing primary research. You might also consider primary research when an existing systematic review indicates that a primary research study is required to answer the research question. However, before performing a primary research study, you might want to search the World Health Organization International Clinical Trials Registry platform and ClinicalTrials.gov trial registers to check if there are ongoing clinical trials that might provide the answer to the research question that you identified. If the search does not reveal any ongoing research study on the topic, or if the ongoing research studies do not address the research questions completely, you can assess the feasibility of performing the primary research study. If the systematic review does not indicate that a primary research study is required and there is moderate or high certainty evidence about the effectiveness of an intervention, you might want to develop clinical practice guidelines if none exist based on up-to-date systematic reviews and methods used for clinical practice guidelines.

Performing a study that is valuable

While performing a research study, it is important to design and conduct the study in such a way that it provides valuable information. This involves choosing methods that provide an unbiased answer to a research question and outcomes that are important to stakeholders. Appropriate follow-up periods are also necessary while performing research. It is obvious that any valuable study should be conducted in an ethical manner.

For example, while performing a systematic review, it is not valuable to search only one database because it is easier to complete. Also, when multiple non-randomised studies show that there is potential for an intervention to work, it is not valuable to perform more non-randomised studies simply on the basis that it is easier to perform such studies than unbiased RCTs. Similarly, avoid choosing outcomes for systematic reviews on the basis that they are reported in primary research studies, and for primary research studies on the basis that they are easy to measure. Outcomes should be chosen on the basis that they are important to stakeholders. You must also avoid a short follow-up period that is meaningless, for example, comparing the beneficial cardioprotective effect of aspirin after one month or measuring the cosmesis of an incision at 24 hours.

Adequate resources

You must ensure that there are adequate resources to perform the research. This includes the knowledge and skills to perform the research, adequate funding and time to complete the research. In primary research studies, you should also consider whether you will be able to obtain the ethical and regulatory approvals, recruit sufficient participants, and follow them adequately in the time that you plan for the project.

What can you do when your initial research plan is deficient in one or more aspects of unmet need, being valuable and feasible?

When the research plan is deficient in one or more aspects addressing an unmet need and being valuable and feasible, it is reasonable to revise your research plan iteratively, until the research project addresses an unmet need, and becomes valuable and feasible.

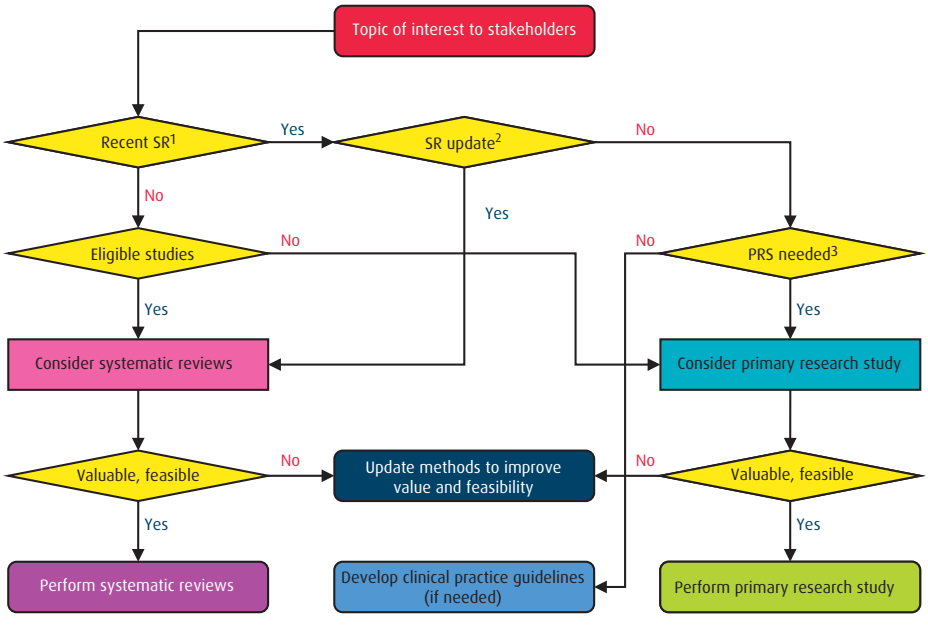
Overview

An overview of the process of choosing a research question of importance to stakeholders is shown in Figure 3.2.

FROM 'RESEARCH QUESTION' TO 'OBJECTIVES'

The research question should guide the objectives of the research. In an intervention review, the primary objective should include details of the population,

Figure 3.2 Choosing a research question of importance to stakeholders: an overview



[1]SR = systematic reviews

[2]SR update is needed if new studies become available or different (improved) methods can be used

[3]PRS = primary research study

intervention, comparator and outcomes (PICO). [8–10] However, this format is suitable when only two interventions are compared at a time and not when multiple interventions are compared simultaneously.

We start with some examples.

Example 1

Here is an example of the primary objective in a standard systematic review of intervention involving comparison of two interventions.

> Evaluate the effectiveness of school-based dental screening versus no screening on improving oral health in children aged 3–18 years by a systematic review and meta-analysis of randomised controlled trials. [11]

In this example, the population is *'school children aged 3–18 years'*, the intervention is *'school-based dental screening'*, the comparator is *'no screening'*, and the outcome is *'oral health'*. Here, although there is no explicit statement in the objective that it is school children, it is reasonable to assume that the population is *'school children (aged 3–18 years)'* because of the nature of the intervention.

Example 2

Another example of the primary objective in a standard systematic review of intervention involving comparison of two interventions is provided below.

> To assess the benefits and harms of laparoscopic distal pancreatectomy versus open distal pancreatectomy for people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma of the body or tail of the pancreas, or both. [12]

In this example, the population is *'people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma of the body or tail of the pancreas, or both'*, the intervention is *'laparoscopic distal pancreatectomy'*, the comparator is *'open distal pancreatectomy'*, and the outcomes are *'benefits and harms'*.

You might note that the outcomes are not as specific in the second example as the first example. This non-specific way of mentioning the outcomes in the objectives is perfectly acceptable.

Additional details in the objectives

While not mandatory in the objectives, it is useful to provide details of the major aspects of the research methods, to give the readers a preliminary idea of the approach that you plan to use. In Example 1, it is clear that you would include RCTs only and that you would perform a meta-analysis; in Example 2, the details

of what type of studies will be included or whether a meta-analysis will be performed are not available.

Alternative format for network meta-analysis

While we suggest the PICO format for standard systematic review of intervention involving comparison of two interventions, we suggest a slightly different format for network meta-analysis in which multiple interventions are compared simultaneously. As a result of comparing multiple interventions simultaneously, it is not possible to specify the intervention and comparator clearly. An example of the primary objective in a network meta-analysis is provided here.

> To assess the benefits and harms of different nutritional supplements for treatment of NAFLD through a network meta-analysis. [13]

In this example, the population is *'people with NAFLD (nonalcohol-associated fatty liver disease)'* and the outcomes are *'benefits and harms'*; however, there is no clearly defined intervention and comparator, apart from *'different nutritional supplements'*.

Secondary objectives

The secondary objectives in a systematic review relate to the use of the data collected as part of the systematic review for reasons other than the primary objectives. In an intervention review, this typically involves the evaluation of the effectiveness of an intervention in a subset of population, comparative effectiveness of a subtype of intervention or comparator, and the effectiveness of an intervention on different outcomes or different ways of measuring an outcome.

When an intervention might work differently in different subsets of populations (for example, people with severe disease or people with disadvantaged backgrounds), you might evaluate the effectiveness of the intervention in a subset of population. If one variation of an intervention or comparator might work differently (for example, different drugs belonging to the same family of drugs, different doses of drugs, or delivery of the intervention by different groups of people such as doctor-led intervention versus nurse-led intervention), you might evaluate the comparative effectiveness of a subtype of intervention or comparator. In cases where there are important secondary outcomes, you might be interested in evaluating the effectiveness of the intervention on such secondary outcomes. And when there are different ways of measuring an outcome, for example, when the observed effect of an intervention might be different when the doctor measures the outcome compared to that reported by the participants themselves, you might evaluate the effectiveness of an intervention on the different ways of measuring an outcome.

FROM 'OBJECTIVES' TO 'ELIGIBILITY CRITERIA'

'Eligibility criteria' refers to which types of studies you would include and exclude from a systematic review. In intervention reviews, as for objectives, the eligibility criteria are also described using the PICO format. [10] However, more details about the population, intervention, comparator and outcomes should be provided in the eligibility criteria.

Population

You need to specify the types of populations who will be included in the review. There is usually restriction (for inclusion in the systematic review) by the disease or health condition. In Example 2, the population of interest for the systematic review is *'people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma of the body or tail of the pancreas, or both'*. There may be restriction by age or other characteristics, as in Example 1, where the population of interest for the systematic review is *'school children aged 3–18 years'*.

In the context of using diseases to define eligibility criteria related to population, you might consider using the 'International Statistical Classification of Diseases and Related Health Problems' (current version: 11th revision (ICD-11)). [14] Other internationally recognised nomenclature, such as the National Library of Medicine's Medical Subject Headings (MeSH) thesaurus [15] or SNOWMED CT [16] are alternative ways of using consistent nomenclature to describe the population. Mapping of SNOWMED CT to ICD-10 is available through Unified Medical Language System® (UMLS®). [17] The use of such terms will enable the end users of the systematic reviews, for example, health professionals, clinical guideline developers, policy makers, researchers and patients, to identify relevant systematic reviews quickly and help with artificial intelligence guided updates of systematic reviews.

In terms of using characteristics of the population as eligibility criteria, you might consider patient-related prognostic factors, such as age or birth sex, and disease-related prognostic factors described in the 'Typical clinical pathway' systematically. These include the symptoms, observed changes, clinical signs, the diagnostic tests used for diagnosing the health condition or disease, and other disease-related prognostic factors that potentially affect the effects of treatment, for example, the severity of disease.

It is highly desirable to describe the effect of an intervention not only upon the whole population but also on the disadvantaged, to potentially reduce socioeconomic inequalities in health. [9] Disadvantaged people can be identified using the PROGRESS-PLUS framework, which refers to place of residence, race/ethnicity/culture/language, occupation, gender/sex, religion, education, socioeconomic status, and social capital, along with any additional context-specific

factors, for example, age, disability, parental education, and so on. [18] Therefore, unless there is a strong reason, the restriction of the population should not be based on gender or ethnicity or other characteristics that exclude disadvantaged people.

Sometimes, the primary research studies may include participants who do not meet the inclusion criteria for the review. In Example 2, the population is *'people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma of the body or tail of the pancreas, or both'*. However, some research studies may include people undergoing distal pancreatectomy for malignant neuroendocrine tumours in addition to those undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma. When separate outcome data for the subset of people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma are available, either directly from the report or by contacting the primary research study authors or sponsors, it is reasonable to include the study and collect data relevant to the subset of the population of interest for the systematic review. However, when the separate data are not available, the two main options are excluding primary research studies where it is not possible to obtain separate data for the subset of the population of interest for the systematic review (you can follow this approach if the scoping search identifies many eligible studies) or including the primary research studies where a certain proportion of the study participants are eligible for inclusion, say 80% or 90% of the primary research study participants are population of interest for the systematic review.

Intervention

You need to specify the types of interventions that will be included in the review. The systematic reviewers should consider using the 'template for intervention description and replication' (TIDieR) checklist [19] to specify the eligibility criteria related to interventions. The TIDieR checklist describes an intervention using the following aspects. [19]

- Brief name.
- Why: this includes a description of any rationale, theory, or goal of the elements essential to the intervention. In writing the protocol and in reporting the findings of the systematic review, this aspect is reported in the background rather than the eligibility criteria, although the essential elements of the inclusion criteria are usually part of the eligibility criteria of the systematic review.
- What: this includes a description of any physical or informational materials used in the intervention, intervention delivery or in training of intervention providers, and any procedures, activities or processes used in the intervention.

- Who: this includes a description of the intervention provider (for example, general practitioner, medical specialist, nurse, physiotherapist), their expertise, background and any additional training given.
- How: this includes a description of the modes of delivery, for example, the route of administration in pharmacological interventions, face-to-face (individual versus group delivery and in-person versus virtual delivery), telephone, websites in non-pharmacological interventions.
- Where: this includes a description of the setting (for example, community-based, primary care, secondary care, specialist centres).
- When and how much: this includes a description of number of times the intervention was delivered, over what period (of time), the schedule, and the duration, intensity or dose.
- Tailoring and modifications: this includes a description of what, why, when and how any personalisation, titration, adaptation or modifications of the intervention, if any, are performed. It is unusual to exclude primary research studies on the basis that an unplanned modification was made in these studies. A more likely scenario is restricting the inclusion of a subset of study participants who received the intervention before (or after) the modification is made, depending upon how specific the research question is with regards to the intervention.
- How well: this includes a description of the extent to which the intervention was delivered as planned (intervention adherence or fidelity), including who assessed this and how, and steps to improve fidelity. Steps to improve fidelity are usually not considered as eligibility criteria for systematic reviews.

Although the TIDieR checklist was mainly targeted at improving the reporting of the intervention (and comparator) in primary research studies, [19] it appears to provide a good framework to develop the eligibility criteria related to interventions (and comparators) for systematic reviews. We recommend that systematic reviewers employ a widely used classification system such as SNOMED CT [16] or MeSH [15] terms to indicate the intervention name. The rationale for using standardised nomenclature is the same as that described in the section on population.

A concept related to eligibility criteria of studies based on intervention is whether to include studies in which a complex intervention is used. An intervention is considered complex when there is more than one component. [20] Other reasons why an intervention is considered a complex intervention can be due to the range of behaviours targeted, expertise and skills required by those delivering and receiving the intervention, when the intervention targets multiple groups or settings, or because of the permitted level of flexibility of the intervention or its components. [20] For example, a drug combination can be considered a complex

intervention because of a combination of drugs. Enhanced recovery protocols after a surgical procedure can be considered a complex intervention as it contains multiple components usually delivered by people with different expertise. [21]

In the context of eligibility criteria related to intervention, if the research question relates to a complex intervention with multiple components, then it is important to specify whether studies with one or more components missing from the intervention will be included. Usually, primary research studies that include fewer components are excluded from the analysis. For example, for the research objective *'Does "recombinant erythropoietin with iron" improve symptoms compared to "no recombinant erythropoietin or iron" in people with chronic renal failure but no anaemia?'*, the complex intervention is *'recombinant erythropoietin with iron'*. In such a systematic review, a primary research study which compares *'recombinant erythropoietin'* versus *'no recombinant erythropoietin or iron'* or *'iron'* versus *'no recombinant erythropoietin or iron'* will not be included. The reason is that we want to assess the effect of the combination rather than each component.

Comparator

You need to specify the types of comparators that will be included in the review. As for interventions, the systematic reviewers should consider using the TIDieR checklist [19] to specify the eligibility criteria related to comparator and use standardised nomenclature when possible. If the comparator was an active complex intervention, the same principles used when the intervention was a complex intervention apply.

Outcomes

You might use outcomes as eligibility criteria for systematic reviews. However, this could introduce bias. [22] Therefore, we recommend some mechanism of identifying the studies that meet all eligibility criteria other than 'outcomes' to allow the exploration of reporting biases. A more detailed description of reporting biases, their impact on the conclusions, how to assess them, and a practical way of identifying and managing the studies that meet all eligibility criteria other than 'outcomes' is available in later chapters of this book.

Other eligibility criteria

Study design

Study design is also commonly used as a criterion for systematic reviews. In Example 1, only RCTs were included. In Example 2, RCTs and NRSI were included (although this information is not available from the objectives). While well-designed and conducted RCTs provide the most reliable estimate of the effect of an intervention compared to a comparator, there may not be many RCTs

to answer the research question or the RCTs may not provide the outcome of interest for the systematic review. Therefore, you might need to include NRSI. The study designs to be included in the systematic review should be specified in advance. This can be decided by a scoping search.

Some related considerations with regards to eligibility criteria about study designs is whether you plan to include variations in RCTs, for example, cross-over RCTs or cluster RCTs. Many non-pharmacological interventions, such as surgery, are not suitable for cross-over RCT design because of their residual effect or because they might cure the disease.

Length of follow-up

It is appropriate to use the length of follow-up as an eligibility criterion depending upon the research question. If the objectives are to obtain the long-term outcomes, it is appropriate to limit the studies that use a long-term follow-up, in whichever way long-term follow-up may be defined. This might have an indirect effect on the study designs included. For example, if residual effect is expected and you are interested in short-term outcomes, you might consider including cross-over RCTs, but only outcomes measured prior to the cross-over; on the other hand, you might exclude such cross-over RCTs if you are interested in the long-term outcomes, as the period prior to cross-over is usually short.

Co-interventions

Co-interventions are interventions that study participants receive in addition to the intervention of interest for the systematic review, and which have the potential to affect the outcomes. Some learners confuse co-interventions with a situation in which the intervention is a complex intervention with multiple components. When the intervention is a complex intervention with multiple components, we are interested in the effect of the combination of the multiple components on the outcome. On the other hand, we are not interested in whether the co-intervention improves or worsens the outcome, although we might be interested in whether the addition of the co-intervention changes the effect of the intervention compared to the comparator on the outcome.

Eligibility criteria of studies in the presence of co-interventions is something you might want to consider while developing the eligibility criteria for systematic reviews. When the intervention and the comparator groups receive the co-intervention equally in primary research studies, such studies are usually included. An exception to this general guidance about co-interventions is when it has been established that the co-intervention causes harm. In such a situation, it is reasonable to exclude the studies that use the harmful co-intervention even if they are used equally. The rationale behind this is that co-intervention will not be used in clinical practice; therefore, it is not useful to find the effect of the

intervention on the outcome in the presence of the co-intervention. When the intervention and the comparator groups do not receive the co-intervention equally, for example, only the intervention group receives the co-intervention, such studies are excluded.

Language of publication

Eligibility criteria should specify whether reports from all languages will be included. In general, we recommend inclusion of reports in all languages. Restriction to English articles has the potential to cause bias (prejudice, [2] that is, preconceived judgement or opinion [24]). There is currently no strong evidence that restriction to English articles leads to highly biased results in systematic reviews of traditional medicine [25–29], although there is some evidence that this can lead to biased results in systematic reviews of complementary and alternate medicine. [27] However, it is generally advisable to include non-English articles in systematic reviews because of the uncertainty in the contribution of non-English articles to the results [27] and the difficulty in predicting whether exclusion of non-English articles can lead to bias and the direction of any bias, that is, overestimation or underestimation of the effect of the intervention on the outcome. [26]

While deciding on the restriction by language, you need to be aware that there is free software that performs optical character recognition of many non-English languages (provided that any scanned image is of high resolution), which can be corrected and then translated using online translators such as Google or Microsoft translator.

Year of publication

Eligibility criteria should specify whether there were any restrictions based on the years of publication. It is not efficient to search for trials before the first-in-human study was conducted.

You might also consider restricting the studies to more recent years if the general supportive care for a health condition or illness has improved over time and you are using supportive care as the comparator. Occasionally, the improvement in the co-interventions can also warrant a restriction based on the years of publication. For example, the survival benefit of a more radical potentially curative surgical procedure compared to a less invasive potentially curative surgical procedure may decrease with improvements in the chemotherapy regimen given as an adjunct to the radical and less invasive procedures.

Another reason that you might want to consider restricting the studies to more recent years is if there has been a considerable improvement in the technology used in a medical device used in the intervention or comparator in the systematic review.

You might also consider applying restrictions based on the years of publication because of the availability of resources to extract data. This is usually not relevant in intervention reviews as there is usually sparsity of studies that provide an unbiased estimate of the effect of the intervention versus comparator. Besides, restriction of studies to year of publication arbitrarily can introduce biased results [30], and therefore should be avoided.

Publication status

There is no consistent way of classifying studies by their publication status. The method proposed by Schmucker et al. to classify studies in this way [31] is shown in Table 3.2.

Systematic reviews investigating whether the publication of a study depends on the direction of results found evidence that studies with positive results are more likely to be published than those with negative results. [32,33] Therefore, exclusion of unpublished studies has the potential to influence the results of a systematic review and can lead to 'reporting bias' (bias resulting from publication of studies based on their results). However, another systematic review concluded that searching unpublished data and grey literature data in addition to published study data had unclear impact on the meta-analyses results as some studies showed considerable change in results while others did not demonstrate a considerable impact on the effect estimates. [31]

Table 3.2 Classification of study data by publication status

Type of study data	Definition	Examples
Unpublished data	Not published at all	1. Supplemental unpublished data related to published trials 2. Data obtained from the Food and Drug Administration (FDA) or other regulatory websites 3. Post-marketing analyses hidden from the public
Grey literature data	Print or electronic information not controlled by commercial or academic publishers	1. Data obtained from trial registers 2. Non-indexed conference abstracts (published in journal collections) 3. Dissertations 4. Press releases 5. Government reports 6. Policy documents 7. Book chapters
Published study data	Published as journal article, usually indexed in electronic databases	Published in journals such as *British Medical Journal*

Source: Based on method proposed by Schmucker et al. [31]

Because of the empirical evidence of reporting bias [32,33] and potential for change in results of systematic reviews, we strongly recommend that the systematic review authors do not place any restrictions based on publication status, that is, published as a full text article, conference abstracts or unpublished.

In a systematic review, Scherer et al. found that only about 45% of research studies initially presented as conference abstracts become available as 'full publications'. [34] In this same study, Scherer et al. found that 'full publication' of research studies initially presented as conference abstracts depended upon the results, with the research studies with positive studies being more likely to be published compared to those with negative studies. [34] Therefore, there is a compelling argument to include results from conference abstracts.

The major concerns about including conference abstracts include the resources required to search conference abstracts, inadequate information from abstracts, and whether the conference abstracts are reliable as they are not peer reviewed. [35] However, the arguments supporting the inclusion of conference abstracts outweigh the arguments against inclusion of conference abstracts. [35] Furthermore, the 'Conference Proceedings Citation Index' includes an electronically searchable index of conference abstracts of major biomedical conferences, making arbitrary decisions about which conference abstracts to search and the highly resource-intensive manual search of conference abstract books redundant. Therefore, we recommend searching conference abstracts routinely, preferably electronically through the 'Conference Proceedings Citation Index'.

FROM 'OBJECTIVES' TO 'OUTCOMES'

Terminology: outcome domains versus outcome measurement instruments

Because of the different ways in which the term 'outcome' is interpreted, it is useful to clarify how we have used the term in this book. We have used the term 'outcome' to represent the 'outcome domain' (a true state or endpoint of interest, irrespective of how it is measured) [36] which may be measured using different tools (outcome measurement instruments, OMI) [37] and may be reported in different ways.

For example, overall HRQoL is an outcome domain. There are many scales used for measuring HRQoL, for example, EQ-5D, [38] SF-36, [39] and FACT, [40] (FACT being specific to cancer). EQ-5D, SF-36 and FACT are outcome measurement instruments. Similarly, the Spencer Children's Anxiety Scale, Pediatric Anxiety Rating Scale, Multi-dimensional Anxiety Scale for Children, Child Anxiety Life Interference Scale and Child Anxiety Impact Scale are all outcome measurement instruments used for measuring anxiety in children. [41]

Even when you use the same scale or measure, it is possible to report the outcomes in different ways. For example, serious adverse events are an important outcome for stakeholders. The ICH-GCP defines serious adverse events as any untoward medical occurrence that results in death, is life-threatening, requires inpatient hospitalisation or prolongation of existing hospitalisation, or results in persistent or significant disability, incapacity, congenital anomaly or birth defect. [42] It is useful for stakeholders to know the average proportion of people who develop a serious adverse event and the average numbers of serious adverse events per participant. In this example, the outcome domain is serious adverse events, which is reported in different ways despite using the same definition.

Classification of outcomes

Before we go on to choose outcomes based on the objectives, it is useful to learn classification of outcomes. Outcomes in clinical research can be classified in many ways. One way of classifying the outcomes is the taxonomy proposed by Dodd et al. [43] In this classification system, clinical outcomes can be classified into five core areas.

1. Mortality or survival
2. Physiological or clinical
3. Life impact
4. Resource use
5. Adverse events or effects

The five core areas have 38 outcome domains. The full list of the 38 outcome domains included in these five core areas is available from Dodd et al. [43]

Choosing outcomes

The choice of outcomes follows the objectives of the research (which in turn depends upon the research question). However, research questions may not be specific in terms of outcomes. In Example 2, the research question did not specify the outcome. In this situation, to ensure that the research question is important to stakeholders, you should choose outcomes that are important to those stakeholders. In this context, we recommend that you start by searching the Core Outcome Measures in Effectiveness Trials (COMET) Initiative database.

Core outcome sets

The COMET Initiative brings together stakeholders (including patients and clinicians) who are interested in developing and applying an agreed standardised sets of outcomes ('core outcome sets' that should be measured and reported in clinical trials of a specific health condition or illness). [44] While the main target

is clinical trials, these core outcome sets should be considered for use in systematic reviews, as the outcomes are usually developed in conjunction with patients and clinicians. Therefore, we recommend using the outcomes listed in core outcome sets, even when the research question specifies an outcome. You can include the outcomes in the core outcome sets but not in the research question as secondary outcomes.

When core outcome sets are not available

Core outcome sets are not available for all illness or health conditions. The systematic reviewer still must choose outcomes that are important to stakeholders to perform a valuable systematic review, but may not have the time, expertise or financial resources to develop the core outcome sets formally. In this situation, it is important to consult with at least the patients and clinicians through discussions, surveys or both. A good starting point for the discussions or surveys is the taxonomy for outcome classification proposed by Dodd et al. [43] Of the 38 items in this classification, only four items were included in 20% or more of the 299 core outcome sets published until 2016. The items included in 20% or more of the 299 core outcome sets, the core area, and the number of core outcome sets in which this item was included are shown in Table 3.3.

It is clear from Table 3.3 that the stakeholders are mainly interested in how long patients lived, how well patients lived, and adverse events of the intervention. Therefore, these items should be considered in the discussion and surveys. It is also useful to note that some methodologists consider it mandatory to include adverse events as outcomes in systematic reviews related to health interventions. [9]

Choosing the appropriate outcome measurement instruments

After choosing the outcomes of interest for the stakeholders, you might be required to choose the appropriate outcome measurement instruments or

Table 3.3 **Outcome items which were included in 20% or more core outcome sets published up to 2016**

Outcome (item)	Core area	Number of core outcome sets which included the item (%)
Global quality of life	Life impact	121 (40%)
Physical functioning	Life impact	111 (37%)
Adverse events or effects	Adverse events or effects	105 (35%)
Mortality or survival	Mortality or survival	99 (33%)

Source: Based on information from Dodd et al. [43]

appropriate ways of measuring the outcomes. For example, global quality of life and physical functioning meet the criteria for patient-reported outcome measures, that is, they measure the status of a patient's health condition using standardised or structured questionnaire and come from direct questioning of the patient. [45] These are measured using many outcome measurement instruments as described in the section 'Terminology: outcome domains versus outcome measurement instruments' for global quality of life, and as summarised by Elsman et al. for physical functioning in type 2 diabetic patients. [46] While the systematic reviewer has no control over what outcome measurement instrument was used in the primary research study, it is reasonable to include only outcome measurement instruments with good measurement properties in the analysis. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) Initiative provides a database of systematic reviews of the measurement properties of different outcome measurement instruments [47] and can help with choosing the best way of measuring outcomes.

Patient-Reported Outcomes Measurement Information System (PROMIS®) Health Measures are validated outcome measurement instruments. [48] The suitable PROMIS Health Measure to measure an outcome can be found by searching online. [49]

Choosing the appropriate follow-up period

The period of follow-up period should be appropriate to answer the research question. For example, if you want to find whether a less invasive treatment provides equivalent survival to a more invasive treatment in a cancer with good prognosis, outcomes should be extracted at a reasonable follow-up time. You might be interested in outcomes at multiple time points, all of which may be clinically relevant. In this scenario, it is reasonable to include outcomes measured at multiple time points.

Misconceptions about choosing outcomes

A common misconception is that the review should include outcomes based on being reported in primary research studies. If the systematic review is performed to guide clinical practice, it is strongly recommended that only outcomes of interest to patients, healthcare professionals and policy makers are included in the systematic review. [9] Therefore, it is not appropriate to choose outcomes solely on the basis that they are reported in primary research studies.

Another misconception is that you should choose an objective measurement such as a blood test over subjective patient-reported outcome measures. As mentioned in the section 'When core outcome sets are not available', patient-reported outcome measures are very important to stakeholders.

Surrogate outcomes

Introduction

A major consideration in choosing an outcome for the systematic review is surrogate outcomes.

Surrogate means substitute. Surrogate outcomes are indirect outcomes used as substitutes to predict important clinical outcomes. Typically, the surrogate outcomes are lab-based outcomes or radiology-based outcomes. Sometimes, surrogate outcomes can be clinical outcomes that predict another important clinical outcome.

Surrogate outcomes are also called surrogate endpoints. Another related term is surrogate biomarker. Biomarkers are physical (external or internal) or genetic characteristics of an individual. Biomarkers are quite commonly used for diagnostic purposes. For example, to diagnose obesity, you use the height and weight of the individual and calculate the Body Mass Index. When such biomarkers are used as outcomes to assess the effect of an intervention, for example, the use of Body Mass Index in assessing the effectiveness of a lifestyle intervention, they are called surrogate biomarkers.

Advantages

Some of the advantages of surrogate outcomes are that they are quicker and cheaper to measure compared to the clinical outcomes. This means that the clinical trials used to assess the effect of an intervention can be completed more quickly. Some endpoints may be used because they can be measured non-invasively compared to the outcome of interest.

Criteria

If the biomarker is used as a substitute for a clinical outcome and is used as means or mechanism to influence the clinical outcome ('therapeutic target'), two conditions have to be met. [50]

- There should be good agreement between the surrogate outcome and the clinical outcome.
- Changing the surrogate outcome should result in an equivalent change in the clinical outcome.

Disadvantages

Despite the common use of surrogate outcomes in clinical and preclinical research studies to find out if an intervention works, surrogate outcomes are notorious for their misuse by researchers and industry, and can lead to wrong conclusions.

For example, Kim et al. found that only about 14% (5/36) of the drugs approved by US Food and Drug Administration (FDA) between 2008 and 2012 on the basis of progression-free survival and disease-free survival actually improved overall survival. [51] For 13/36 approvals, the evidence was not sufficient to conclude whether the drug improved survival, and in the remaining drugs (18/36), the drug did not improve survival. [51]

Rupp et al. found that, of the 18 drugs that were proven not to have survival benefit by Kim et al., HRQoL was available for seven drugs. [52] Of these, the HRQoL was worse than placebo or no treatment for two drugs; there was no difference in HRQoL between the drug and placebo or no treatment for four drugs; one drug had mixed results in terms of HRQoL. The average annual costs of the drugs with no survival or HRQoL advantage was US$88,000.

Ideal scenario for surrogate outcomes

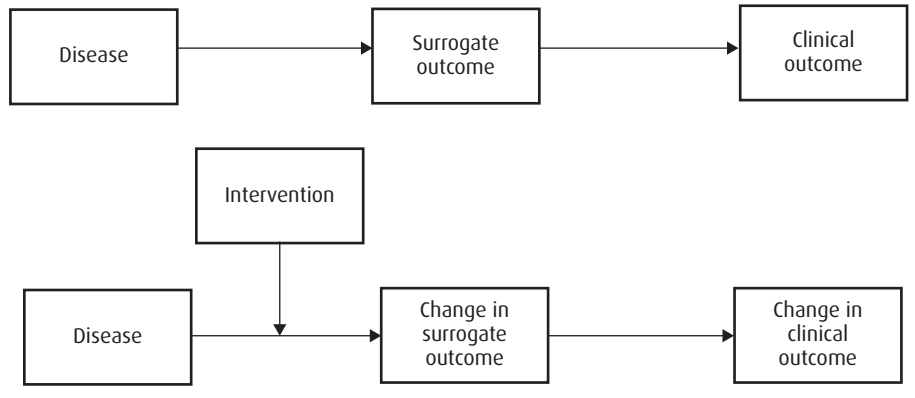
The ideal scenario for surrogate outcomes is when the surrogate outcome is the only pathway by which the disease can cause the clinical outcome, and the intervention acts in this pathway and causes a change in surrogate outcome, leading to a change in the clinical outcome. [53] This is shown in Figure 3.3.

Reasons for poor accuracy

The reasons for poor accuracy of surrogate outcomes are shown in Figure 3.4.

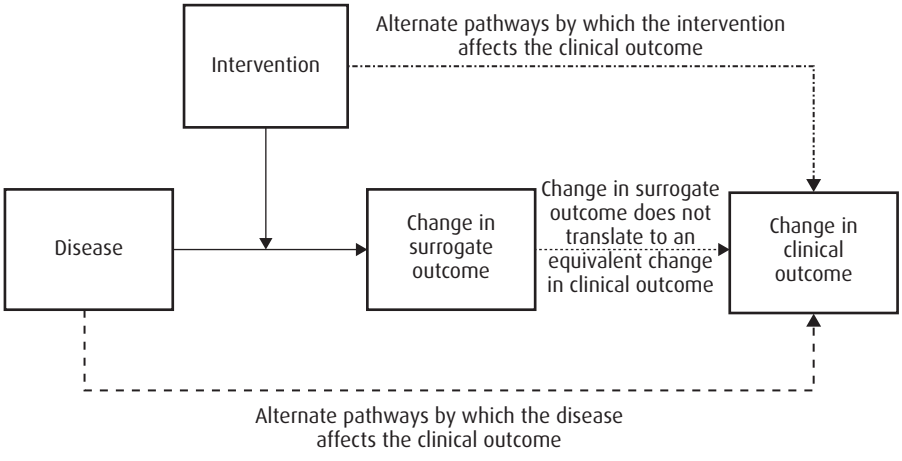
Essentially, if there are other pathways (which are not affected by the intervention) through which the disease can cause the clinical outcome, then the validity of the surrogate outcome will be decreased. If the intervention affects

Figure 3.3 Ideal scenario for surrogate outcomes



Source: Adapted from Gurusamy et al. [54] (contains public sector information licensed under the Open Government Licence v3.0).

Figure 3.4 Reasons for poor accuracy of surrogate outcomes



Source: Adapted from Gurusamy et al. [54] (contains public sector information licensed under the Open Government Licence v3.0).

Table 3.4 Non-healthcare example of surrogate outcome

Aspect	Valencia	Real Madrid
Possession	35%	65%
Shots on target	4	6
Corners	1	8

Source: Data from: www.bbc.com/sport/football/54859896

the clinical outcome through pathways unrelated to the surrogate outcome, again the validity of the surrogate outcome will be decreased.

How can the intervention improve a surrogate outcome but not result in an equivalent change in the clinical outcome? The reasons for this were mentioned in the previous paragraph. A non-healthcare example is also provided to explain this concept further.

Let us say you are asked to guess which team won in football based on ball possession, shots on target and corners, as shown in Table 3.4.

As a background for people who do not follow football, ball possession, shots on target and corners can be considered indirect measures of the domination of one team over another. Of course, simply having high ball possession, more shots on target or more corners do not mean that the team wins. The opposite team's goalkeeper may have defended the goal successfully on all the attempts. The team that has less possession and fewer shots on target or corners may have more goals as they shoot more accurately.

Can you guess the winner from the information shown in Table 3.4? Everything points to Real Madrid being the dominant team. So, you can be forgiven for thinking that Real Madrid must have won this game or at least, this was a draw. However, Valencia thrashed Real Madrid 4–1.

Similarly, an intervention may improve the surrogate outcome but not the clinical outcome, leading to measurement error.

Inference

So, what inference can we make from the information provided? Considerable measurement error is possible when surrogate outcomes are used to assess interventions, leading to wrong conclusions. This can result in interventions causing harm to people despite showing improvement in surrogate outcomes.

Composite outcomes

Introduction

Composite outcomes consist of two or more component outcomes. When a patient has at least one of the component outcomes, they are considered to have the composite outcome.

Advantages

The main reason for the use of a composite outcome is that it increases statistical efficiency, that is, because of higher event rates, the sample size required to demonstrate differences between two treatments is decreased. [55] This decreases the resources and the time required to conduct the trial. Other reasons for the use of the composite outcome is that it includes more than one aspect of health status and avoids having to choose arbitrarily between outcomes of equal importance. [55]

Disadvantages

The major problem with composite outcomes is that the components that make up the composite outcome should be of similar importance. [55] If this condition is not met, this can lead to significant problems, particularly if the intervention does not result in similar reductions in the components. For example, an intervention may decrease the proportion of patients who developed the complication component of less clinical importance (say, mild stroke) but increase the proportion of patients who developed the complication component of greater clinical importance (say, severe stroke). Even if there is an overall decrease in the composite outcome, stroke in this example, can the intervention be recommended?

If the different components are not of similar importance to the patient and reductions in the different components of the primary outcome are not equal, it can lead to confusion in interpretation. [56,57]

SUMMARY

In this chapter, we have provided some guidance on determining the research question for a systematic review.

PRACTICE QUESTIONS

1. From the following research objective 'To determine the effect of exercise in addition to dietary advice compared with dietary advice on prevention of stroke in people with type II diabetes mellitus', identify the following.
   a. Population
   b. Intervention
   c. Comparator
   d. Outcome(s)
2. From the following research objective 'To determine the effect of heart bypass surgery versus stent in improving the longevity of life of people with heart attack', identify the following.
   a. Population
   b. Intervention
   c. Comparator
   d. Outcome(s)

Answers to the practice questions can be found in the Appendix.

REFERENCES

1. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009; 374(9683):86–9.
2. James Lind Alliance. Top 10s of Priorities for Research. www.jla.nihr.ac.uk/top-10-priorities/ (accessed on 4 September 2023). 2023.
3. Smith DJ, Gold JAW, Benedict K, Wu K, Lyman M, Jordan A, et al. Public health research priorities for fungal diseases: a multidisciplinary approach to save lives. Fungi (Basel). 2023; 9(8).
4. Raichurkar P, Denehy L, Solomon M, Koh C, Pillinger N, Hogan S, et al. Research priorities in prehabilitation for patients undergoing cancer

surgery: an international Delphi study. Ann Surg Oncol. 2023; 30(12):7226–35.
5. National Institute for Health and Care Excellence. Research Recommendations. www.nice.org.uk/about/what-we-do/science-policy-research/research-recommendations (accessed on 5 September 2023). 2023.
6. Garner P, Hopewell S, Chandler J, MacLehose H, Schünemann HJ, Akl EA, et al. When and how to update systematic reviews: consensus and checklist. BMJ. 2016; 354: i3507.
7. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016; 69:225–34.
8. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Method. 2018; 18(1):5.
9. Higgins J, Lasserson T, Chandler J, Tovey D, Churchill R. Methodological Expectations of Cochrane Intervention Reviews. Cochrane: London. https://community.cochrane.org/sites/default/files/uploads/MECIR%20PRINTED%20BOOKLET%20FINAL%20v1.01.pdf (accessed on 7 September 2023). 2016.
10. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021; 372: n160.
11. Joury E, Bernabe E, Sabbah W, Nakhleh K, Gurusamy K. Systematic review and meta-analysis of randomised controlled trials on the effectiveness of school-based dental screening versus no screening on improving oral health in children. Dent. 2017; 58:1–10.
12. Riviere D, Gurusamy KS, Kooby DA, Vollmer CM, Besselink MG, Davidson BR, et al. Laparoscopic versus open distal pancreatectomy for pancreatic cancer. Cochrane Database Syst Rev. 2016; 4(4): Cd011391.
13. Komolafe O, Buzzetti E, Linden A, Best LM, Madden AM, Roberts D, et al. Nutritional supplementation for nonalcohol-related fatty liver disease: a network meta-analysis. Cochrane Database Syst Rev. 2021; 7(7): Cd013157.
14. World Health Organization. International Statistical Classification of Diseases and Related Health Problems 11th Revision. www.who.int/standards/classifications/classification-of-diseases (accessed on 31 March 2024). 2022.
15. National Library of Medicine. Welcome to Medical Subject Headings. www.nlm.nih.gov/mesh/meshhome.html (accessed on 10 September 2023). 2023.

16. SNOWMED International. SNOMED International SNOMED CT Browser. https://browser.ihtsdotools.org/? (accessed on 10 September 2023). 2023.
17. National Library of Medicine. Unified Medical Language System (UMLS). www.nlm.nih.gov/research/umls/index.html (accessed on 31 March 2024). 2023.
18. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol. 2014; 67(1):56–64.
19. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014; 348: g1687.
20. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021; 374: n2061.
21. Bond-Smith G, Belgaumkar AP, Davidson BR, Gurusamy KS. Enhanced recovery protocols for major upper gastrointestinal, liver and pancreatic surgery. Cochrane Database Syst Rev. 2016; 2(2): Cd011382.
22. 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. 2010; 340: c365.
23. Merriam-Webster Dictionary. Bias. www.merriam-webster.com/dictionary/bias (accessed on 14 March 2024). 2023.
24. Merriam-Webster Dictionary. Prejudice. www.merriam-webster.com/dictionary/prejudice (accessed on 9 September 2023). 2023.
25. Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR, et al. What contributions do languages other than English make on the results of meta-analyses? J Clin Epidemiol. 2000; 53(9):964–72.
26. Jüni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002; 31(1):115–23.
27. Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. Health Technol Assess. 2003; 7(41):1–90.
28. Morrison A, Polisena J, Husereau D, Moulton K, Clark M, Fiander M, et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. Int J Technol Assess Health Care. 2012; 28(2):138–44.
29. Dobrescu AI, Nussbaumer-Streit B, Klerings I, Wagner G, Persad E, Sommer I, et al. Restricting evidence syntheses of interventions to

English-language publications is a viable methodological shortcut for most medical topics: a systematic review. J Clin Epidemiol. 2021; 137:209–17.

30. Marshall IJ, Marshall R, Wallace BC, Brassey J, Thomas J. Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. J Clin Epidemiol. 2019; 109:30–41.

31. Schmucker CM, Blümle A, Schell LK, Schwarzer G, Oeller P, Cabrera L, et al. Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. PLoS One. 2017; 12(4): e0176210.

32. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One. 2008; 3(8): e3081.

33. 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. 2010; 14(8):iii,ix–xi,1–193.

34. Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database Syst Rev. 2007(2): Mr000005.

35. Scherer RW, Saldanha IJ. How should systematic reviewers handle conference abstracts? A view from the trenches. Systematic Reviews. 2019; 8(1):264.

36. Sterne J, Higgins J, Elbers R, Reeves B. Development group for ROBINS-I. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance. www.riskofbias.info (accessed on 10 September 2023). 2016.

37. Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. Guideline for selecting outcome measurement instruments for outcomes included in a core outcome set. https://cosmin.nl/wp-content/uploads/COSMIN-guideline-selecting-outcome-measurement-COS.pdf (accessed on 10 September 2023). 2016.

38. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013; 22(7):1717–27.

39. McHorney CA, Ware JE, Jr., Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care. 1993; 31(3):247–63.

40. Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. Clin Oncol. 1993; 11(3):570–9.

41. Linden A, Best L, Elise F, Roberts D, Branagan A, Tay YBE, et al. Benefits and harms of interventions to improve anxiety, depression, and other mental health outcomes for autistic people: a systematic review and network meta-analysis of randomised controlled trials. Autism. 2023; 27(1):7–30.
42. ICH. Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice. E6(R2). https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf (accessed on 9 September 2023). 2016.
43. Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. J Clin Epidemiol. 2018; 96:84–92.
44. COMET Initiative. Core Outcome Measures in Effectiveness Trials. www.comet-initiative.org/ (accessed on 10 September 2023). 2023.
45. Cella D, Hahn EA, Jensen SE, Butt Z, Nowinski CJ, Rothrock N, et al. Patient-Reported Outcomes in Performance Measurement. www.ncbi.nlm.nih.gov/books/NBK424378/pdf/Bookshelf_NBK424378.pdf (accessed on 10 September 2023). 2015.
46. Elsman EBM, Mokkink LB, Langendoen-Gort M, Rutters F, Beulens J, Elders PJM, et al. Systematic review on the measurement properties of diabetes-specific patient-reported outcome measures (PROMs) for measuring physical functioning in people with type 2 diabetes. BMJ Open Diabetes Res Care. 2022; 10(3).
47. COSMIN Initiative. Find the Right Tool. www.cosmin.nl/finding-right-tool/select-best-measurement-instrument/ (accessed on 10 September 2023). 2023.
48. PROMIS Health Organization. What Is PROMIS? www.promishealth.org/57461-2/ (accessed on 10 September 2023). 2023.
49. Health Measures. Search & View Measures. www.healthmeasures.net/search-view-measures (accessed on 10 September 2023). 2023.
50. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. 1989; 8(4):431–40.
51. Kim C, Prasad V. Cancer drugs approved on the basis of a surrogate end point and subsequent overall survival: an analysis of 5 years of US Food and Drug Administration approvals. JAMA Intern Med. 2015; 175(12):1992–4.
52. Rupp T, Zuckerman D. Quality of life, overall survival, and costs of cancer drugs approved based on surrogate endpoints. JAMA Intern Med. 2017; 177(2):276–7.
53. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996; 125(7):605–13.
54. Gurusamy KS, Moher D, Loizidou M, Ahmed I, Avey MT, Barron CC, et al. Clinical relevance assessment of animal preclinical research (RAA) tool: development and explanation. PeerJ. 2021; 9:e10673.

55. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ. 2010; 341: c3920.

56. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003; 289(19):2554–9.

57. Montori VM, Permanyer-Miralda G, Ferreira-González I, Busse JW, Pacheco-Huergo V, Bryant D, et al. Validity of composite end points in clinical trials. BMJ. 2005; 330(7491):594–6.

4 Select the studies

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- *Choose the databases that you need to search*
- *Understand the principles of designing formal search strategies*

OVERVIEW

In many types of systematic reviews, the aim is to identify all the studies related to a topic. Multiple sources must be searched to achieve this aim. The minimum databases that must be searched for systematic reviews of intervention should be MEDLINE, EMBASE, Cochrane Library and trial registers. If the appropriate sources are not searched, there is a possibility of getting a wrong answer for the research question. To avoid reporting bias, there should be no restrictions based on language or publication status while designing the search strategy.

The first step in designing a formal search strategy is to split the research question into different parts. These different parts are called 'domains' or 'key concepts' or simply 'concepts'. In an intervention review, the domains are based around the population, intervention, comparator and outcomes (PICO) format used for determining the eligibility criteria. Except under exceptional circumstances, we recommend that the search strategy does not include 'outcomes' as one of the domains, so that exploration of reporting biases is possible.

Each domain or key concept can be further subdivided into sub-domains and sub-concepts. Each concept or sub-concept should include a combination of controlled vocabulary terms and free text terms. The concepts and sub-concepts are combined with each other using Boolean terms (such as 'OR', 'AND') depending upon whether you want to identify articles that contain at least one of the concepts (or sub-concepts) or only those articles that contain all the concepts (or sub-concepts).

Search filters are validated combinations of search terms related to a concept and can be used when appropriate. Most search filters available relate to the study design.

Once a search strategy is designed, at least two independent researchers should make the study selection. All reports of a study should be selected and collated. There is uncertainty about the role of automated searching and screening tools.

WHAT IS STUDY SELECTION?

You may recollect from Chapter 2 that once the research question is determined the next steps are to identify suitable methods for study selection, data extraction, analysis and reporting, and to complete protocol registration.

As the name indicates, study selection involves selecting the studies to be included in the systematic review.

WHY IS STUDY SELECTION IMPORTANT?

Since data are extracted from the included studies (and subsequently analysed to obtain the results and arrive at conclusions in the systematic review), study selection is a key step in a systematic review and can affect the results of the review. To illustrate this, we provide an example. In 2012, GlaxoSmithKline LLC (GSK) agreed to pay US$3 billion to the US Government to resolve its criminal and civil liability arising from the unlawful actions of the company. [1] Among other things, the company was alleged to have unlawfully promoted an antidepressant (paroxetine) in children and adolescents. [1] While the UK Government did not decide to prosecute GSK with regards to paroxetine, the report of the investigation by the Medicines and Healthcare Products Regulatory Agency (MHRA) is available from the National Archives of the UK Government. [2] As per this report, GSK provided analyses that did not reveal any evidence of increased suicidal risk because of the use of paroxetine. However, this analysis included adult and paediatric trials: when a meta-analysis of clinical trial data on children and adolescents was performed, it revealed an increased risk of suicide related to paroxetine in children and adolescents. [2]

Furthermore, by definition, a systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria to answer a specific research question. [3] This inevitably means that attempts should be made to identify all the studies that provide the answer to the specific research question. In addition,

the empirical evidence that we have provided in Chapter 3 highlights the importance of not restricting studies based on the publication status or language.

STUDY SELECTION PROCESS

Overview of study selection

An overview of the process of study selection is shown in Figure 4.1. Further details about each aspect of this process are explained in the remainder of this chapter.

Step 1 Choose the criteria for study selection

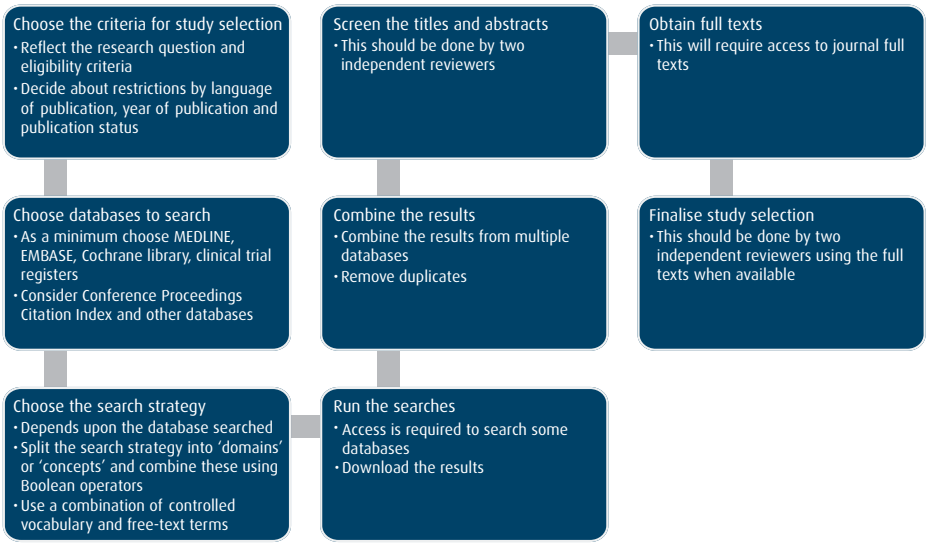
There are certain general considerations and research question-specific considerations in choosing the criteria for study selection. These relate to the eligibility criteria based on PICO, and other eligibility criteria such as restrictions by study design, length of follow-up, co-interventions, language of publication, year of publication and publication status as described in Chapter 3. To avoid reporting bias, there should be no restrictions based on language or publication status while designing the search strategy.

Step 2 Choose databases to search

MEDLINE, EMBASE, CINAHL, PsycINFO and LILACS are examples of biomedical literature databases that index biomedical literature. Each article published

Figure 4.1 Overview of the process of study selection

This figure shows an overview of the study selection process.



in an indexed journal is added to these searchable literature databases: some journals are included in multiple biomedical literature databases, while others are included in a limited number of databases. Therefore, the articles included in these literature databases are different and more than one database should be searched in a systematic review. These databases can be searched through various platforms. For example, MEDLINE, EMBASE and PsycINFO can be searched through the Ovid® platform, a subscription-based service. CINAHL can be searched through the EBSCOhost platform.

MEDLINE can also be searched through PubMed. In fact, MEDLINE is the largest component of PubMed, a free resource for searching biomedical literature. [4] However, restriction of articles to MEDLINE alone can result in relevant studies being missed, resulting in incorrect results in the systematic review. [5,6] Therefore, as a minimum, for systematic reviews in health and disease, you should search MEDLINE and EMBASE.

The Cochrane Central Register of Controlled Trials (CENTRAL) is a register provided by Cochrane Library and contains RCTs and quasi-randomised studies identified by searching PubMed, EMBASE and CINAHL. [7] Searching the Cochrane Library is free, although downloading the results is a subscription-based service. There are no studies that investigate whether searching Cochrane CENTRAL can mitigate the risk of missing studies resulting from a failure to search EMBASE.

The Conference Proceedings Citation Index contains abstracts of many biomedical conferences and appears to be a viable alternative to the resource-intensive manual searching of conference abstract publications. This can be searched through the Web of Science™ platform and is a subscription-based service. We recommend searching the Conference Proceedings Citation Index when possible, to identify conference abstracts.

You must consider searching other databases depending upon the type of population, intervention, or comparator. For example, if you perform a systematic review of nursing or allied health intervention, you should consider searching CINAHL; if the population is people with psychiatric illnesses, you should consider searching PsycINFO; if the intervention involves Chinese medicines, you should consider searching one of the Chinese biomedical literature databases; [8] and if the intervention involves treatment of illnesses mainly prevalent in Latin America and Caribbean countries, you should consider searching the LILACS database. [9]

You must search trial registers routinely to decrease the risk of reporting bias. The International Committee of Medical Journal Editors recommends that the journals should consider accepting only clinical trials with publicly accessible registration in any registry that is a primary register of the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP) or

in ClinicalTrials.gov. [10] Therefore, we recommend that these clinical trial registries are searched as a minimum to identify unpublished trials.

Step 3 Choose the search strategy

Search strategy or 'formal search strategy' refers to a set of terms used to search the different biomedical databases.

Why should you use a 'formal search strategy'?

One of the key characteristics of a systematic review is the use of reproducible methodology. [3] Therefore, it is necessary that a 'formal search strategy' is used (to allow others to replicate the searches) rather than finding the studies using ad hoc search terms.

How do you design a formal search strategy?

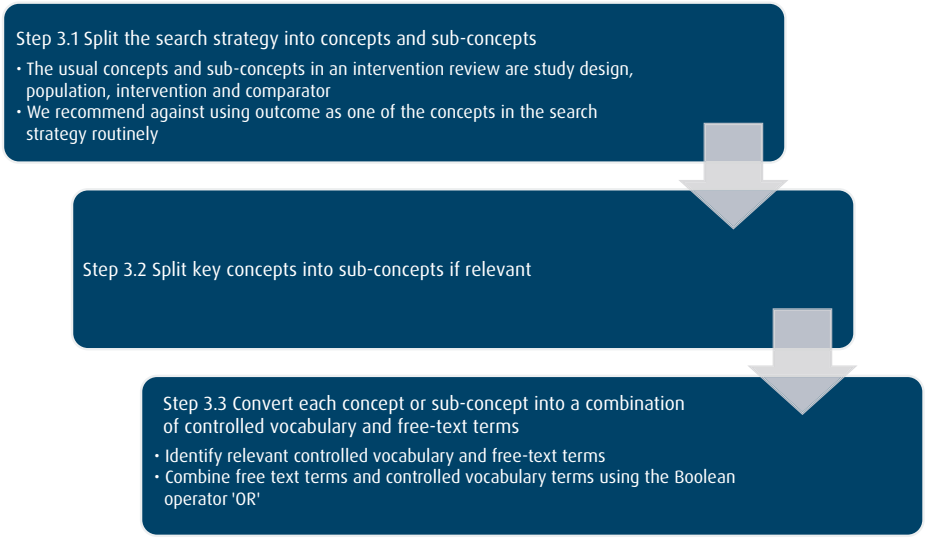
Overview of designing a formal search strategy

An overview of designing a formal search strategy is provided in Figure 4.2.

Step 3.1 Split the search strategy into 'domains' or 'key concepts'

The first step in designing a formal search strategy is to split this into different parts. These different parts are called 'domains' or 'key concepts' or simply 'concepts'. In an intervention review, the domains are based around the PICO format used for determining the eligibility criteria. However, except under exceptional

Figure 4.2 Overview of designing a formal search strategy



circumstances, we recommend that the search strategy does not include 'outcomes' as one of the domains, so that exploration of reporting biases is possible.

Example 1

The search strategy 'domains' or 'concepts' in the systematic review whose primary objective is *'Evaluate the effectiveness of school-based dental screening versus no screening on improving oral health in children aged 3–18 years by a systematic review and meta-analysis of randomised controlled trials'* [11] are (a) population: *'children'*, (b) intervention: *'school-based dental screening'*, and (c) comparator: *'no screening'*.

Example 2

The search strategy 'domains' or 'concepts' in the systematic review whose primary objective is *'To assess the benefits and harms of laparoscopic distal pancreatectomy versus open distal pancreatectomy for people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma of the body or tail of the pancreas, or both'* [12] are (a) population: *'people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma'*, (b) intervention: *'laparoscopic distal pancreatectomy'*, and (c) comparator: *'open distal pancreatectomy'*.

If you compare the PICO that we stated in Chapter 3, the search concepts are much less specific. This is because we do not want to be too restrictive in the search concepts. Being too restrictive at the search concept level can result in missing studies which include some concepts only in the full text level: full text is not searched when you run a search in the common databases mentioned above, although the subject headings and keywords are based on full text.

Step 3.2 Split key concepts into sub-concepts if relevant

Once you have identified key concepts, you may split them into sub-concepts. The intervention *'school-based dental screening'* in Example 1 can be split further into *'school'* and *'dental screening'*. The population *'people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma'* in Example 2 can be further split into *'distal pancreatectomy'* and *'pancreatic ductal adenocarcinoma'*; the intervention in the same example could be split into *'laparoscopic'* and *'distal pancreatectomy'*.

Step 3.3 Convert each concept or sub-concept into a combination of controlled vocabulary and free text terms

'Controlled vocabulary terms' are a set of standardised terms or 'codes' to describe various medical concepts. The underlying reason for using such 'standardised terms' is that it is possible that the same medical concept is described in different ways in different publications.

Example 3

'Chronic kidney failure' and 'chronic renal failure' refer to the same medical concept. Some journal publications may use the term 'chronic kidney failure' while others may use the term 'chronic renal failure'. In the absence of the concept of 'controlled vocabulary', a search for 'chronic kidney failure' will not retrieve articles which used the term 'chronic renal failure' to describe the same medical illness; similarly, a search for 'chronic renal failure' will not retrieve articles which used the term 'chronic kidney failure' to describe the same medical illness (in the absence of the concept of 'controlled vocabulary').

Example 4

'Liver transplantation' and 'liver grafting' refer to the same medical concept. Some journal publications may use the term 'liver transplantation' while others may use the term 'liver grafting'. In the absence of the concept of 'controlled vocabulary', a search for 'liver transplantation' will not retrieve articles which used the term 'liver grafting' to describe the same medical procedure; similarly, a search for 'liver grafting' will not retrieve articles which used the term 'liver transplantation' to describe the same procedure (in the absence of the concept of 'controlled vocabulary').

The use of 'controlled vocabulary' helps in retrieving articles which use different words for the same concept. In Example 3, using the controlled vocabulary term 'Kidney Failure, Chronic' will allow retrieval of articles which use either 'chronic kidney failure' or 'chronic renal failure'. In Example 4, using the controlled vocabulary term 'Liver Transplantation' allows articles which used the term 'liver grafting' to be retrieved. Therefore, the use of controlled vocabulary terms is particularly useful when more than one term exists for a disease or condition. It is also useful for identifying foreign articles.

How do you identify the controlled vocabulary terms for a concept or sub-concept?

In Examples 3 and 4, we mentioned that the controlled vocabulary terms are 'Kidney Failure, Chronic' and 'Liver Transplantation'. These terms are specific to MEDLINE, although other databases may use the same controlled vocabulary terms for this concept. It is important that you verify the controlled vocabulary terms for each database before you use it. But how do we identify these controlled vocabulary terms? This is answered step by step using screenshots.

Step 3.3.1 Identifying controlled vocabulary terms for MEDLINE (through PubMed)

The controlled vocabulary terms are called Medical Subject Headings (MeSH) in MEDLINE. The steps below show how to obtain MeSH terms through the PubMed platform. (Note: The screenshots are intended to provide instructive and

practical guidance on processes as they currently stand, however I acknowledge that some of these pages may change over time.)

Step 3.3.1.1

Go to PubMed initial screen at https://pubmed.ncbi.nlm.nih.gov/ (see Figure 4.3). Click 'MeSH database' (red ellipse). This takes you to the screen shown in Figure 4.4.

Figure 4.3 Identifying controlled vocabulary terms for MEDLINE (through PubMed): landing page

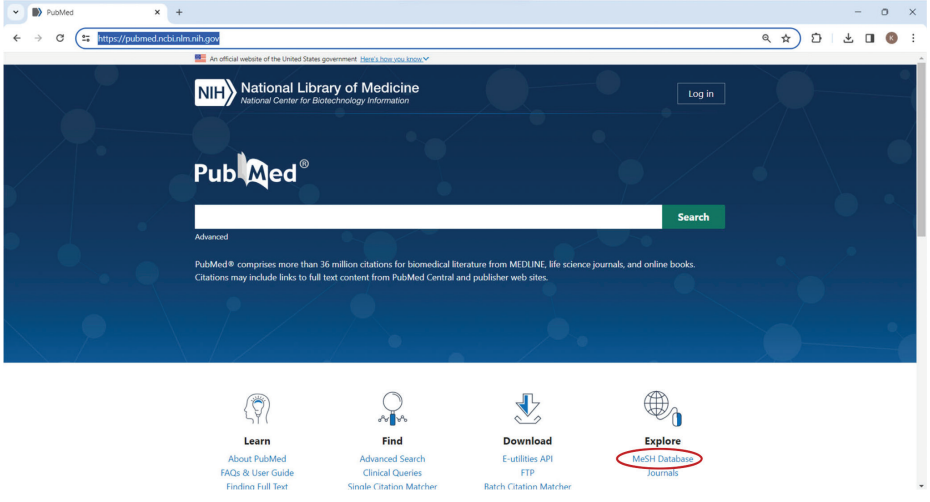
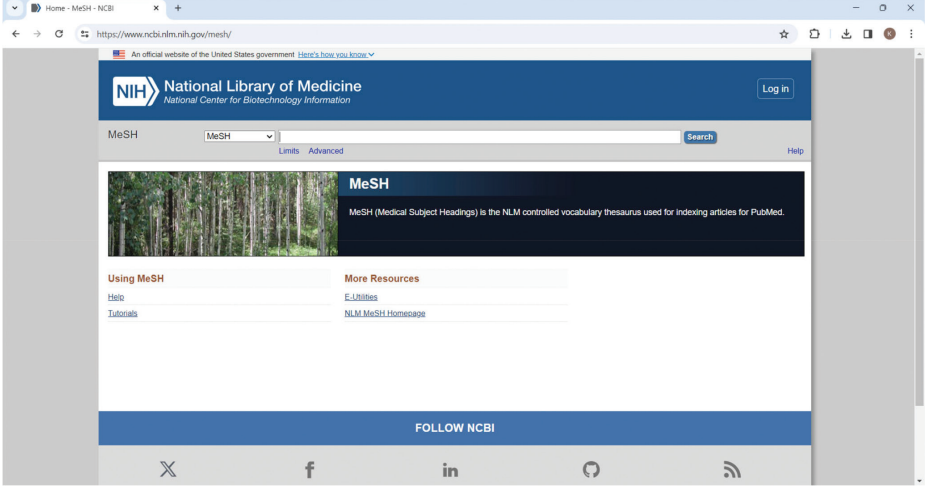


Figure 4.4 Identifying controlled vocabulary terms for MEDLINE (through PubMed): MeSH search interface



Step 3.3.1.2

Enter the medical concept or sub-concept. You might get text predictions. Click 'Search' (red ellipse), shown in Figure 4.5.

Step 3.3.1.3

This gives you the MeSH term (red ellipse shown in Figure 4.6).

Figure 4.5 Identifying controlled vocabulary terms for MEDLINE (through PubMed): search terms entered

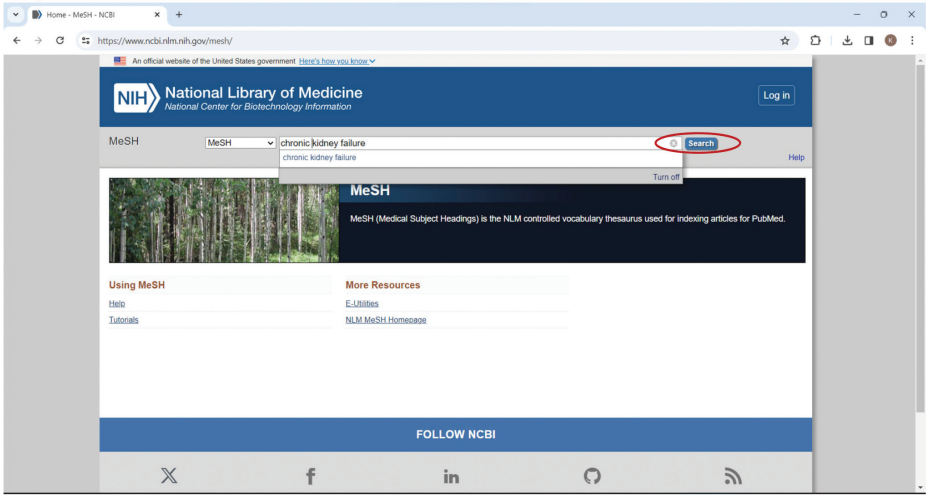


Figure 4.6 Identifying controlled vocabulary terms for MEDLINE (through PubMed): search results (singleMeSH term)

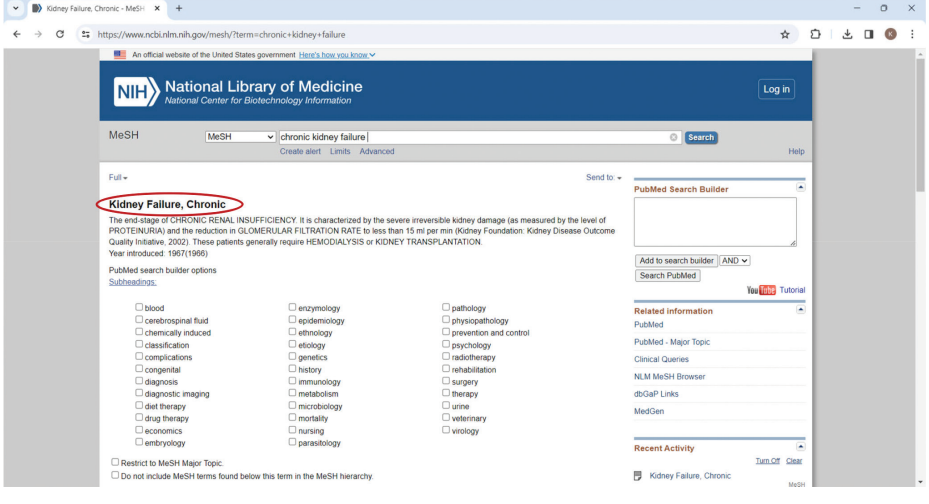
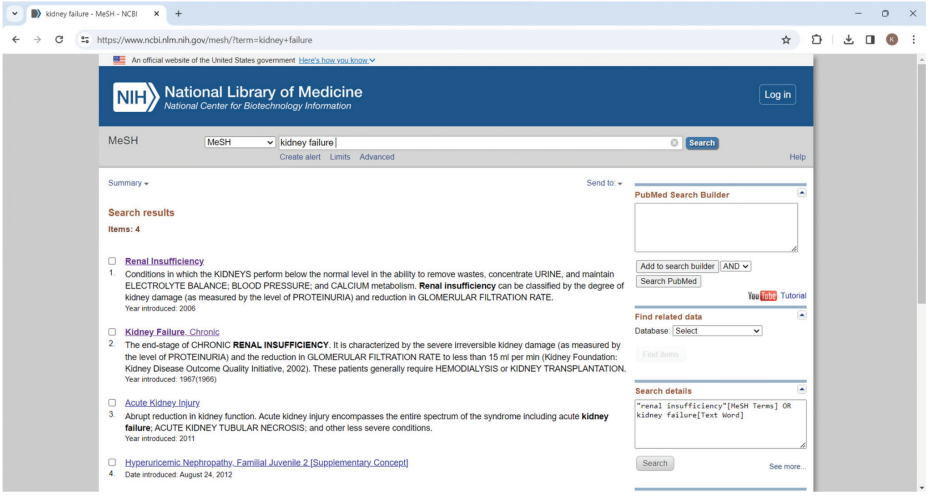


Figure 4.7 Identifying controlled vocabulary terms for MEDLINE (through PubMed): search results (multiple MeSH terms)



Sometimes, you might get many MeSH terms for a single concept or sub-concept. For example, if you used the term 'kidney failure' rather than 'chronic kidney failure', you would get multiple MeSH terms as shown in the screenshot in Figure 4.7. In such a situation, you must choose the appropriate term. You may sometimes have to click on each term to find the closest match.

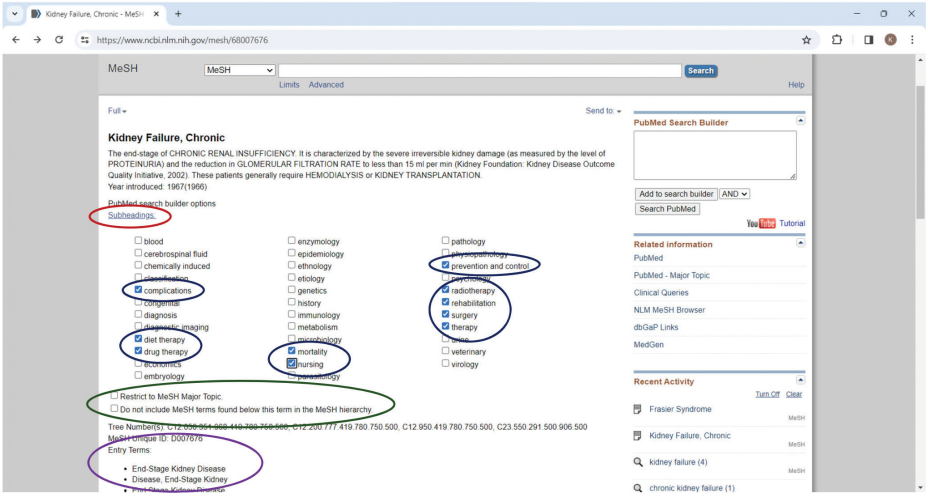
Step 3.3.1.4

You might have noticed that when you obtain the MeSH term, there are subheadings (red ellipse in Figure 4.8). You can select the appropriate subheadings depending on the focus of your systematic review. In the screenshot, I have selected all the treatment and clinical outcome-related subheadings (blue ellipses).

You might also have noticed that I did not select the boxes next to 'Restrict to MeSH Major Topic' or 'Do not include MeSH terms below the term in MeSH hierarchy' (green ellipse). Restricting the search to MeSH Major Topic refers to the topic being identified as a major topic in the article and should be considered only if there is an unmanageable number of references that make it impossible to perform a systematic review. We strongly recommend against choosing the option 'Do not include MeSH terms below the term in MeSH hierarchy' as this can lead to relevant articles included under narrower MeSH terms being missed. Please discuss the implications of selecting this option with an information specialist or librarian with expertise in designing search strategies.

You might have also noticed that there are 'Entry terms' (purple ellipse). These are the free text terms for the concept or sub-concept. The importance of

Figure 4.8 Identifying controlled vocabulary terms for MEDLINE (through PubMed): subheadings



searching databases using free text terms is explained in the section 'Step 3.3.8 Identifying free text terms'.

Step 3.3.1.5

Sometimes, you might want to use broader or narrower terms (depending upon whether you want to use a broader or narrower definition of the medical concept compared to the term that you identified). For example, when you scroll down the page that you obtained by clicking 'Kidney Failure, Chronic', you can find broader and narrower terms (see Figure 4.9).

You might want to consider using 'Renal Insufficiency, Chronic' (red ellipse), if your systematic review had a broader focus than 'Renal Failure, Chronic'. On the other hand, you may only be interested in Frasier syndrome (blue ellipse), a medical condition that has chronic renal failure as one feature in addition to other features. In such a case, you can consider including only Frasier syndrome as the MeSH term. Please note that there are multiple such instances of these terms. Clicking any of these takes you to the term in which you are interested.

Step 3.3.1.6

After making all the appropriate selections, click 'Add to search builder' (red ellipse in Figure 4.10). This results in some terms in the PubMed Search Builder (blue ellipse). These are the controlled vocabulary search terms for the concept or sub-concept. You can click 'Search PubMed' (green ellipse) to search PubMed using these terms.

Figure 4.9 Identifying controlled vocabulary terms for MEDLINE (through PubMed): broader and narrower terms

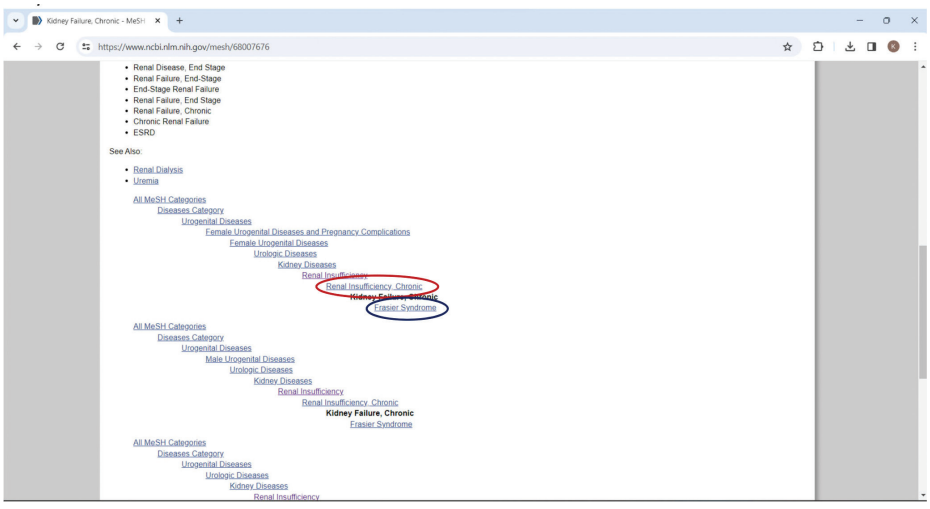
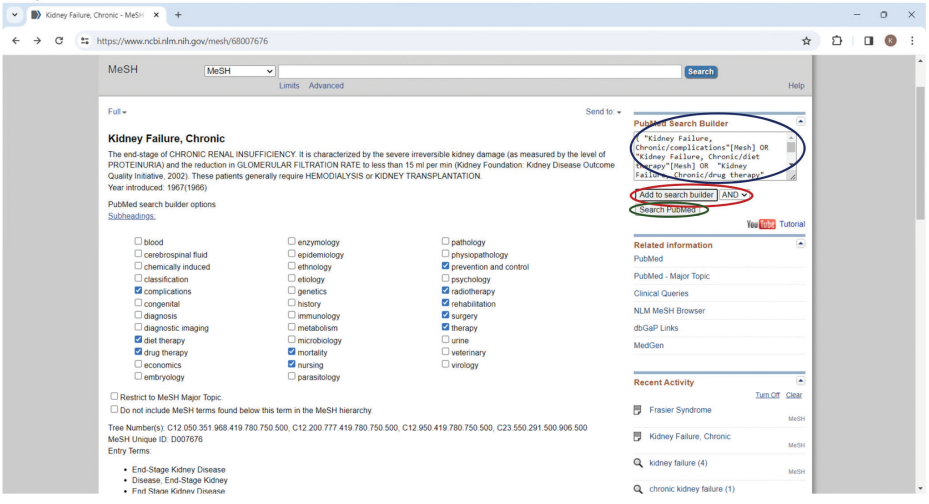


Figure 4.10 Identifying controlled vocabulary terms for MEDLINE (through PubMed): search builder



Step 3.3.2 Identifying controlled vocabulary terms for MEDLINE (through Ovid)

Step 3.3.2.1

Log in to Ovid and select Ovid MEDLINE (red ellipse on Figure 4.11). Click 'Continue' (blue ellipse).

Figure 4.11 Identifying controlled vocabulary terms for MEDLINE (through Ovid): landing page (for Ovid)

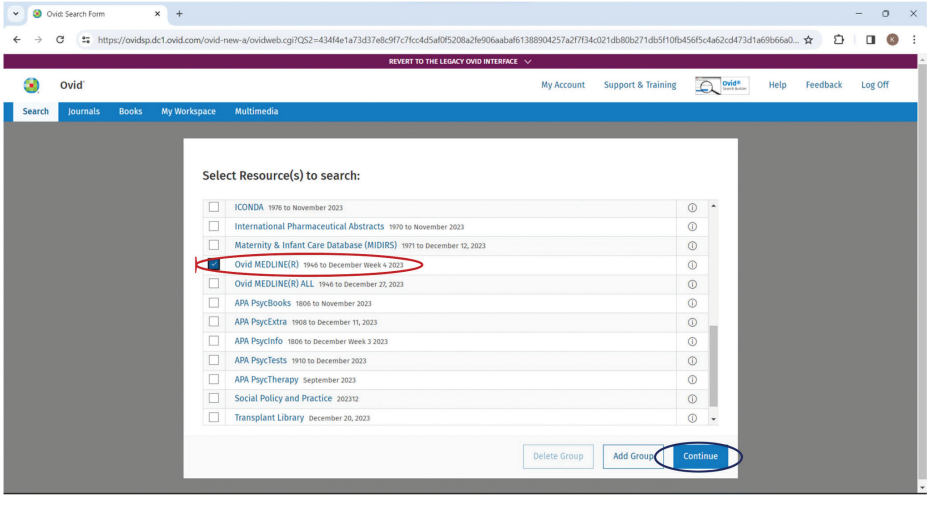
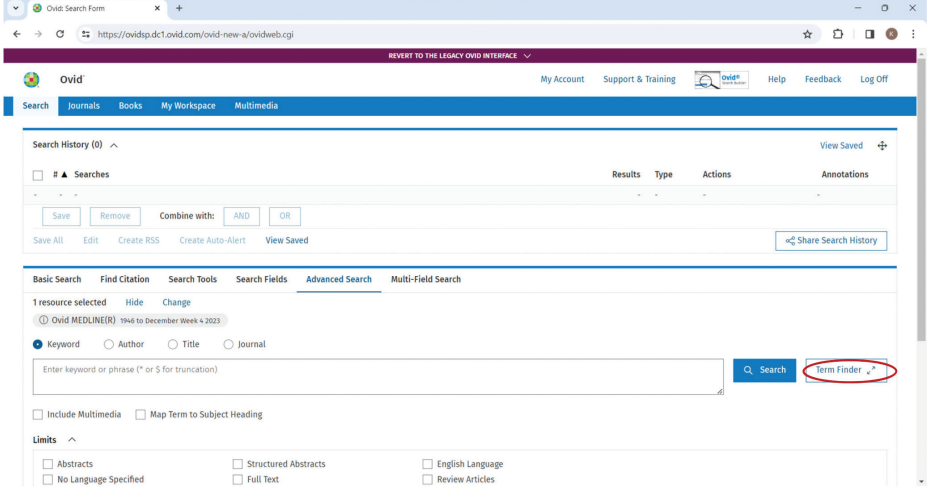


Figure 4.12 Identifying controlled vocabulary terms for MEDLINE (through Ovid): search interface for Ovid MEDLINE



Although the 'Term Finder' (red ellipse in Figure 4.12) can be useful for finding controlled vocabulary terms in MEDLINE (MeSH terms), it does not allow the same level of control as that offered by 'Map Term to Subject Heading'. Therefore, we recommend using the same steps as those for EMBASE which involves using 'Map Term to Subject Heading'. Please see Step 3.3.3 for further information on using 'Map Term to Subject Heading' for identifying controlled

vocabulary terms in Ovid. It should also be noted that the MeSH terms identified through Ovid are the same as those identified through PubMed.

Step 3.3.3 Identifying controlled vocabulary terms for EMBASE (through Ovid)

Step 3.3.3.1

The initial screen when you log into EMBASE is identical to that of MEDLINE, but there is no 'Term Finder'. Enter the search term into the search box (red ellipse in Figure 4.13). Select 'Map Term to Subject Heading' (blue ellipse) and click 'Search' (green ellipse).

Step 3.3.3.2

In the resulting screen, any subject headings (called Emtree terms and equivalent to MeSH terms in MEDLINE), are shown as clickable links (red ellipse in Figure 4.14). Please note there is also a search term with '.mp.' added to its end (blue ellipse). At this stage, we are not interested in this term.

Step 3.3.3.3

Clicking the relevant Emtree link results in the screen shown in Figure 4.15. There are broader terms and narrower terms (red ellipses) for you to consider depending upon the research question. There are also some terms under '[Used For]' (blue ellipse). These are additional free text terms. Click 'Explode' (green ellipse) and then 'Continue' (purple ellipse).

Figure 4.13 Identifying controlled vocabulary terms for EMBASE (through Ovid): search interface for Ovid EMBASE

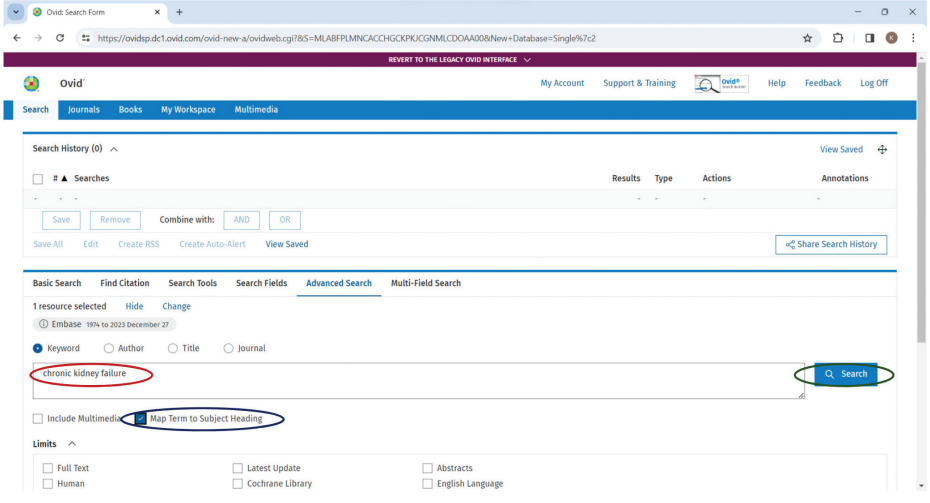


Figure 4.14 Identifying controlled vocabulary terms for EMBASE (through Ovid): search results

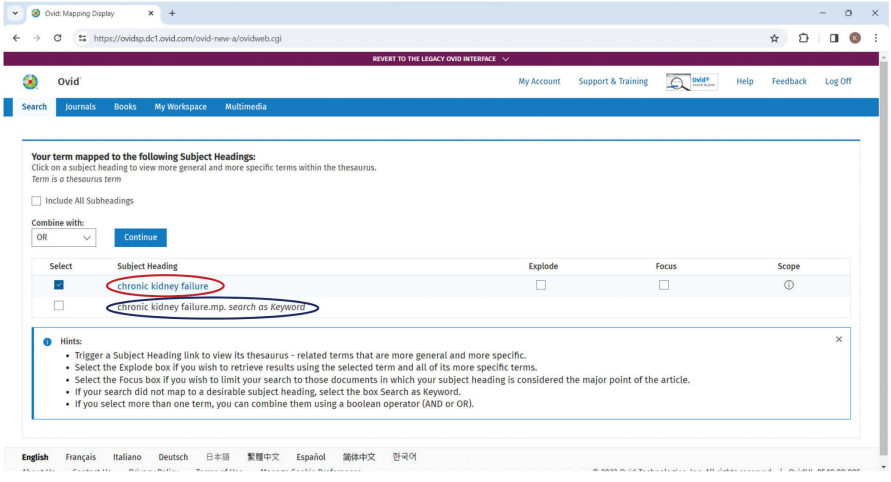
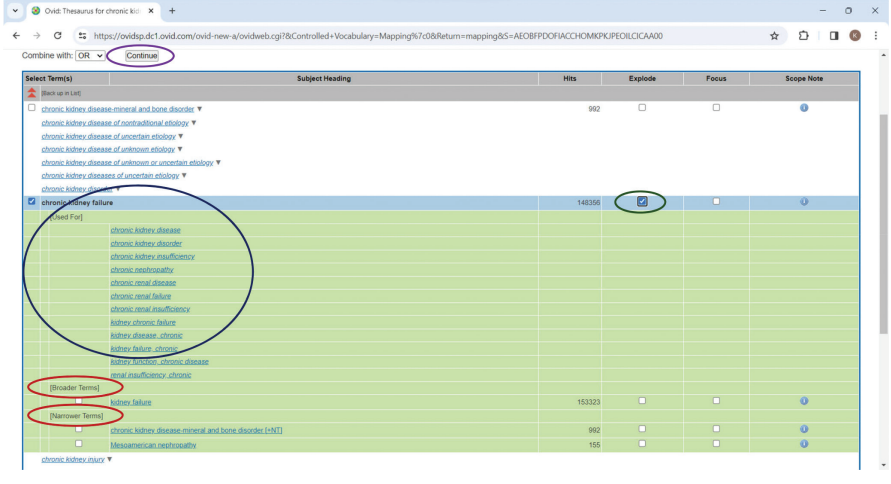


Figure 4.15 Identifying controlled vocabulary terms for EMBASE (through Ovid): Emtree terms



Step 3.3.3.4

In the resulting screen (Figure 4.16), choose the relevant subheadings (red ellipses) and click 'Continue' without altering the 'OR' in the dropdown menu (blue ellipse).

Step 3.3.3.5

This results in the search terms directly included in the search (red ellipse in Figure 4.17).

Figure 4.16 Identifying controlled vocabulary terms for EMBASE (through Ovid): subheadings

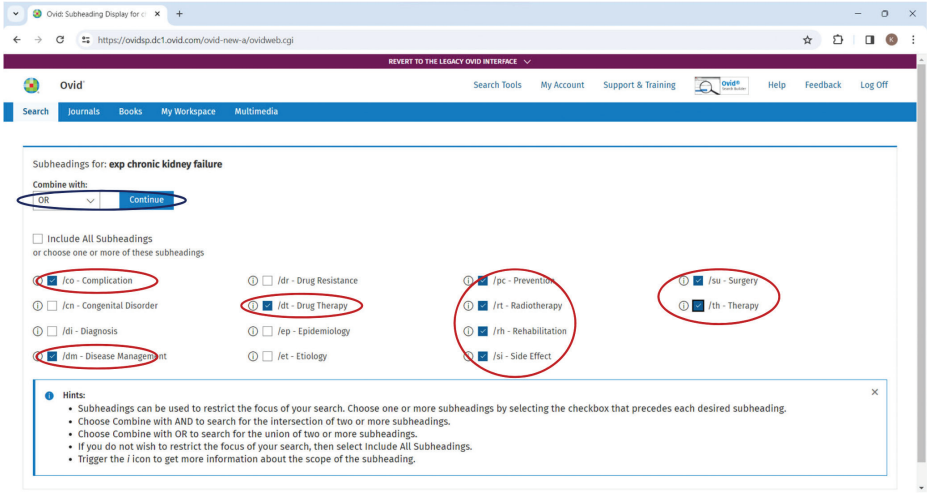
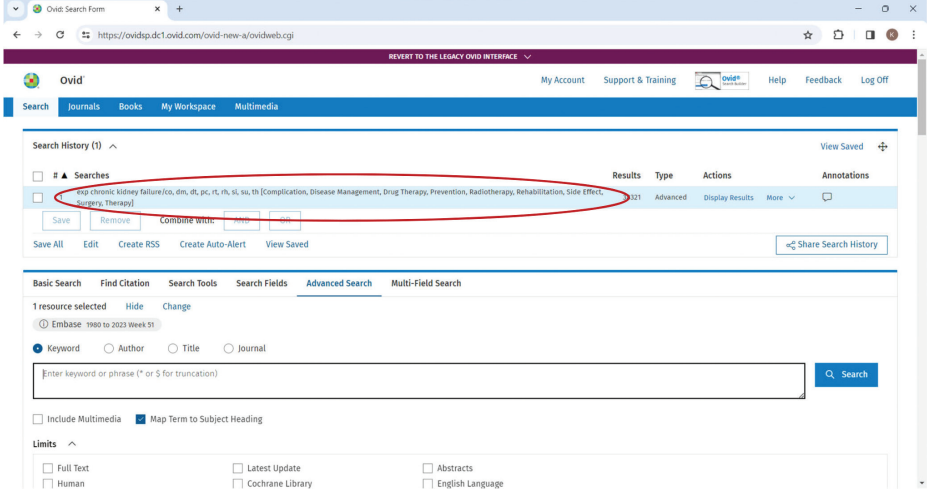


Figure 4.17 Identifying controlled vocabulary terms for EMBASE (through Ovid): final search results



Step 3.3.4 Identifying controlled vocabulary terms for Cochrane Library (through Wiley)

Step 3.3.4.1

The initial screen when you visit the Cochrane Library is shown in the screenshot (Figure 4.18). Click 'Advanced Search' (red ellipse).

Figure 4.18 Identifying controlled vocabulary terms for Cochrane Library (through Wiley): landing page

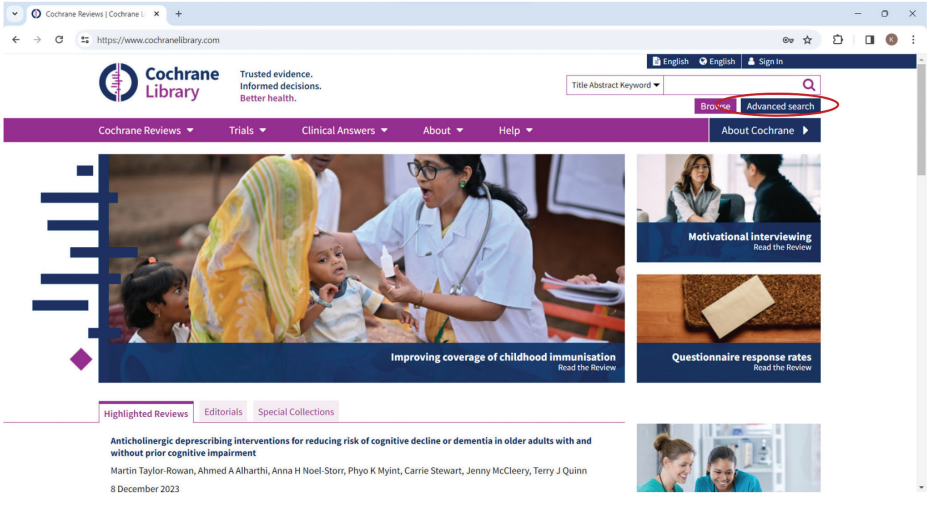
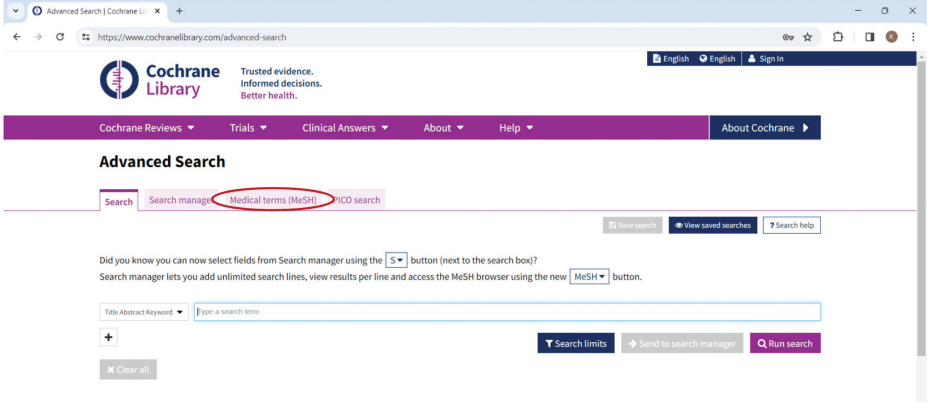


Figure 4.19 Identifying controlled vocabulary terms for Cochrane Library (through Wiley): advanced search interface



Step 3.3.4.2

In the resulting screen (Figure 4.19), choose the 'Medical terms (MeSH)' tab (red ellipse).

Step 3.3.4.3

Enter the search term (red ellipse in Figure 4.20) and click 'Look up' (blue ellipse).

Figure 4.20 Identifying controlled vocabulary terms for Cochrane Library (through Wiley): Advanced Search interface with search terms entered

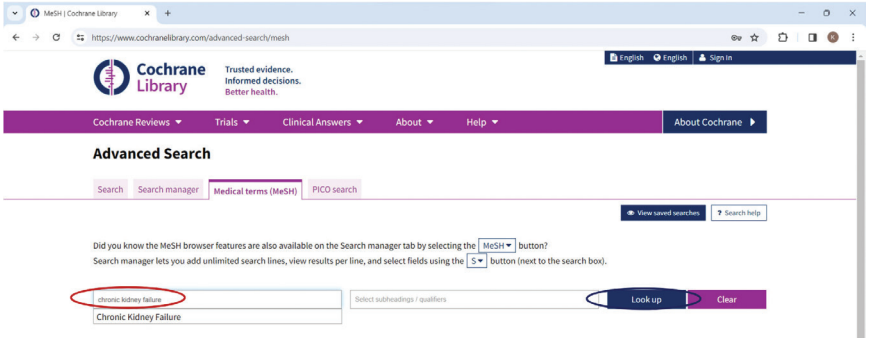
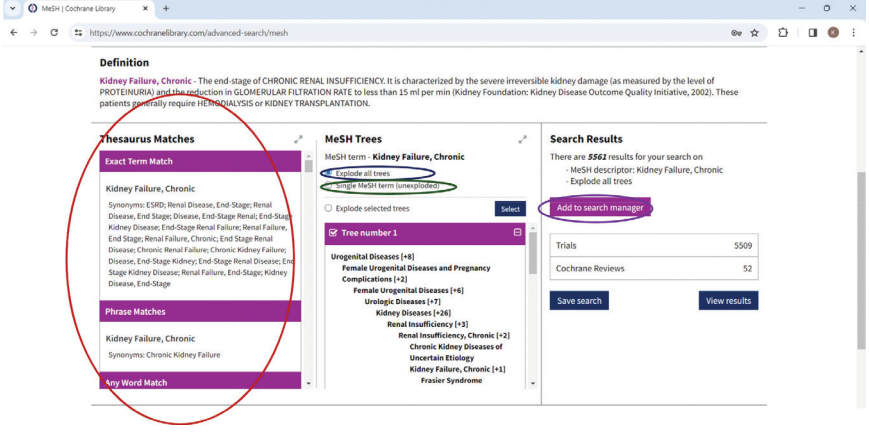


Figure 4.21 Identifying controlled vocabulary terms for Cochrane Library (through Wiley): thesaurus matches



Step 3.3.4.4

In the resulting screen (Figure 4.21), some 'Thesaurus Matches' are available (red ellipse). Retain the default option of 'Explode all trees' (blue ellipse). Selecting 'Single MeSH term (unexploded)' (green ellipse) can result in fewer results, but this may exclude some relevant studies. Therefore, consult with an information specialist or librarian with expertise in designing search strategies before selecting the option. Click 'Add to search manager' (purple ellipse).

Step 3.3.4.5

This results in the MeSH terms being added to the search manager (red ellipse in Figure 4.22). In the screenshot, you might notice that the MeSH terms in the Cochrane Library are identical to those in MEDLINE.

Figure 4.22 Identifying controlled vocabulary terms for Cochrane Library (through Wiley): search results

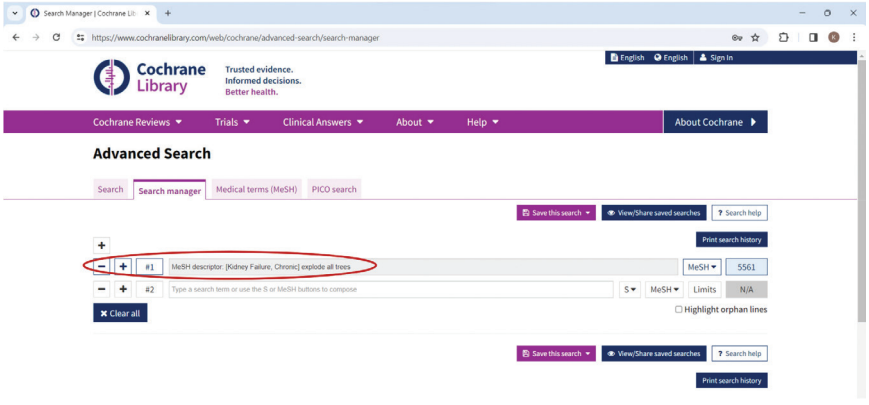
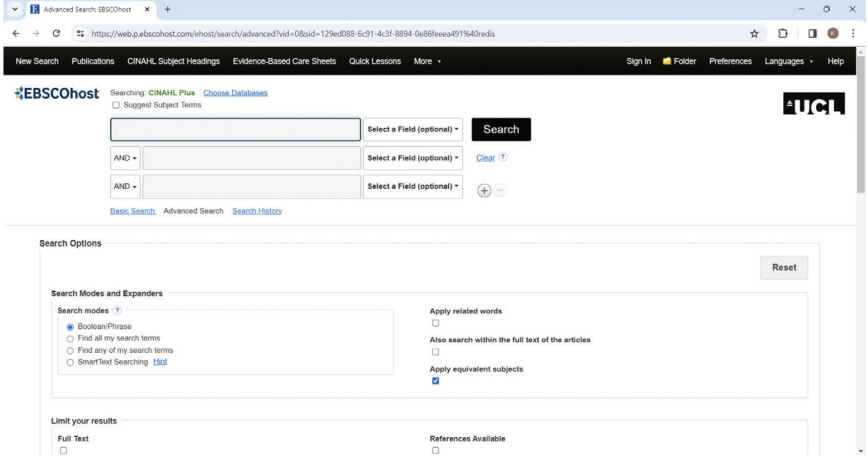


Figure 4.23 Identifying controlled vocabulary terms for CINAHL (through EBSCOhost): landing page



Step 3.3.5 Identifying controlled vocabulary terms for CINAHL (through EBSCOhost)

Step 3.3.5.1

The initial screen on logging into CINAHL database through EBSCO is shown in the screenshot (Figure 4.23).

Step 3.3.5.2

In the resulting screen (Figure 4.24), enter the search term in the search box (red ellipse), check the 'Suggest Subject Terms' box (blue ellipse), and click 'Search' (green ellipse).

Figure 4.24 Identifying controlled vocabulary terms for CINAHL (through EBSCOhost): search terms entered

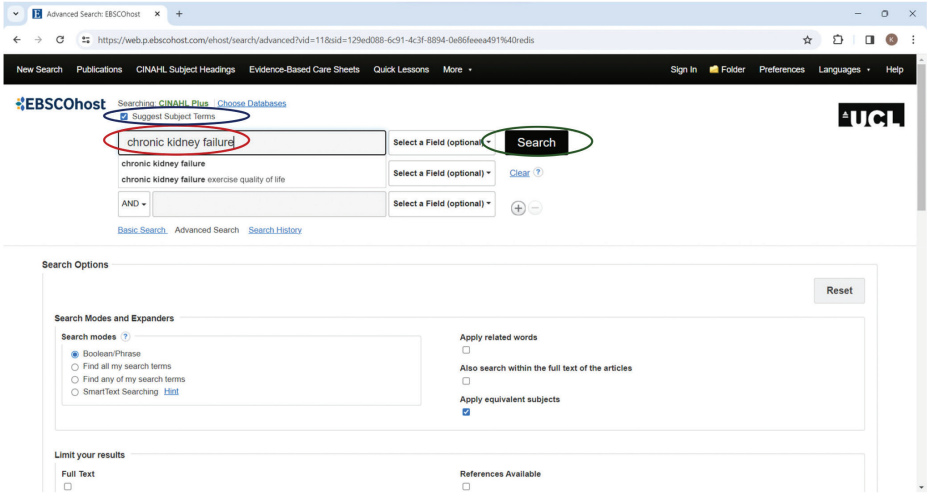
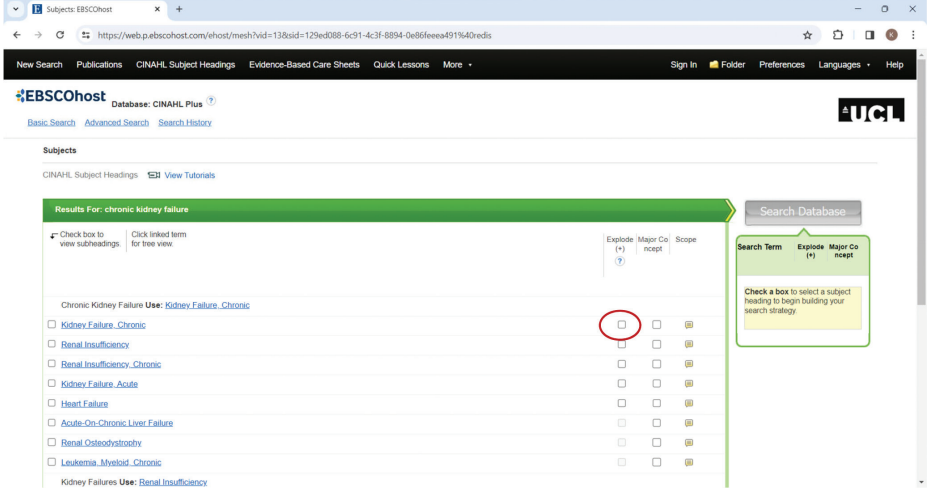


Figure 4.25 Identifying controlled vocabulary terms for CINAHL (through EBSCOhost): search results



Step 3.3.5.3

This results in the screen shown in Figure 4.25. Click 'Explode' for the appropriate term (red ellipse).

Step 3.3.5.4

In the resulting screen (Figure 4.26), choose the relevant 'Subheadings' (red ellipse) and click 'Search Database' (blue ellipse). You can restrict this to 'Major

Figure 4.26 Identifying controlled vocabulary terms for CINAHL (through EBSCOhost): subheadings

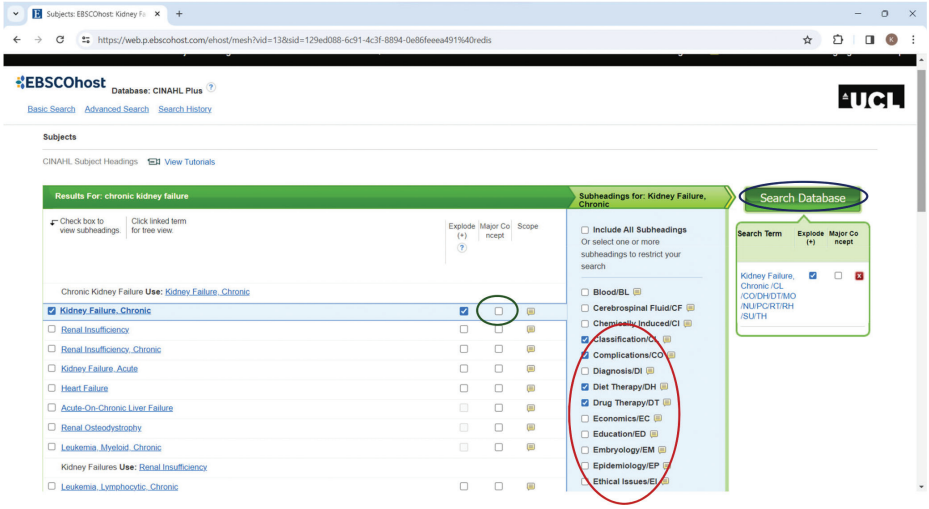
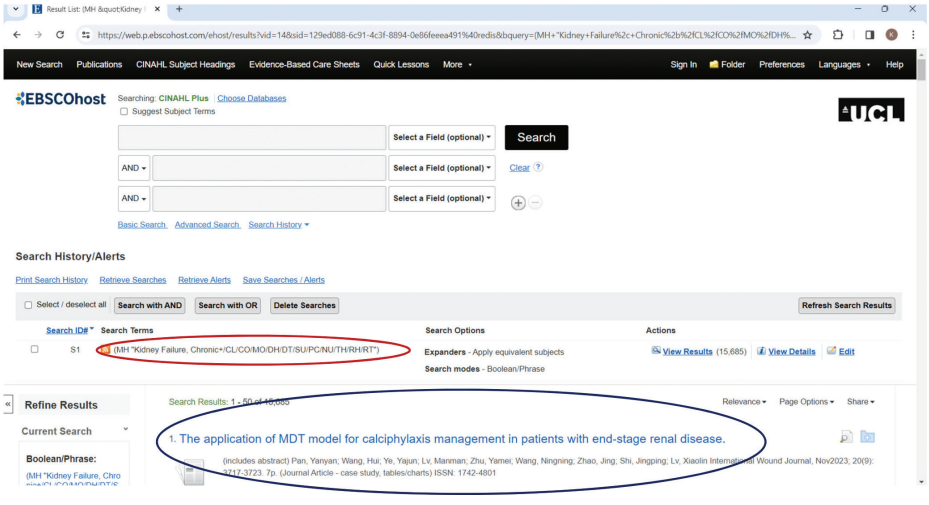


Figure 4.27 Identifying controlled vocabulary terms for CINAHL (through EBSCOhost): final search results

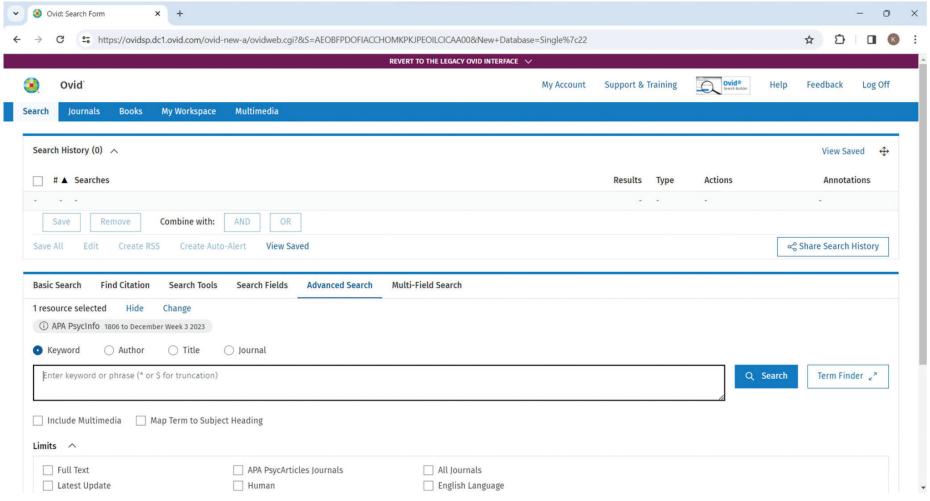


Concept' to obtain manageable results (green ellipse). In the demonstration, this has not been selected.

Step 3.3.5.5

This results in the screen shown in Figure 4.27 with the search terms (red ellipse) and the results (blue ellipse).

Figure 4.28 Identifying controlled vocabulary terms for PsycInfo (through Ovid): landing page



Step 3.3.6 Identifying controlled vocabulary terms for PsycInfo (through Ovid)

The initial screen when you log into PsycInfo through Ovid is shown in the screenshot (Figure 4.28). The subsequent steps in identifying the controlled vocabulary terms are like those in EMBASE.

Step 3.3.7 Identifying controlled vocabulary terms for the Conference Proceedings Citation Index (through Web of Science) and trial registers

There are no controlled vocabulary terms for the Conference Proceedings Citation Index, ClinicalTrials.gov and WHO ICTRP. Only free text terms are used.

Step 3.3.8 Identifying free text terms

Why should you search using free text terms?

Free text terms (those terms that are not part of the controlled vocabulary terms of the database) should also be used to search the databases. Studies have shown that free text terms and controlled vocabulary terms complement each other and that fewer studies are missed when these are used in combination. [13,14] One reason could be that it may take time for a journal article to be fully indexed. Therefore, using only controlled vocabulary terms alone may miss the newer publications. [15]

How can free text terms be identified?

Free text terms can be identified while communicating with the stakeholders. Further free text terms can be identified by reviewing the terms used for a

concept (or sub-concept) in some of the articles already identified without a formal search strategy. You can identify further free text terms at the same time as searching for controlled vocabulary terms. The screenshots under the sections 'Step 3.3.1.4', 'Step 3.3.3.3', and 'Step 3.3.4.4' show some terms. These are the free text terms that were mapped on to the controlled vocabulary terms. Therefore, you can consider using these terms as free text terms. You can also consider using the search terms in SNOMED CT, [16] ICD-11 [17] and Unified Medical Language System® (UMLS®) terms if you have used these terms to define the population, intervention and comparator. [18]

How do you enter the free text terms?

The free text terms can be entered as identified, but any variations in spelling should also be included in the searches. The use of variations in spellings is best illustrated using an example.

Example 5

If you are performing a systematic review of the role of haematopoietic stem cell transplantation in children with lymphoma, you should search for 'haematopoietic stem cell transplantation' in addition to 'hematopoietic stem cell transplantation' to account for the variations in the spellings used in different regions of the world.

Wildcard search

Wildcard search refers to use of '?' in the search string. The databases process the character '?' as indicating any character (including alphabets, numbers, hyphens, space and no character). For example, rather than using the two spellings as described above, you can use 'h?ematopoietic stem cell transplantation' which will retrieve both 'haematopoietic stem cell transplantation' and 'hematopoietic stem cell transplantation'. While not relevant (as no such words exist), the terms 'hbematopoietic stem cell transplantation', 'hcematopoietic stem cell transplantation', and so on, will also be searched using the wildcard as the second character.

Truncation

The free text terms can also be truncated (shortened) using the character '*' in most databases. Ovid also allows the use of '$' to truncate search terms. The advantage of using such truncation is that any words starting with the letters before '*' or '$' (called 'word stem') are retrieved by the 'truncated' search term. However, the disadvantage is that if there are lots of variations of the word stem this can result in retrieval of unnecessary articles. Furthermore, in some databases, only a limited number of variations of the word stem are searched; this may result in relevant articles being missed.

Because of the potential influence of using wildcard and truncated searches, if you have used terms that contain wildcard or truncated terms, we recommend that you consult with an information specialist or a librarian with expertise in designing search strategies to understand the impact of using such terms.

Step 3.3.9 Combining search terms through Boolean operators

The next step is to combine the search terms through Boolean operators. The common Boolean operators available in the databases are 'OR' and 'AND'. Some systematic reviewers get confused as to which of these Boolean operators should be used to combine search terms. The confusion arises from the way these terms are used in spoken language. A simple way to remember which of these Boolean operators should be used is to translate 'OR' to 'any of' and 'AND' to 'all of': if you wanted to retrieve articles with any of the search terms, you would use 'OR' to combine the terms; if you wanted to retrieve articles with all of the search terms, you would use 'AND' to combine the terms.

Usually, you would use the Boolean operator 'OR' for combining controlled vocabulary terms and free text terms of the same concept (or sub-concept) and the Boolean operator 'AND' to combine different concepts (or sub-concepts).

In Example 1, the search concepts for designing the search strategy were (a) population: *'children'*, (b) intervention: *'school-based dental screening'*, and (c) comparator: *'no screening'*. If we split the intervention into two sub-concepts: 'school-based' and 'dental screening', the different concepts and sub-concepts to be combined with Boolean operator 'AND' are as follows.

1. Concept_1: *'children'*
2. Concept_2a: *'school-based'*
3. Concept_2b: *'dental screening'*
4. Concept_3: *'no screening'*

Within each of the concept and sub-concept, a combination of controlled vocabulary and free text terms should be used. Therefore, the search strategy in each database will be as follows.

> "((controlled vocabulary terms for *'children'*) OR (free text terms for *'children'*)) AND
> ((controlled vocabulary terms for *'school-based'*) OR (free text terms for *'school-based'*)) AND
> ((controlled vocabulary terms for *'dental screening'*) OR (free text terms for *'dental screening'*)) AND
> ((controlled vocabulary terms for *'no screening'*) OR (free text terms for *'no screening'*))"

While the controlled vocabulary terms may differ between the databases, the free text terms can be similar across databases. However, the way the different search platforms and databases handle wildcard characters, truncated search terms, and a phrase may be different; therefore, extra caution is necessary when you use wildcard characters, truncated search terms or phrases in free text terms.

When combining terms, it is important to ensure that you have used the brackets correctly; otherwise, the searches may be performed erroneously. For example, "Concept_1 AND (Concept_2 OR Concept_3)" results in different search results compared to "(Concept_1 AND Concept_2) OR Concept_3" although both sets of search terms include the same concepts and Boolean operators in the same order of appearance: the differences are in the way that brackets appear. In the first situation, Concept_2 is combined with Concept_3 using the Boolean operator 'OR' first; this 'first set of results' is then combined with Concept_1 using the Boolean operator 'AND'. In the second situation, Concept_1 is combined with Concept_2 using the Boolean operator 'AND' first; this 'first set of results' is then combined with Concept_3 using the Boolean operator 'OR'. This is because the databases process the terms within brackets first before processing the remaining terms. Brackets can also be 'nested' within other brackets. For example, in the search strategy, "((Concept_1 OR Concept_2)) AND (Concept_3 AND (Concept_4 OR Concept_5))", the innermost brackets are processed first, that is, "(Concept_1 OR Concept_2)" ('first set of results') and "(Concept_4 OR Concept_5)" ('second set of results'). Then "Concept_3" is combined with "(Concept_4 OR Concept_5)" (that is, the second set of results) to get the 'third set of results'. Finally, the first and third sets of results are combined to obtain the final results. The order in which the databases process the search terms within brackets is the same as the way expressions within brackets are processed in mathematical equations: mathematical expressions in the innermost brackets are processed first and those in outer brackets are processed next, with multiple levels of nesting possible.

Other Boolean operators

Some databases allow Boolean operators 'NOT' and proximity search operators, for example 'ADJ'. The Boolean operator 'NOT' can be translated as 'but not containing'. For example, "Concept_1 NOT Concept_2" retrieves all articles containing Concept_1 but not containing Concept_2. This is acceptable when the articles contain only one of the two concepts. However, articles can contain more than one concept. In such cases, the articles containing both the concepts are not retrieved. This is usually not desirable. Therefore, we recommend against using the Boolean operator 'NOT' without discussing the implications with an information specialist or librarian with expertise in designing search strategies.

The Boolean operator 'ADJ' can be translated as 'all of the terms in close proximity to each other'. For example, the search terms "Concept_1 ADJ5 Concept_2" will retrieve articles in which Concept_1 appears within five words of Concept_2. While 'ADJ' is the Boolean operator for Ovid and 'NEAR/' in Web of Science, in PubMed, this is indicated as Concept_1 Concept_2[tiab:~5], noting that proximity search is only available for searching the title, abstract and affiliation fields of the article. [19] Boolean operators for proximity can be used to restrict the results to manageable limits.

Step 3.3.10 Search filters

Search filters are validated combinations of search terms related to a concept and can be used when appropriate. The InterTASC Information Specialists' Subgroup Search Filter Resource [20] is an excellent resource for search filters. We recommend searching this resource to identify the appropriate search filter. Most search filters available relate to the study design, although other filters are also available. The purpose of these search filters is to improve the precision of the retrieved searches, that is, to decrease the number of unwanted results. For example, if a systematic review includes only RCTs, you can use a search filter for RCTs to filter out (exclude) the non-randomised studies, which do not meet the eligibility criteria of the review.

Step 3.3.11 Revising the search strategy

Developing formal search strategies is an iterative process. Therefore, search strategies may require revisions before being finalised. But how do you revise a search strategy?

Step 3.3.11.1 Evaluation of the draft search strategy

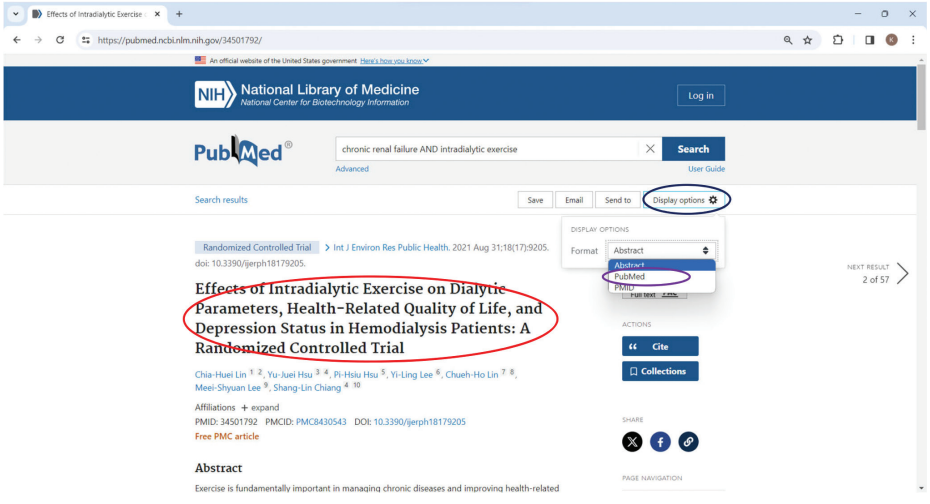
The first step is to evaluate the draft search strategy. One way of evaluating the draft search strategy is check whether the strategy identifies articles relevant to your systematic review from a previous systematic review on the topic. For example, you might be interested only in RCTs while the previous systematic review might include non-randomised studies too. You would not expect your draft search strategy to retrieve non-randomised studies if you have used a filter for RCTs, but you could expect that the RCTs are identified.

If there are no relevant systematic reviews, you can check whether the search results include some of the major studies of which you are aware.

Step 3.3.11.2 Finding additional controlled vocabulary and free text terms

If the search strategy misses some of the studies, you can look at the controlled vocabulary and free text terms of the missed studies and consider including them in the revised search strategy. This can be done in the following way in PubMed.

Figure 4.29 Finding additional controlled vocabulary and free text terms: missed reference in PubMed



Step 3.3.11.2.1

Identify the missed reference in PubMed (red ellipse in Figure 4.29). Click 'Display options' (blue ellipse) and choose 'PubMed' from the dropdown menu (purple ellipse).

Step 3.3.11.2.2

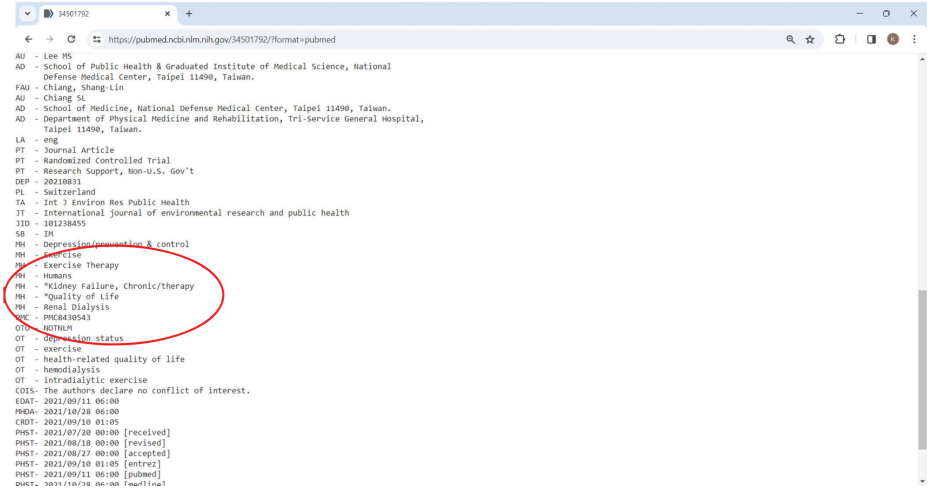
You can identify the controlled vocabulary terms with the tag 'MH' (red ellipse) in the resulting screen (Figure 4.30).

Other issues related to formal search strategies

Are there any alternatives to designing a formal search strategy?

An alternative to designing your own formal search strategy is to use a previously existing formal search strategy. This might be a possibility if you are updating a previous high-quality systematic review. You should reference such a search strategy. However, even if you are not updating a previous high-quality systematic review, you may be able to use part of the search strategies used in high-quality systematic reviews. A good knowledge of how concepts and sub-concepts are combined using Boolean operators is necessary if you use this approach. In Example 2, the concepts for the formal search strategy are (a) population: *'people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma'*, (b) intervention: *'laparoscopic distal pancreatectomy'*, and (c) comparator: *'open distal pancreatectomy'*. Let us say that there are no high-quality systematic reviews on the topic, but that

Figure 4.30 Finding additional controlled vocabulary and free text terms: missed reference in PubMed with PubMed display option



there is a high-quality systematic review comparing laparoscopic and open distal pancreatectomy for chronic pancreatitis and another one comparing pylorus sparing pancreaticoduodenectomy versus classical Whipple for pancreatic ductal adenocarcinoma. You can obtain the search strategy related to the concepts *'laparoscopic and open distal pancreatectomy'* from the first review and that related to the concept *'pancreatic ductal adenocarcinoma'* from the second review. These concepts can then be combined using the appropriate Boolean operators. If you choose to use search strategies from other systematic reviews fully or partially, you should ensure that you indicate clearly that you used search strategies from other systematic reviews and cite the sources.

Absence of controlled vocabulary terms for a concept

There may not be controlled vocabulary terms for all concepts. In such cases, sometimes, controlled vocabulary terms may exist for some sub-concepts. In such a situation, the controlled vocabulary terms and free text terms of the sub-concept for which controlled vocabulary terms are available can be combined with the free text terms of the sub-concept for which there are no controlled vocabulary terms, using the appropriate Boolean operator 'AND'. For example, a search of MeSH database for the concept '*robotic cholecystectomy*' does not reveal any MeSH term. However, a MeSH term for *'cholecystectomy'* is available. In such a situation, the search strategy for *'robotic cholecystectomy'* includes "(free text terms for '*robotic*') AND (controlled vocabulary terms for *'cholecystectomy'* OR free text terms for *'cholecystectomy'*)". We used the Boolean operator

'AND' to combine the sub-concepts because we are interested in only articles containing both the sub-concepts. Sometimes, there are no controlled vocabulary terms for any of the sub-concepts. In such cases, the search terms for the concept include only free text terms: "(free text terms for sub-concept 1) AND (free text terms for sub-concept 2)".

Tools to aid developing search strategies

There are currently no tools available that automatically generate the search strategy if a concept (or sub-concept) is entered. Some prototype and pilot tools are available to help with managing already identified search strategies. For example, Kamdar et al. provide a tool to help with the appropriate tags and parentheses (brackets); [21] Wanner et al. provide a tool to help with conversion of MEDLINE strategies between PubMed and Ovid platforms; [22] Bramer et al. provide conversion between EMBASE and MEDLINE search strategies; [23] and Ovid provides its own tool to translate PubMed-based search strategy to the Ovid platform (https://tools.ovid.com/ovidtools/translate.html).

Mapping of SNOWMED CT to ICD-10 is available through Unified Medical Language System® (UMLS®). [18] MeSH terms are also available through the UMLS® system, but the MeSH terms retrieved do not always appear to be related. Therefore, further research is necessary to determine if automated development of search strategies is possible through the mapping of different systems available for coding medical concepts.

Supplementing formal search strategies

Contacting experts

We recommend that experts in the field are contacted to identify further trials. The term 'contacting experts' may be interpreted in different ways. A practical way of implementing this recommendation is to contact the trialists who investigate the research question and guideline developers who incorporate the results of the research question in their clinical practice guidelines.

Reference search

It is recommended that references of included articles and any relevant systematic reviews are searched to identify further trials. Searching references of systematic reviews can also be used to evaluate the draft search strategy as mentioned in Step 3.3.11.1.

Similar articles and cited reference search

The 'Similar articles' function in PubMed or 'Find Similar' function in Ovid identify articles similar to another article. The 'Cited by' function in PubMed or the

'Find Citing articles' function in Ovid identify articles that cite a study. The value of supplementing formal search strategies with searching for 'similar articles' or 'citing articles' is not known. These functions may be used to find more relevant studies while evaluating the draft search strategy.

Web search engines

Managing searches of web search engines can be time consuming. The value of supplementing formal search strategies with searching through web search engines is not known.

Step 4 Run the searches

Having identified the search terms for each concept (and sub-concept), the next step is to run these searches in the different databases and obtain the results. This is demonstrated step by step, refining one of the examples that we used to develop controlled vocabulary terms and free text terms (Example 3). In Example 3, we described only one concept: chronic kidney failure. Let us consider a full research question for explaining how to run the searches and obtain results. Only the first draft of the results is presented. Some specific aspects to be considered for revision if studies are missed are also mentioned.

Example 6

A researcher wants to:

> *Evaluate the effectiveness of intradialytic exercise versus no intradialytic exercise on improving mortality and health-related quality of life in children or adults with chronic kidney failure by a systematic review and meta-analysis of randomised controlled trials*.

In this example, the population is *'children or adults with chronic kidney failure'*, the intervention is *'intradialytic exercise'*, the comparator is *'no intradialytic exercise'*, and the outcomes are *'mortality and health-related quality of life'*.

Since we do not usually use outcomes as eligibility criteria, the search concepts that we should use for identifying relevant studies are as follows.

1. Population: Concept_1: *'children or adults with chronic kidney failure'*
2. Intervention: Concept_2: *'intradialytic exercise'*
3. Comparator: Concept_3: *'no intradialytic exercise'*
4. Study design: Concept_4: *'randomised controlled trial'*

If you further convert the concepts to sub-concepts, the search concepts and sub-concepts are as follows.

1. Population
    a. Sub-concept_1: *'children or adults'*
    b. Sub-concept_2: *'chronic kidney failure'*
2. Intervention:
    a. Sub-concept_3: *'intradialytic'*
    b. Sub-concept_4: *'exercise'*
3. Comparator:
    a. Sub-concept_5: *'no'*
    b. Sub-concept_6: *'intradialytic'*
    c. Sub-concept_7: *'exercise'*
4. Study design: Concept_4: *'randomised controlled trial'*

We can exclude Sub-concept_1 as the population includes all age groups. If this was confined to one age group (say, adults), you might consider using this as a sub-concept. However, this is not the case in this example. We can also exclude Sub-concept_5 since it is difficult to design any search terms for absence of an intervention. We can exclude Sub-concept_6 and Sub-concept_7 since they are the same sub-concepts as Sub-concept_3 and Sub-concept_4. This leaves us with the following concepts and sub-concepts: *'chronic renal failure'* (Sub-concept_2), *'intradialytic'* (Sub-concept_3), *'exercise*' (Sub-concept_4), and *'randomised controlled trial'* (Concept_4).

Table 4.1 shows the controlled vocabulary and free text terms for the different databases obtained by going through the steps described in the section 'Step 3 Choose the search strategy'. This is a draft search strategy, and the free text terms were based on the 'entry' terms found in PubMed MeSH headings, the '[Used for]' terms in Ovid EMBASE, and the common variations in spellings in different countries. Please note that the entry terms were revised so that the terms appear in the way they are usually stated in articles. For example, 'dialysis, renal' was revised to 'renal dialysis'. The search strategy can be refined further depending on whether some eligible studies have been missed and by discussing with clinical experts. Specifically, consider whether the controlled vocabulary terms and free text terms for exercise rather than exercise therapy should be used and whether the appropriate subheadings have been included.

Executing the searches in different databases

Step 4.1 Running the searches in PubMed

Step 4.1.1

The initial screen in PubMed is shown in Figure 4.31. Click 'Advanced' (red ellipse) or 'Advanced Search' (blue ellipse).

Table 4.1 Controlled vocabulary and free text terms for Example 6

Concept or sub-concept[1]	Controlled vocabulary terms (PubMed)	Controlled vocabulary terms (Ovid MEDLINE)	Controlled vocabulary terms (Ovid EMBASE)	Controlled vocabulary terms (Cochrane Library)	Free text terms[2]
Sub-concept_2: *'chronic kidney failure'*	("Kidney Failure, Chronic/ complications"[Mesh] OR "Kidney Failure, Chronic/ mortality"[Mesh] OR "Kidney Failure, Chronic/ nursing"[Mesh] OR "Kidney Failure, Chronic/ rehabilitation"[Mesh] OR "Kidney Failure, Chronic/ therapy"[Mesh])	exp Kidney Failure, Chronic/ co, mo, nu, rh, th	exp chronic kidney failure/co, dm, dt, rh, si, th	MeSH descriptor: [Kidney Failure, Chronic] explode all trees	(chronic kidney failure OR chronic renal failure OR end stage kidney disease OR end-stage kidney disease OR end-stage renal disease OR end stage renal disease OR ESRD OR end-stage renal failure OR end stage renal failure OR chronic kidney disease OR chronic kidney insufficiency OR chronic nephropathy OR chronic renal disease OR chronic renal insufficiency OR kidney chronic failure OR kidney function, chronic disease)
Sub-concept_3: *'intradialytic'*	("Renal Dialysis/ adverse effects"[Mesh] OR "Renal Dialysis/ mortality"[Mesh] OR "Renal Dialysis/ nursing"[Mesh])	exp Renal Dialysis/ae, mo, nu, rh	exp renal replacement therapy/rh	MeSH descriptor: [Renal Dialysis] explode all trees	(extracorporeal dialyses OR extracorporeal dialysis OR renal dialyses OR renal dialysis OR hemodialyses OR hemodialysis OR haemodialyses OR haemodialyses OR dialysis therapy OR dialysis treatment OR kidney dialysis OR kidney replacement therapy OR kidney support OR renal support)

Sub-concept_4: *'exercise'*[3]	("Exercise Therapy/ adverse effects"[Mesh] OR "Exercise Therapy/ mortality"[Mesh] OR "Exercise Therapy/ nursing"[Mesh])	exp Exercise Therapy/ae, co, mo, th	exp kinesiotherapy/	MeSH descriptor: [Exercise Therapy] explode all trees	(rehabilitation exercise OR rehabilitation exercises OR exercise therapies OR exercise therapy OR remedial exercise OR remedial exercises OR corrective exercise OR exercise movement techniques OR exercise therapy OR exercise treatment OR kinesiotherapeutic intervention OR kinesiotherapeutic method OR kinesiotherapeutic procedure OR kinesiotherapeutic technique OR kinesiotherapeutical treatment OR kinesitherapeutic exercises OR kinesitherapeutic intervention OR kinesitherapeutic method OR kinesitherapeutic methodology OR kinesitherapeutic procedure OR kinesitherapeutic technique OR kinesitherapeutic treatment OR kinesitherapeutical treatment OR kinesitherapy OR SKTM OR specialised kinesitherapeutic methodology OR specialized kinesitherapeutic methodology OR therapeutic exercise)

(*Continued*)

Table 4.1 Controlled vocabulary and free text terms for Example 6 (*Continued*)

Concept or sub-concept[1]	Controlled vocabulary terms (PubMed)	Controlled vocabulary terms (Ovid MEDLINE)	Controlled vocabulary terms (Ovid EMBASE)	Controlled vocabulary terms (Cochrane Library)	Free text terms[2]
Concept_4: 'randomised controlled trial'[4]	#1 randomized controlled trial [pt] #2 controlled clinical trial [pt] #3 randomized [tiab] #4 placebo [tiab] #5 drug therapy [sh] #6 randomly [tiab] #7 trial [tiab] #8 groups [tiab] #9 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 #10 animals [mh] NOT humans [mh] #11 #9 NOT #10	1 exp randomized controlled trial/ 2 controlled clinical trial.pt. 3 randomized.ab. 4 placebo.ab. 5 drug therapy.fs. 6 randomly.ab. 7 trial.ab. 8 groups.ab. 9 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 10 exp animals/ not humans.sh. 11 9 not 10	1 exp randomized controlled trial/ 2 controlled clinical trial/ 3 random$.ti,ab. 4 randomization/ 5 intermethod comparison/ 6 placebo.ti,ab. 7 (compare OR compared OR comparison).ti,ab. 8 ((evaluated OR evaluate OR evaluating OR assessed OR assess) AND (compare OR compared OR comparing OR comparison)) 9 (open adj label).ti,ab. 10 ((double OR single OR doubly OR singly) adj (blind OR blinded OR blindly)).ti,ab. 11 double blind procedure/ 12 parallel group$1.ti,ab. 13 (crossover OR cross over).ti,ab. 14 ((assign$ OR match OR matched OR allocation) adj5 (alternate OR group$1 OR intervention$1 OR patient$1 OR subject$1 OR population$1)).ti,ab. 15 (assigned OR allocated).ti,ab. 16 (controlled adj7 (study OR design OR trial)).ti,ab. 17 (volunteer OR volunteers).ti,ab. 18 human experiment/ 19 trial.ti. 20 or/1–19 21 (random$ adj sampl$ adj7 ("cross section$" OR questionnaire$1 OR survey$ OR database$1)).ti,ab. NOT (comparative study/ OR controlled	No filter	Not applicable as this is part of filter If the number of references in Citation Proceedings or Trial registries is high, consider using the term 'random*'

			study/ OR randomi?ed controlled.ti,ab. OR randomly assigned.ti,ab.) 22 cross-sectional study/de NOT (exp randomized controlled trial/ OR controlled clinical trial/ OR controlled study/ OR randomi?ed controlled.ti,ab. OR control group$1.ti,ab.) 23 ((case adj control$) AND random$.ti,ab.) NOT randomi?ed controlled.ti,ab. 24 (systematic review.ti,ab. NOT (trial OR study)).ti. 25 (nonrandom$ NOT random$).ti,ab. 26 "random field$".ti,ab. 27 (random cluster adj3 sampl$).ti,ab. 28 (review.ab. AND review.pt.) NOT trial.ti. 29 ("we searched".ab. AND (review.ti. OR review.pt.)) 30 "update review".ab. 31 (databases adj4 searched).ab. 32 (rat OR rats OR mouse OR mice OR swine OR porcine OR murine OR sheep OR lambs OR pigs OR piglets OR rabbit OR rabbits OR cat OR cats OR dog OR dogs OR cattle OR bovine OR monkey OR monkeys OR trout OR marmoset$1).ti. AND animal experiment/ 33 animal experiment/ NOT (human experiment/ OR human/) 34 or/21–33 35 20 NOT 34		

The table shows the controlled vocabulary terms and free text terms for some concepts and sub-concepts.

Notes

1. Please see the text for information about why search strategies were developed only for some concepts and sub-concepts.
2. Free text terms were based on the 'entry terms' in PubMed and '[Used For]' terms in Ovid EMBASE.
3. The controlled vocabulary terms for 'Exercise therapy' rather than exercise were used since the controlled vocabulary terms for 'Exercise' appears to relate to physiology rather than a treatment option in people with illness.
4. The sensitive search filters from (24) were used for PubMed, MEDLINE, and EMBASE. It should be noted that the numbers in front of the lines should not be used when these terms are entered in the search boxes.

Figure 4.31 Running the searches in PubMed: landing page

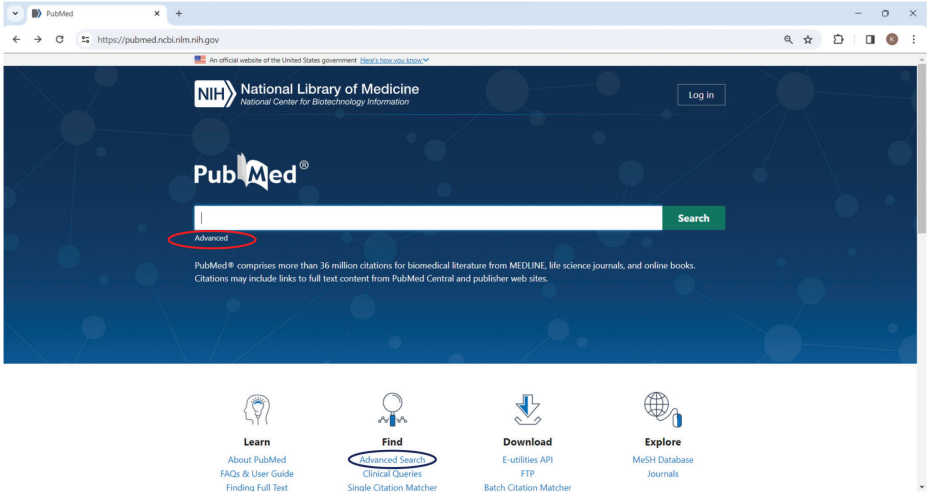
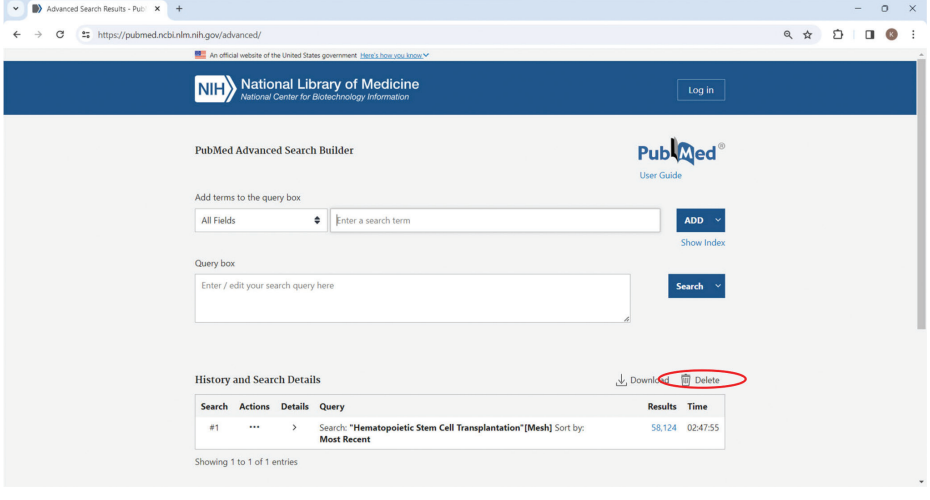


Figure 4.32 Running the searches in PubMed: advanced search interface



Step 4.1.2

Ensure that 'History and Search Details' is empty, to avoid any confusion. In the screenshot (Figure 4.32), there is an existing query. We can delete the previous query by clicking 'Delete' (red ellipse).

The screenshot after deletion is shown in Figure 4.33.

Step 4.1.3

Enter the controlled vocabulary terms for 'chronic kidney failure' (Sub-concept_2) in the 'Query box' (red ellipse in Figure 4.34). Click 'Search' (blue ellipse).

Figure 4.33 Running the searches in PubMed: advanced search interface after deletion of previous searches

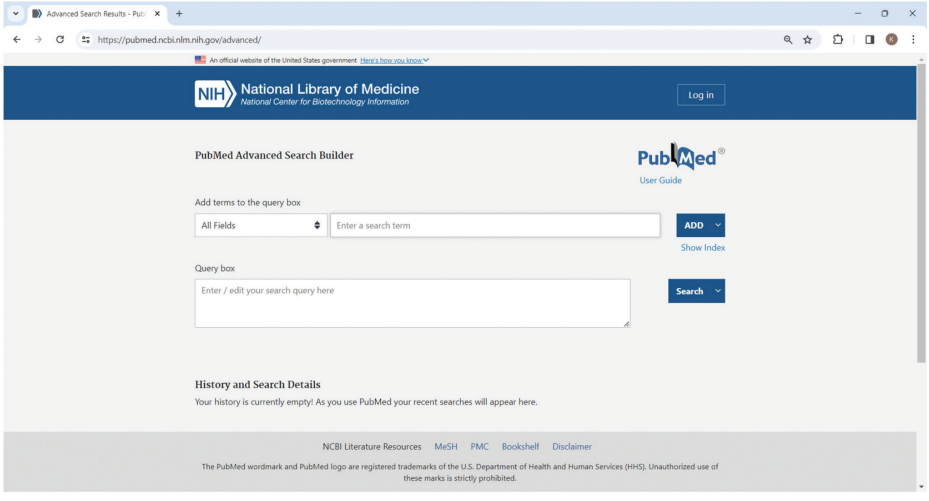
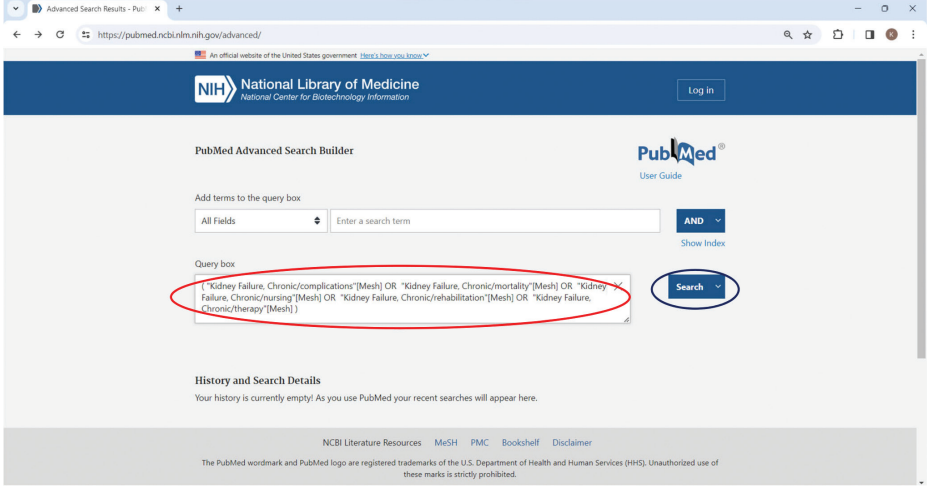


Figure 4.34 Running the searches in PubMed: query box



Step 4.1.4

This results in the screen shown in Figure 4.35. You can clear the search box (red ellipse) and enter the next set of terms in the search box (blue ellipse). However, this usually causes confusion, as it is not clear whether the search results correspond to the latest set of terms. Therefore, we recommend that you click 'Advanced' (purple ellipse) and enter the searches in the 'Query box' as shown in Figure 4.36.

Figure 4.35 Running the searches in PubMed: search results after first set of search terms

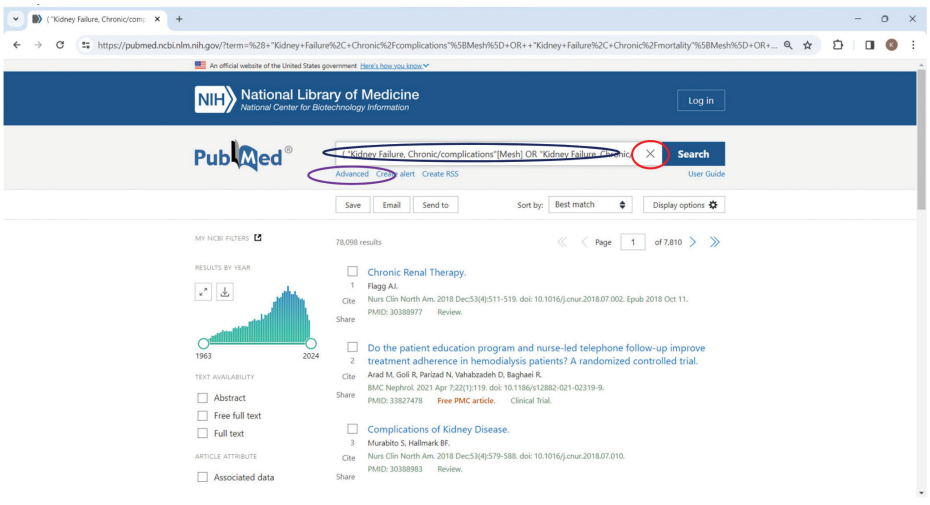
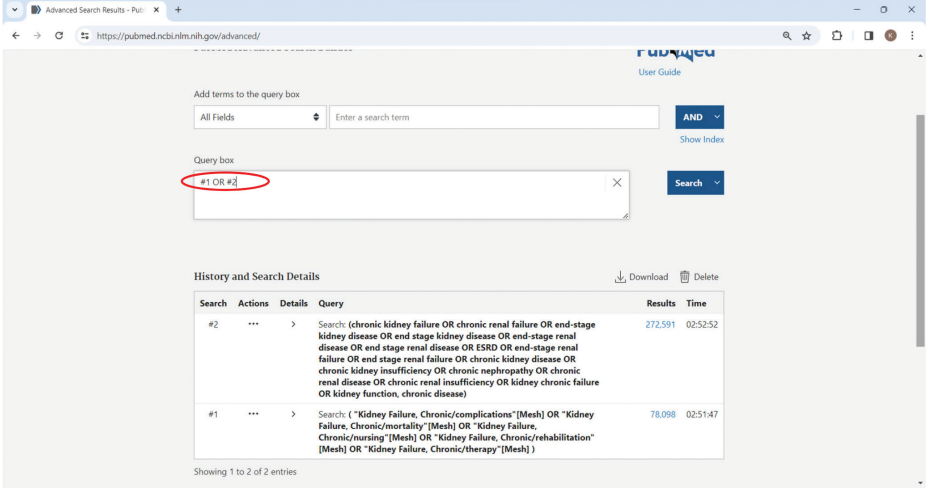


Figure 4.36 Running the searches in PubMed: combining controlled vocabulary and free text terms



Step 4.1.5

Figure 4.36 shows the controlled vocabulary and free text terms entered for 'chronic kidney failure'. We should now combine these controlled vocabulary and free text terms using the Boolean operator 'OR'. However, there is no requirement to type the searches again; instead, we combine the search

number #1 with search number #2 using the Boolean operator 'OR' in the 'Query box' as shown in the screenshot (red ellipse). Note the '#' in front of the search number.

Step 4.1.6

Figure 4.37 shows all the search concepts and sub-concepts entered. The important thing to note is that the controlled vocabulary and free text terms of each sub-concept are combined with the Boolean operator 'OR'. Also note the search numbers #18 and #20 (red ellipse). These are different from the numbers in Table 4.1. This is because the search numbers in Table 4.1 refer to the search numbers within RCT (Concept_4) and not across all the concepts or sub-concepts. The search numbers #1 to #8 in Table 4.1 correspond to search numbers #10 to #17 in the screen. If adequate care is not taken in revising the search numbers, it can lead to wrong results. Therefore, you might want to enter the search filter for RCTs first and then the remaining concepts. The final results can be obtained by clicking on the results column of the final search terms corresponding to search number #21 (blue ellipse).

Step 4.1.7

The results are shown along with any warnings. The warning in this screenshot (Figure 4.38) shows that there were no studies using the term 'kinesitherapeutical' (red ellipse). You might consider removing this term from the searches, but it does not change the results. Besides, a future study may use this term, which

Figure 4.37 Running the searches in PubMed: advanced search interface after entry of all search terms

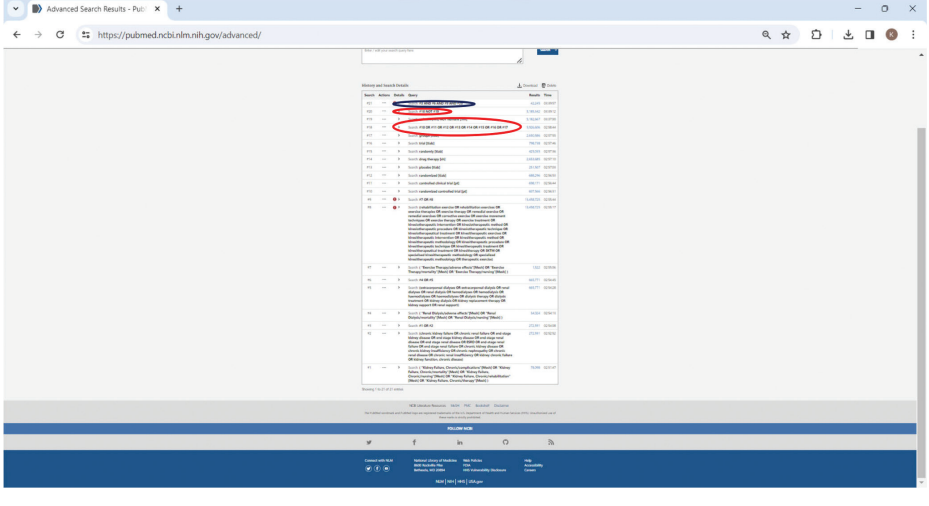
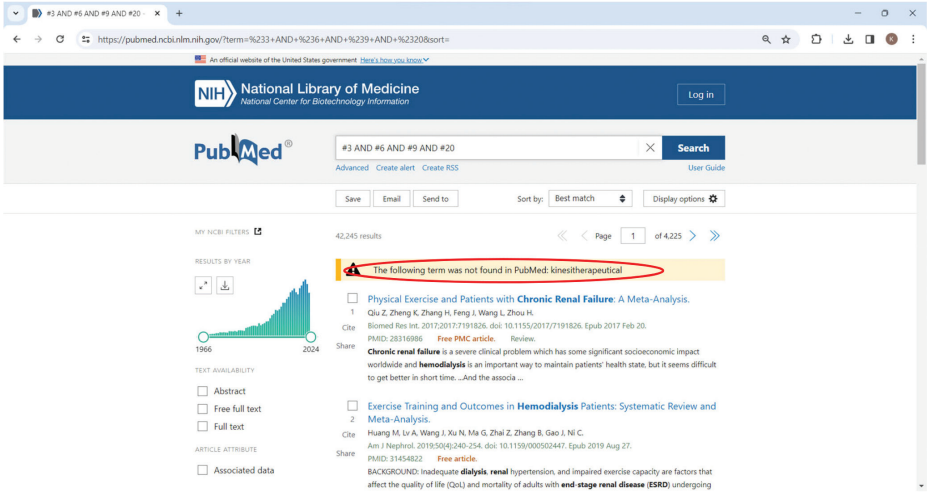


Figure 4.38 Running the searches in PubMed: search results with warnings



was identified from '[Used For]' terms in Ovid EMBASE. Therefore, you can retain this term. However, there is a major problem with the search: the number of results returned is unmanageable. You might want to consider whether all the subheadings are relevant and whether all the free text terms are appropriate and restrict the free terms to only the relevant terms. You might also consider restricting the search results to those that contain the term 'intradialytic' (including the common variation 'intra-dialytic'). This might result in loss of some relevant references; however, you need to balance between attempting a systematic review which cannot be completed easily without mammoth resources versus accepting that you might miss some studies that do not use these terms ('intradialytic' or 'intra-dialytic') in the title, abstract or keywords.

Restricting the searches to studies that contain the terms ('intradialytic' or 'intra-dialytic') results in a more manageable number of references (red ellipse in Figure 4.39).

Step 4.1.8

For exporting references, click on the results as before. In the resulting screen (Figure 4.40), click 'Send to' (red ellipse) and then 'Citation manager' in the resulting options (blue ellipse).

Step 4.1.9

This results in the screen shown in Figure 4.41. Click 'Selection: All results on this page' (red ellipse) and change the 'Selection' to 'All results' (not shown in the figure). Then click 'Create file' (blue ellipse). This sends the search results to a

Figure 4.39 Running the searches in PubMed: advanced search interface after restricting searches

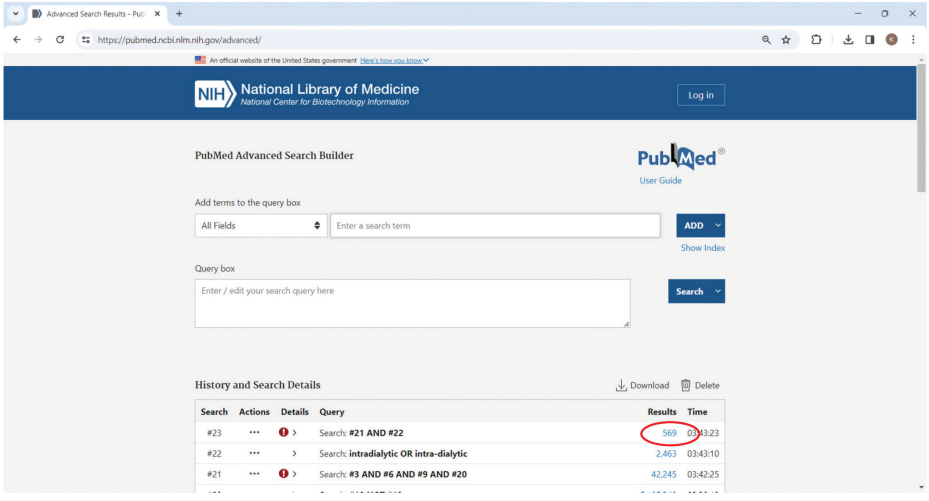
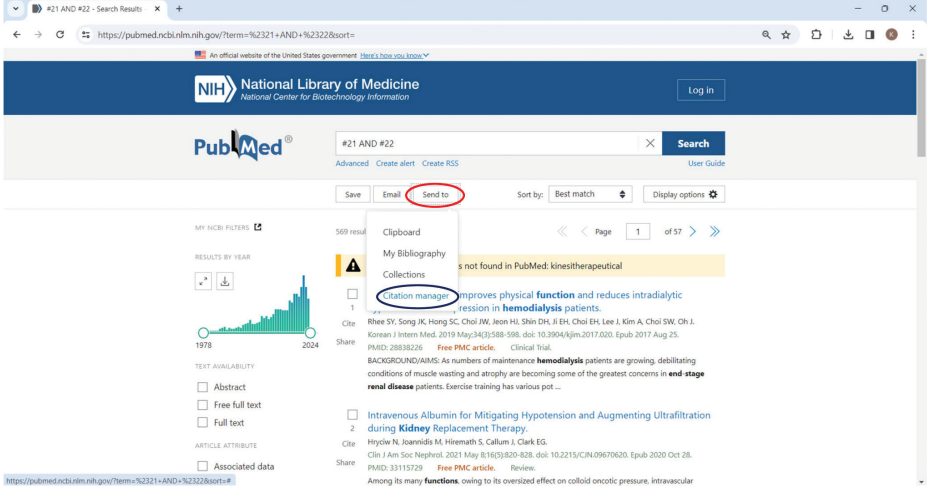


Figure 4.40 Running the searches in PubMed: exporting search results interface



file, usually found in the 'Download' folder, unless the settings in the computer have been changed.

Step 4.2 Running the searches in other platforms

The general principles and process for entering the search terms in the different databases and platforms are similar. Only the major differences are shown in the following.

Figure 4.41 Running the searches in PubMed: exporting search results final step

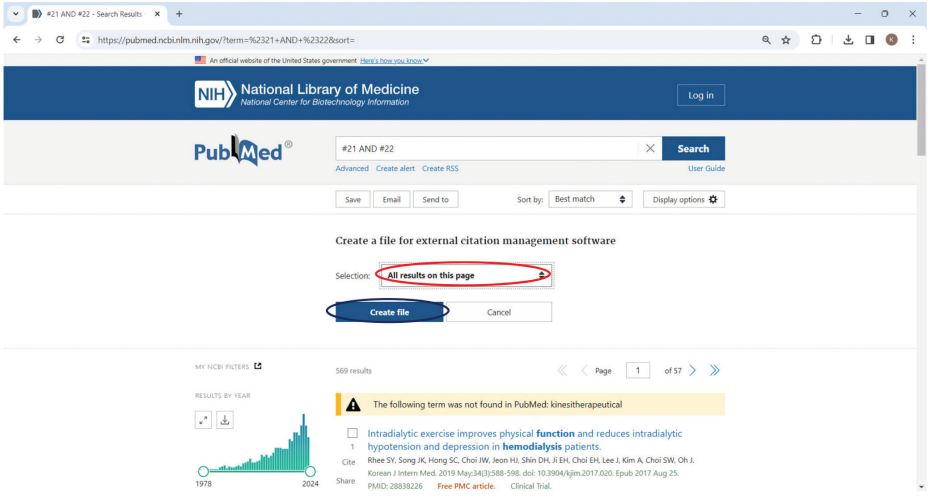
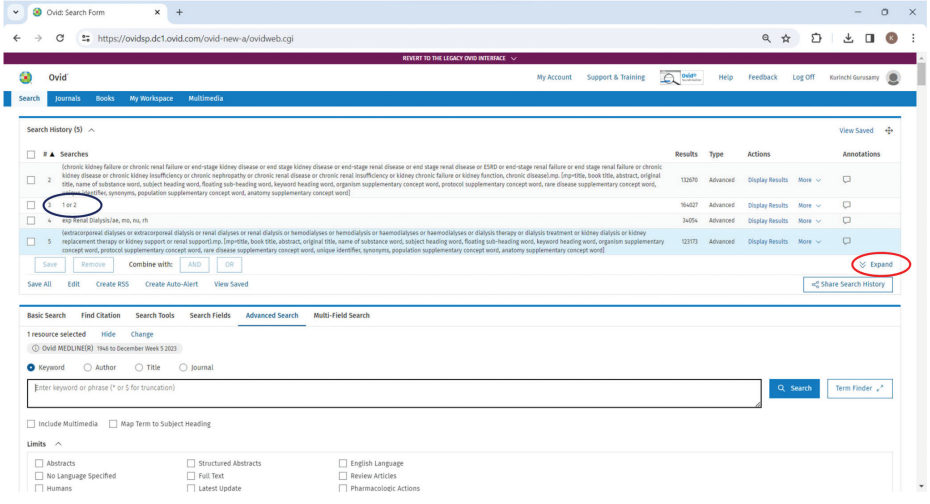


Figure 4.42 Running the searches in Ovid: advanced search interface with multiple search terms entered



Step 4.2.1 Ovid

Step 4.2.1.1

In Ovid, by default, only a small portion of the screen is allocated to search lines (See Figure 4.42.). This means that the full searches may not be visible. You can view the full search lines by clicking 'Expand' (red ellipse). Also note the search lines are referred to without a '#' before the search line number in Ovid. This is why the search line 3 has '1 or 2' (blue ellipse) and not '#1 or #2'.

Step 4.2.1.2

The results of running the searches are shown in Figure 4.43. Surprisingly, there are only 128 references (red ellipse) returned, compared to an equivalent search strategy in PubMed which returned more than 40,000 references. If there is such a huge discrepancy, you must recheck whether the correct database has been selected and whether the searches have been entered correctly. The searches seem to have been entered correctly in both the databases. The differences may be due to the way that different words in the phrases in the free text terms are combined.

Once you have run all the search strings and combined them appropriately, you can click 'Display Results' (blue ellipse).

Step 4.2.1.3

To export the references, enter the references to export (red ellipse in Figure 4.44). Click 'Export' (blue ellipse).

Usually, there is a limit of exporting 3,000 references at a time. If your searches return more than 3,000 references, you can click 'Clear' (purple ellipse) and enter the new set of references, for example, 3001–6000.

Step 4.2.1.4

In the resulting window, shown in Figure 4.45, click 'Export' ('Export' window), choose 'Excel sheet' as the format (red ellipse) and 'Complete reference' as the field (blue ellipse) from the dropdown menus, then click 'Export' (purple ellipse).

Figure 4.43 Running the searches in Ovid: advanced search interface after entry of all search terms

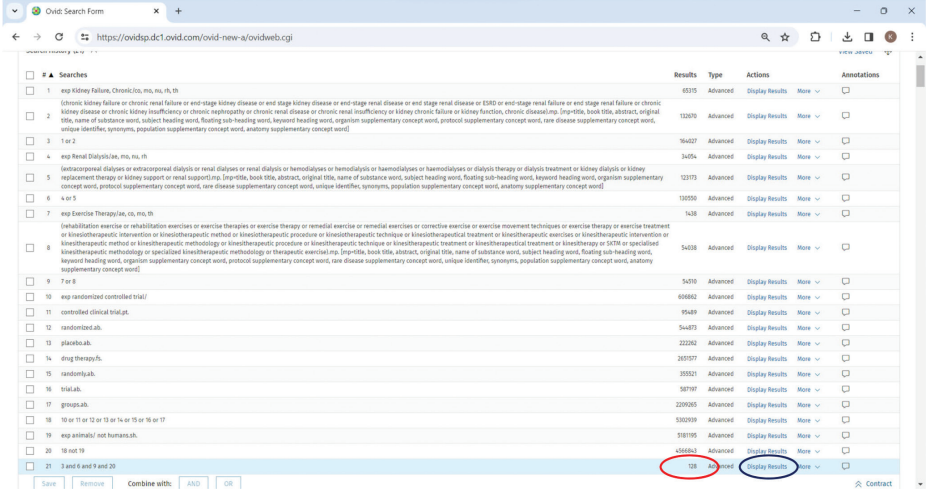


Figure 4.44 Running the searches in Ovid: selecting references for export

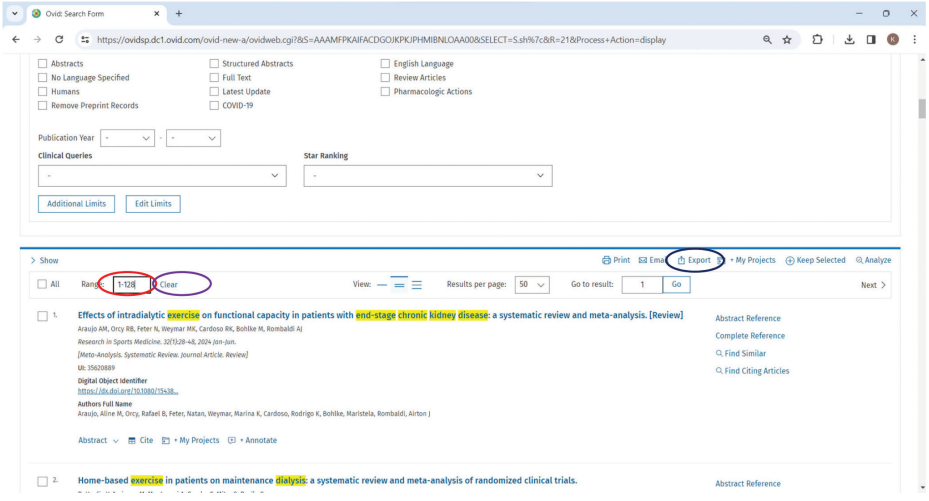
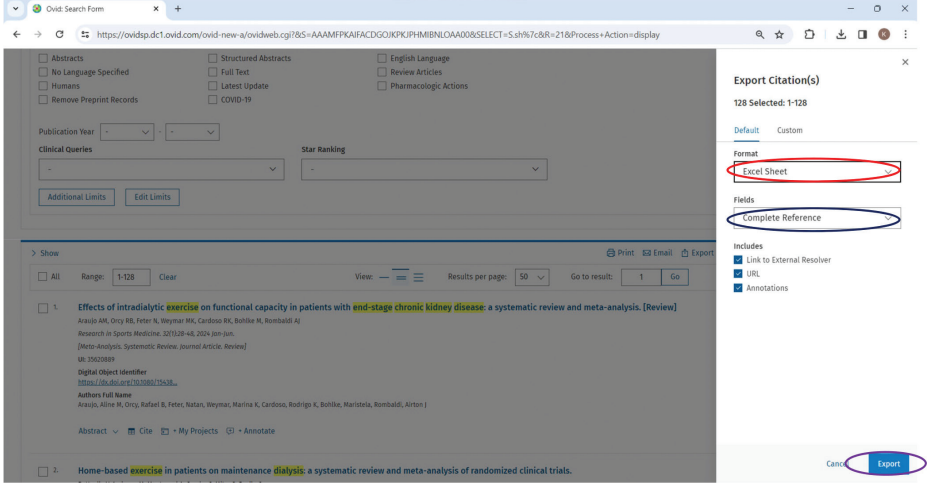


Figure 4.45 Running the searches in Ovid: exporting search results interface

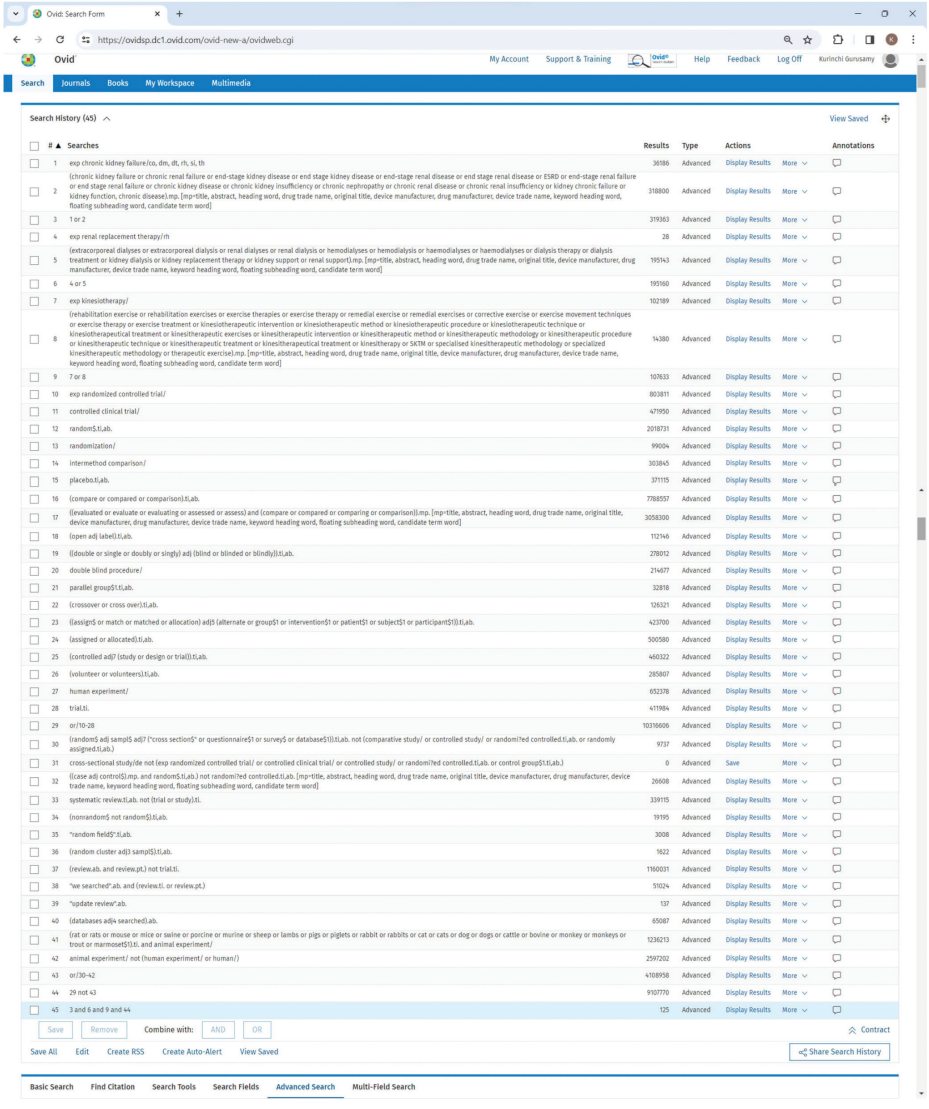


This results in an Excel file which is usually in the 'Downloads' folder. Please note that there is no option to change the references that you want to export from the 'Export' window. You have to change this as described in Step 4.2.1.3.

Step 4.2.1.5

The EMBASE search strategy in the Ovid platform is shown in Figure 4.46. The export is similar to that in Ovid MEDLINE. As before, choose 'Excel sheet' as the format and 'Complete reference' as the field.

Figure 4.46 Running the searches in Ovid: advanced search interface after entry of all search terms



Step 4.2.2 Cochrane Library

Step 4.2.2.1

For Cochrane Library, even if the MeSH terms had been previously identified, they have to be re-entered as in Step 3.3.4. In other words, you cannot simply copy and paste the search strings as for the PubMed or Ovid platforms.

Step 4.2.2.2

Each time you want to include a controlled vocabulary term, you need to follow the steps outlined in Step 3.3.4.

Step 4.2.2.3

You can directly enter the free text terms in the search box (red ellipse in Figure 4.47) and click 'Continue' (blue ellipse).

Step 4.2.2.4

To combine the searches, you can enter the search numbers with a '#' in front of the search number, as for PubMed (red ellipse in Figure 4.48).

Step 4.2.2.5

An alternative to entering the MeSH terms in the MeSH tab is to click the drop-down box in MeSH (red ellipse in Figure 4.49) which opens a new window as shown, which is like the window that appears when the MeSH tab is clicked.

Step 4.2.2.6

Figure 4.50 shows the searches entered in the Cochrane Library. To view the results, click on the references in the last line (red ellipse). This results in the references being displayed.

Step 4.2.2.7

To export the references, click the 'Trials' tab (red ellipse in Figure 4.51) and click 'Select all' (blue ellipse). Then click 'Export selected references' (purple ellipse).

Figure 4.47 Running the searches in Cochrane: free text search

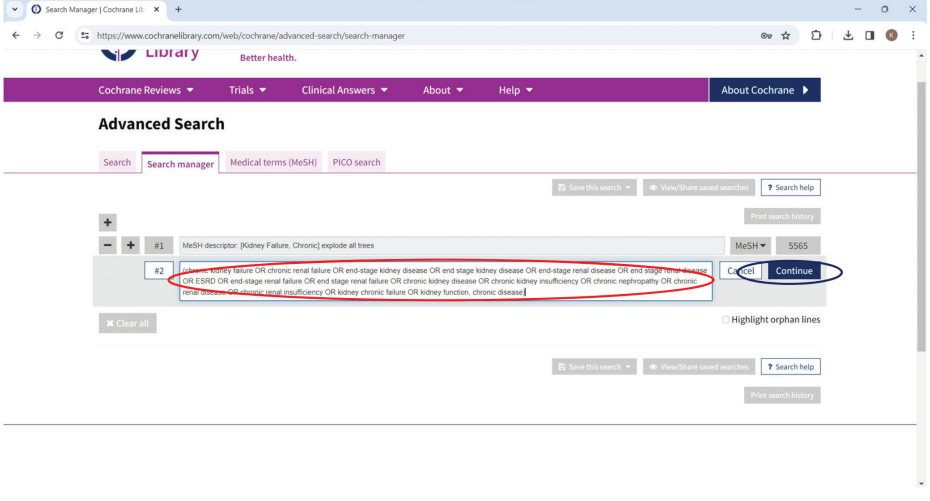


Figure 4.48 Running the searches in Cochrane: combining controlled vocabulary and free text terms

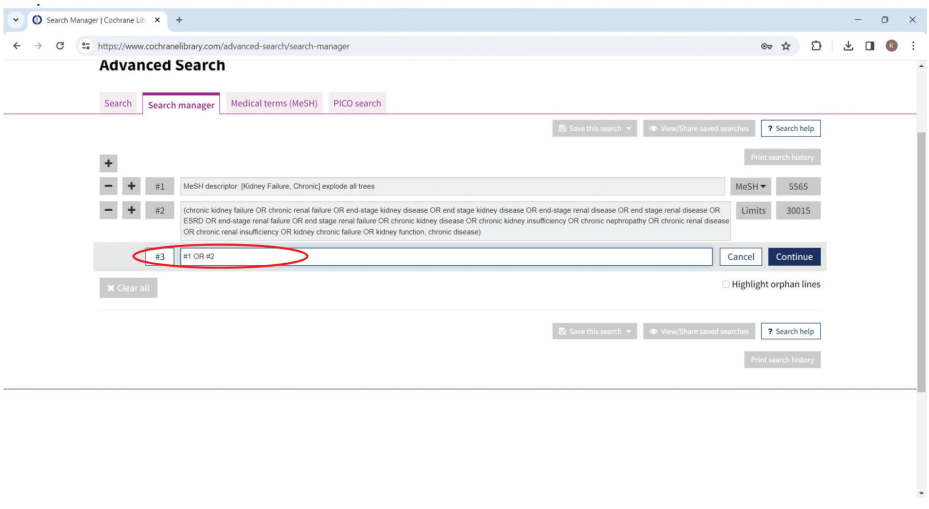
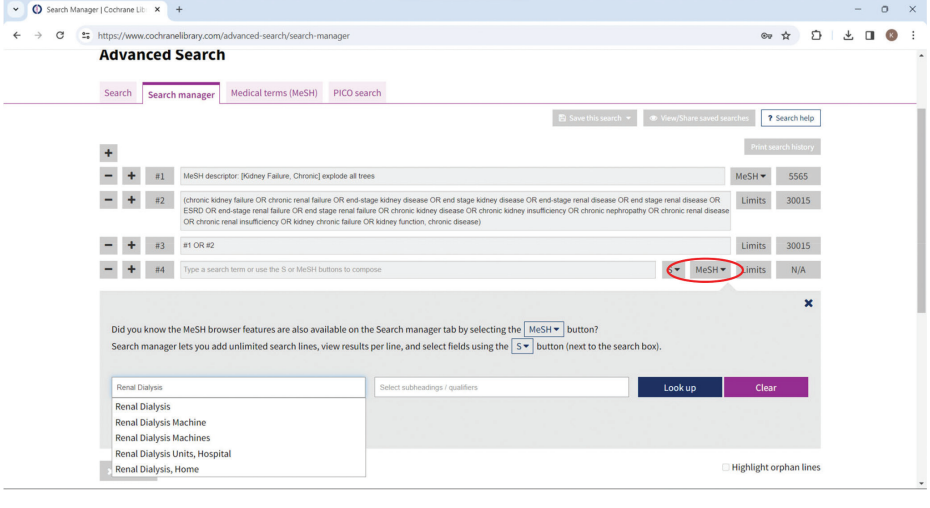


Figure 4.49 Running the searches in Cochrane: MeSH search through search manager



On the resulting screen (Figure 4.52), click 'CSV(Excel)' (red ellipse) and ensure that 'Include abstract' (blue ellipse) is selected. This is selected by default. Then click 'Download' (purple ellipse).

FIGURE 4.50 Running the searches in Cochrane: search results

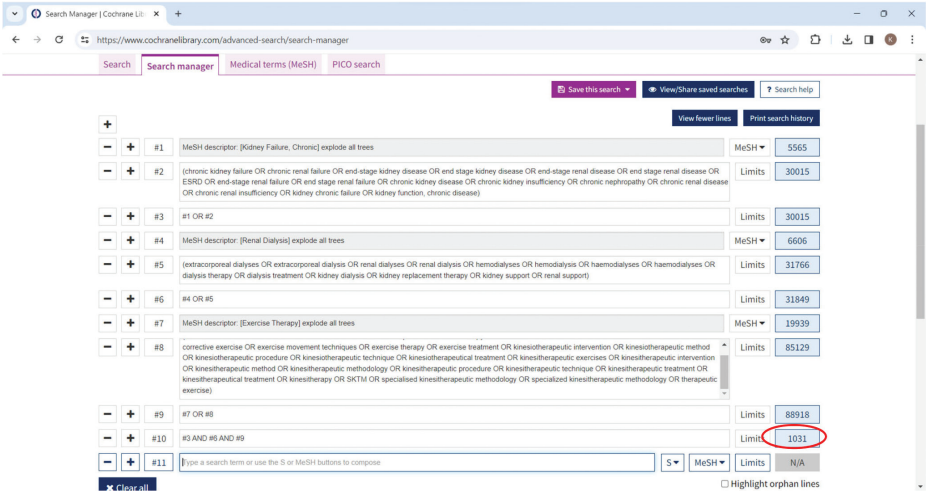
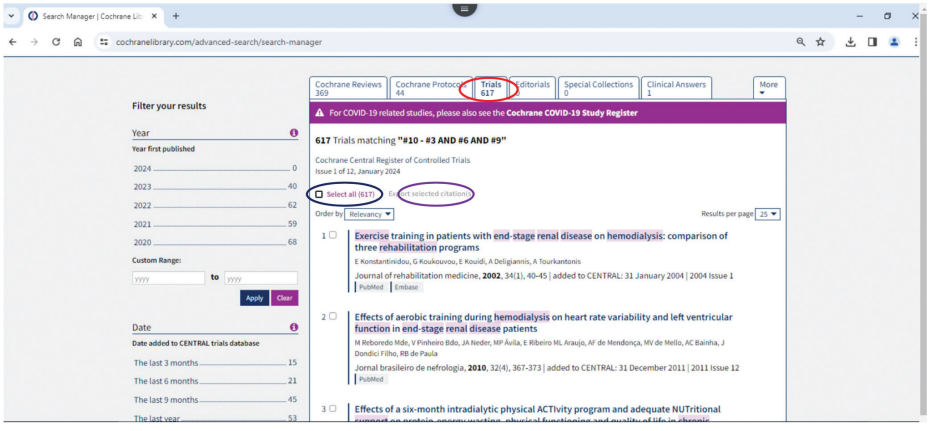


Figure 4.51 Running the searches in Cochrane: exporting search results interface



Step 4.2.3 Web of Science

Step 4.2.3.1

Figure 4.53 shows the basic search page when you log into Web of Science. Click 'Advanced Search' (red ellipse).

Step 4.2.3.2

Click 'Editions' (red ellipse in Figure 4.54), uncheck 'Select All' (blue ellipse), and then select 'Conference Proceedings Citation Index – Science' (purple ellipse). This results in only the 'Conference Proceedings Citation Index – Science' being searched.

Figure 4.52 Running the searches in Cochrane: exporting search results final step

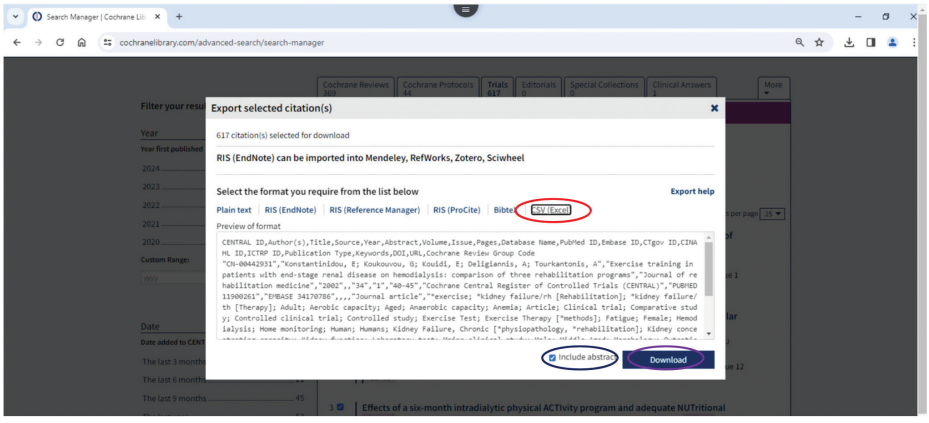
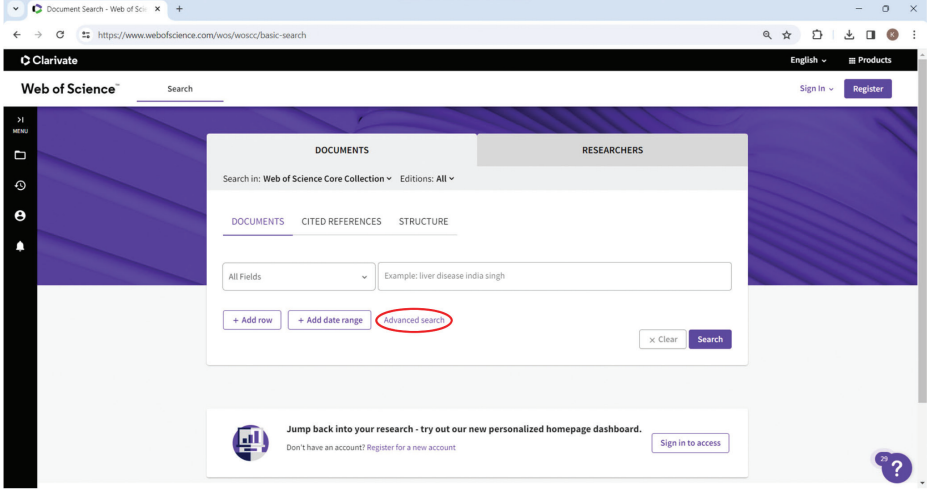


Figure 4.53 Running the searches in Web of Science: landing page



Step 4.2.3.3

Only the free text terms are entered in the 'Conference Proceedings Citation Index'. Enter the free text terms in the search box labelled 'Add terms to the query preview' (red ellipse in Figure 4.55), change the 'All Fields' (blue ellipse) to 'Topic' and click 'Add to query' (purple ellipse).

Step 4.2.3.4

This results in the terms being entered into the search box, and the program has added 'TS=' (red ellipse in Figure 4.56) in front of the search terms. You can add 'TS=' manually and skip the step of entering the search into the 'Add terms to the

Figure 4.54 Running the searches in Web of Science: selecting the databases to be searched

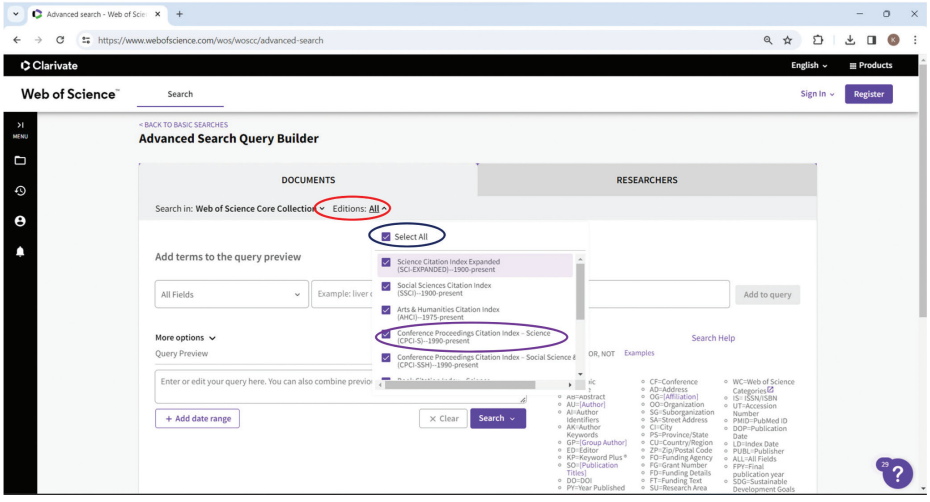
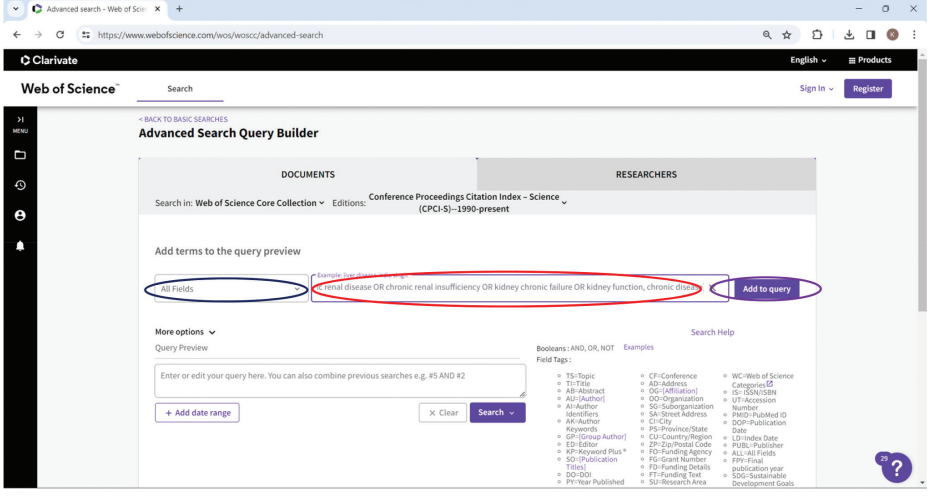


Figure 4.55 Running the searches in Web of Science: adding search terms to query preview



query preview' box if you prefer to do so. However, if you enter the search terms without the 'TS=' in front, it can result in error. Therefore, we recommend you enter the search terms to the query preview search box and let the program add the correct tag in front. Click 'Search' (blue ellipse).

Figure 4.56 Running the searches in Web of Science: results of query preview

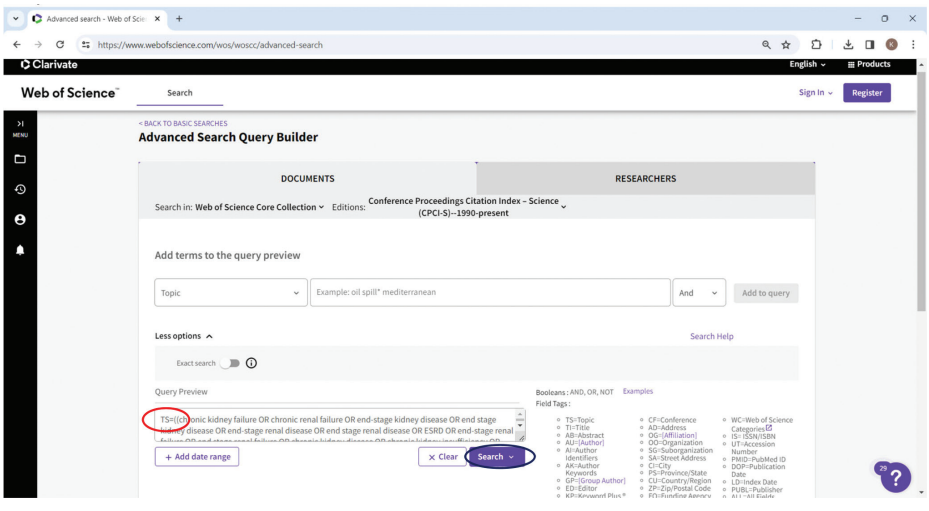
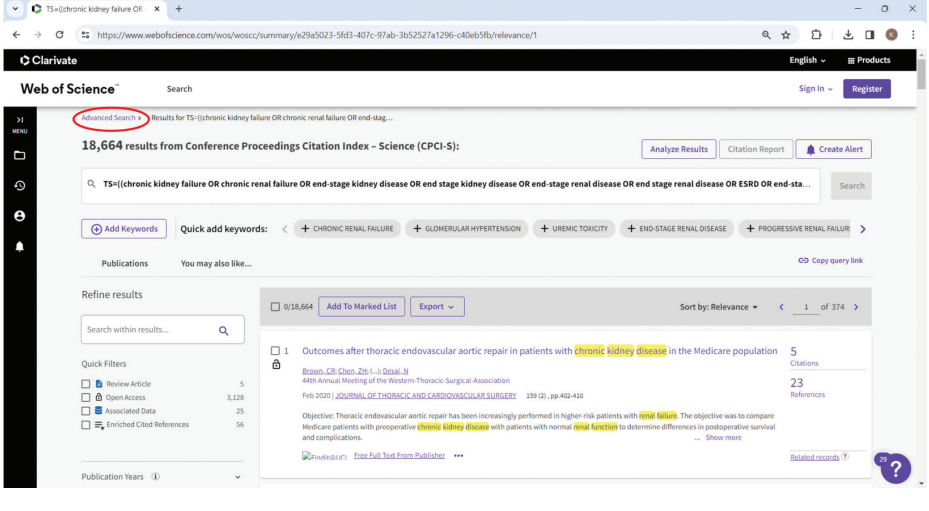


Figure 4.57 Running the searches in Web of Science: search results (single concept or sub-concept)



Step 4.2.3.5

This gives you the search results (see Figure 4.57). However, this is only for one sub-concept. You can add other concepts or sub-concepts by clicking 'Advanced Search' (red ellipse).

Step 4.2.3.6

In the resulting window (Figure 4.58), click 'Clear' (red ellipse) and repeat Steps 4.2.3.3, 4.2.3.4 and 4.2.3.5 for the other concepts.

Step 4.2.3.7

After entering all the concepts and sub-concepts, scroll down the 'Advanced Search' page (Figure 4.59). This shows all the concepts and sub-concepts that

Figure 4.58 Running the searches in Web of Science: clearing query preview

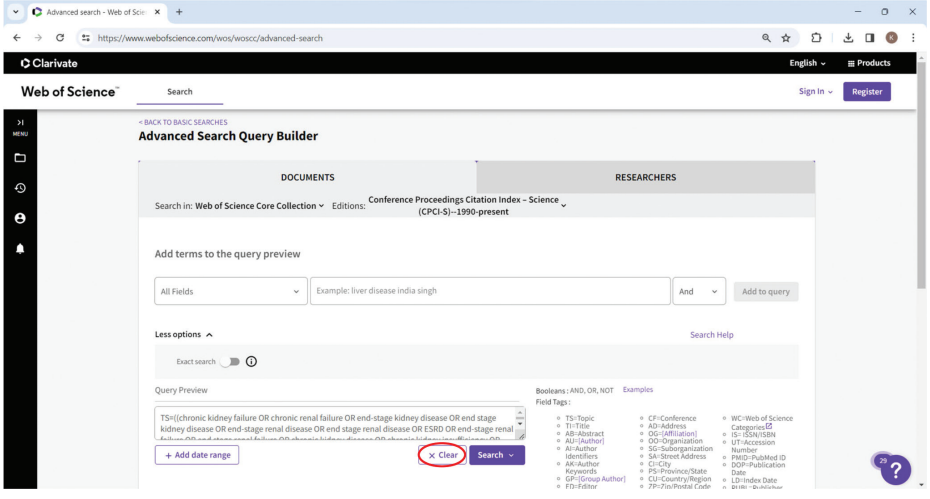
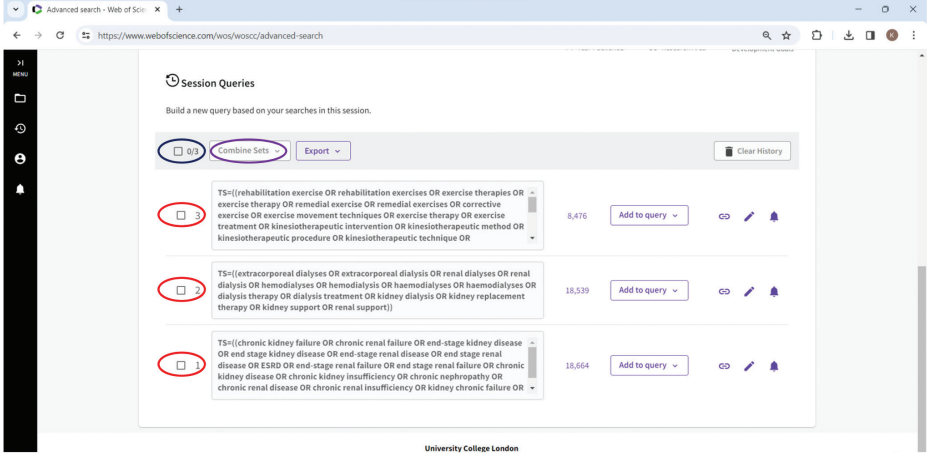


Figure 4.59 Running the searches in Web of Science: search results (multiple concepts or sub-concepts)



have been entered. Now, select the searches that you want to combine (red ellipses). You can also select all the searches in a single click by clicking above the searches (blue ellipse). Then click 'Combine Sets' (purple ellipse).

Step 4.2.3.7

Choose the appropriate Boolean operator, which is usually 'And' (red ellipse in Figure 4.60) for combining different concepts or sub-concepts.

Step 4.2.3.8

You can view the results by clicking on the number in the row for the search (red ellipse in Figure 4.61). You can export the search strategy by clicking 'Export' (blue ellipse). If the references are unmanageable, you can consider adding terms such as 'random* OR placebo* OR trial*' to identify RCTs.

Step 4.2.3.9

To export the results, click 'Export' (red ellipse in Figure 4.62) and then 'Tab delimited file' (blue ellipse).

Step 4.2.3.10

In the resulting screen (Figure 4.63), choose the records to export (red ellipse) and 'Full record' (blue ellipse) as the 'Record Content'. Then click 'Export' (hidden by the dropdown menu for 'Record Content').

Figure 4.60 Running the searches in Web of Science: combining multiple concepts or sub-concepts

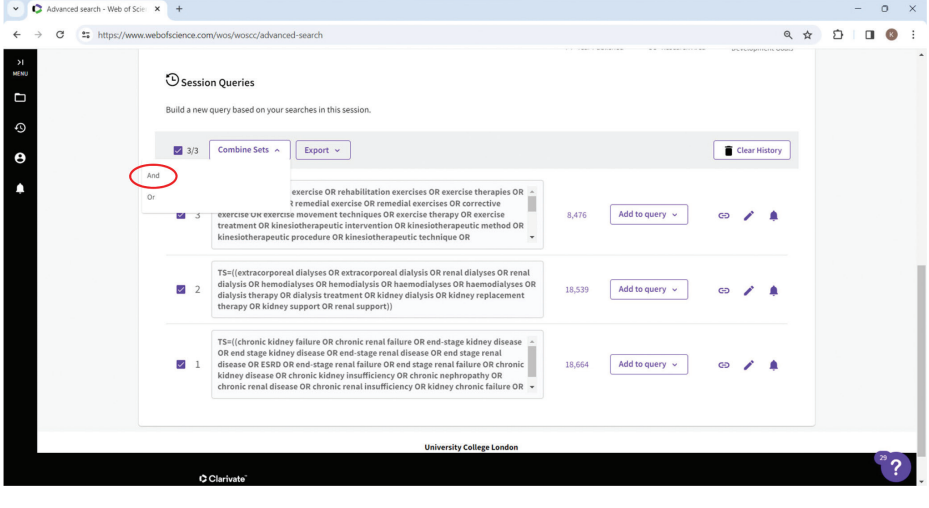


Figure 4.61 Running the searches in Web of Science: final search results

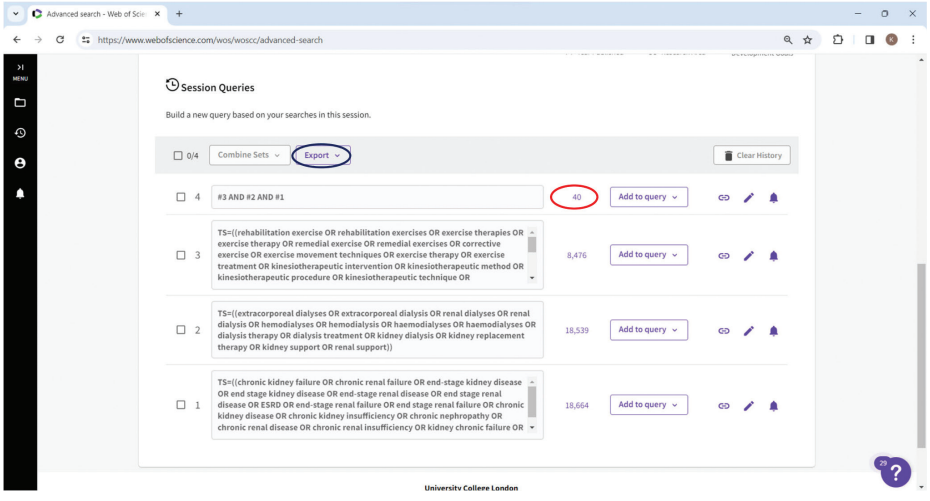
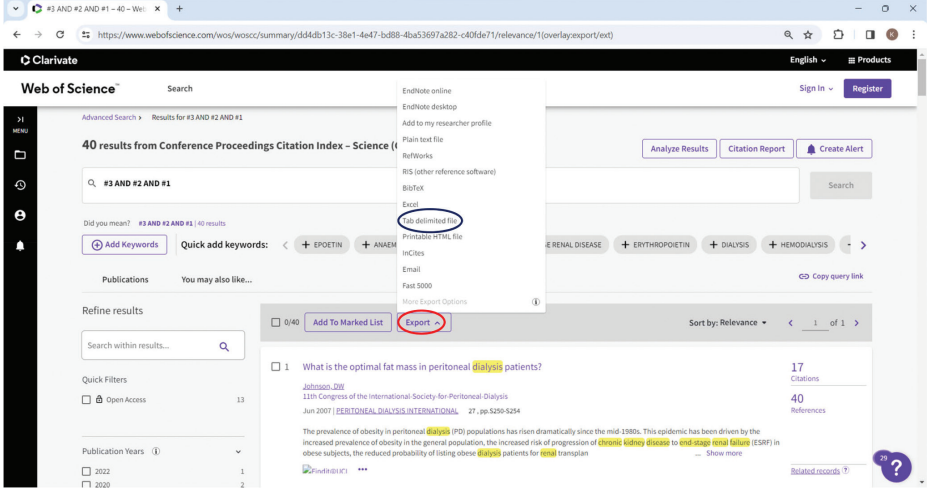


Figure 4.62 Running the searches in Web of Science: exporting search results interface



Step 4.2.4 ClinicalTrials.gov

Step 4.2.4.1

In ClinicalTrials.gov initial page, enter the terms in the relevant search boxes (Figure 4.64). Note that I have entered the terms related to dialysis in the 'Other terms' (red ellipse). Click 'Search' (blue ellipse). This takes you to the search results.

Figure 4.63 Running the searches in Web of Science: exporting search results final step

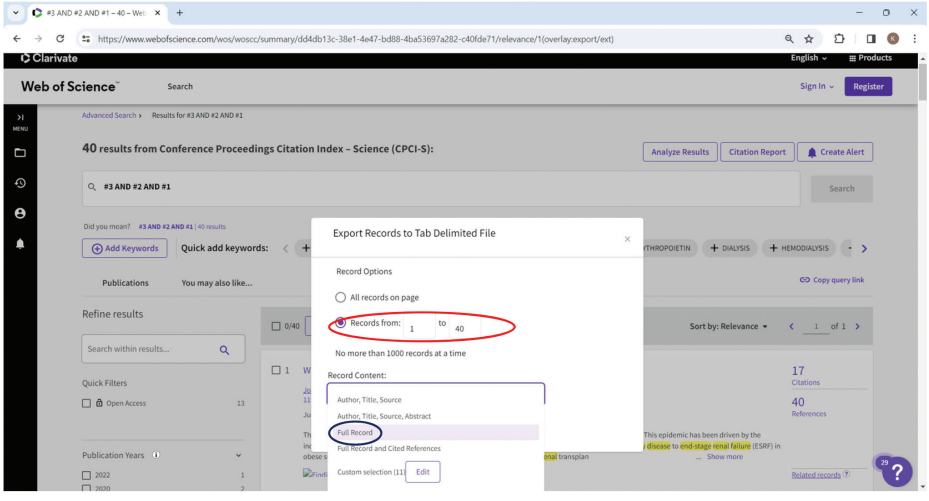
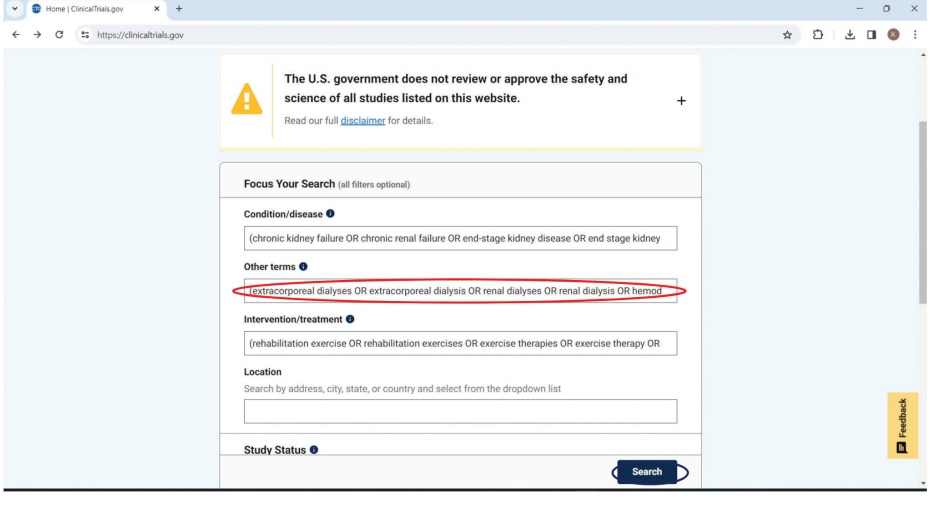


Figure 4.64 Running the searches in ClinicalTrials.gov: landing page



Step 4.2.4.2

To download the results, click the 'Download' symbol (red ellipse in Figure 4.65).

Step 4.2.4.3

In the resulting window (Figure 4.66), choose 'CSV' (red ellipse) as the 'File Format' and 'All studies' (blue ellipse) in 'Results to Download'. Ensure that all 'Data Fields' are selected by clicking 'Select all' (purple ellipse), scroll down, and click 'Download' (not shown).

Figure 4.65 Running the searches in ClinicalTrials.gov: search results

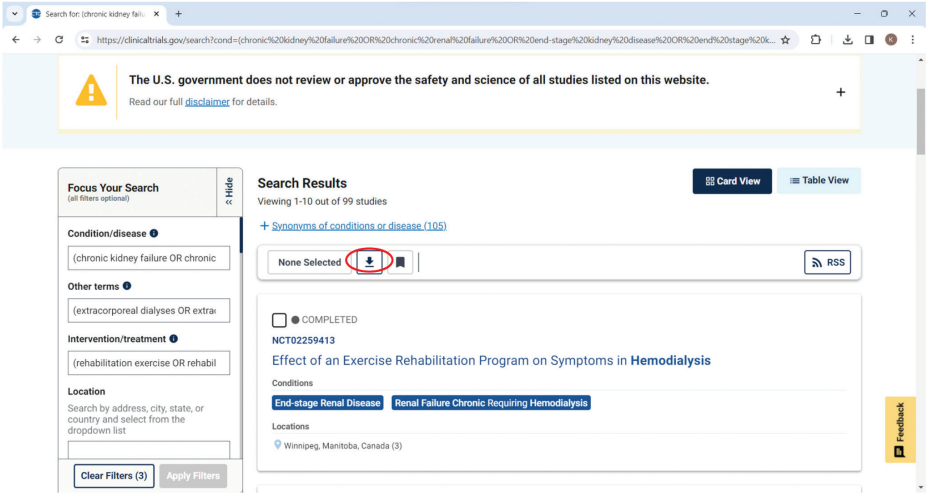
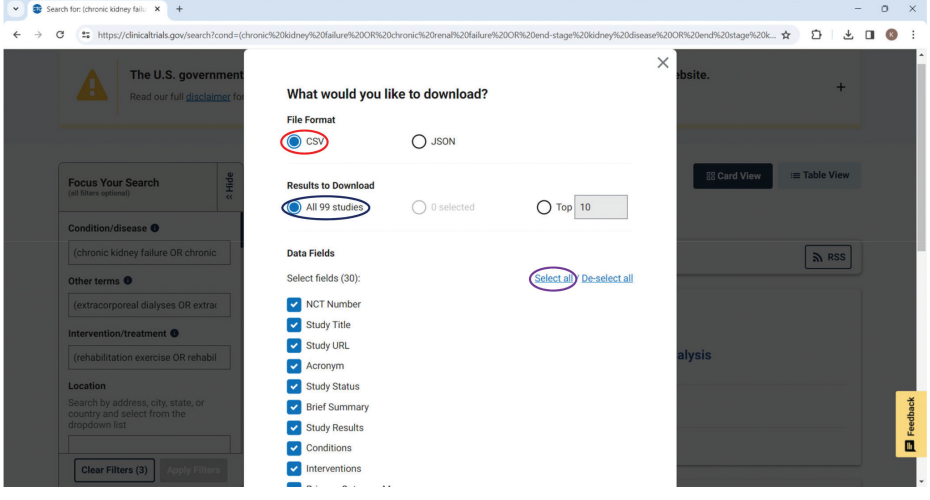


Figure 4.66 Running the searches in ClinicalTrials.gov: exporting search results interface



Step 4.2.5 WHO ICTRP

Step 4.2.5.1

The initial page in WHO ICTRP is shown in Figure 4.67. Click 'Advanced Search' (red ellipse).

Figure 4.67 Running the searches in WHO ICTRP: landing page

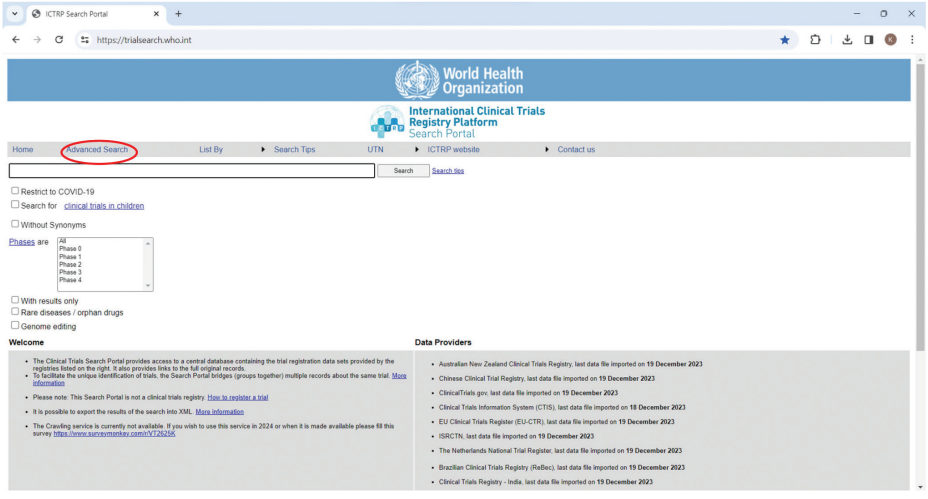
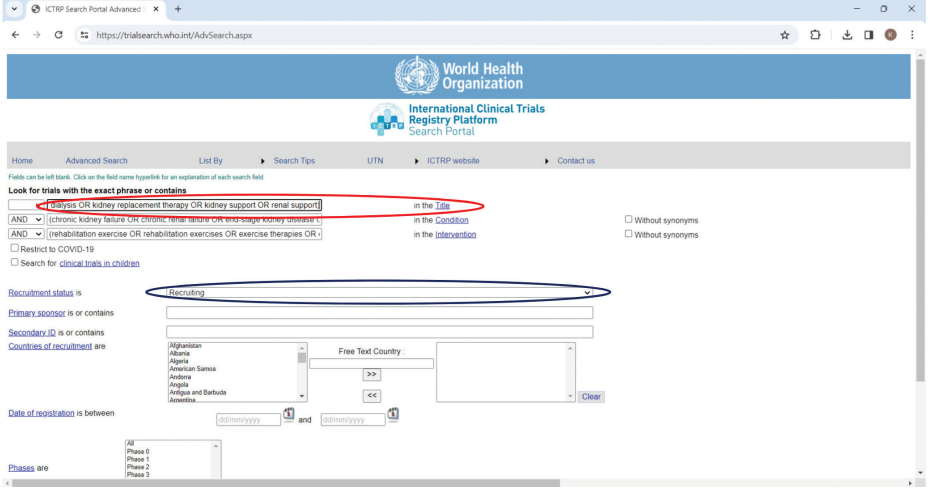


Figure 4.68 Running the searches in WHO ICTRP: advanced search interface



Step 4.2.5.2

The search terms are entered in the relevant box (see Figure 4.68). Note that the search terms related to dialysis have been entered in the 'Title' field (red ellipse). Before you search, it is important to change the 'Recruitment status' from 'Recruiting' (blue ellipse) to 'ALL'. This takes you to the search results directly.

Step 4.2.5.3

To download the results, click 'Export results to XML' (Figure 4.69).

Figure 4.69 Running the searches in WHO ICTRP: search results

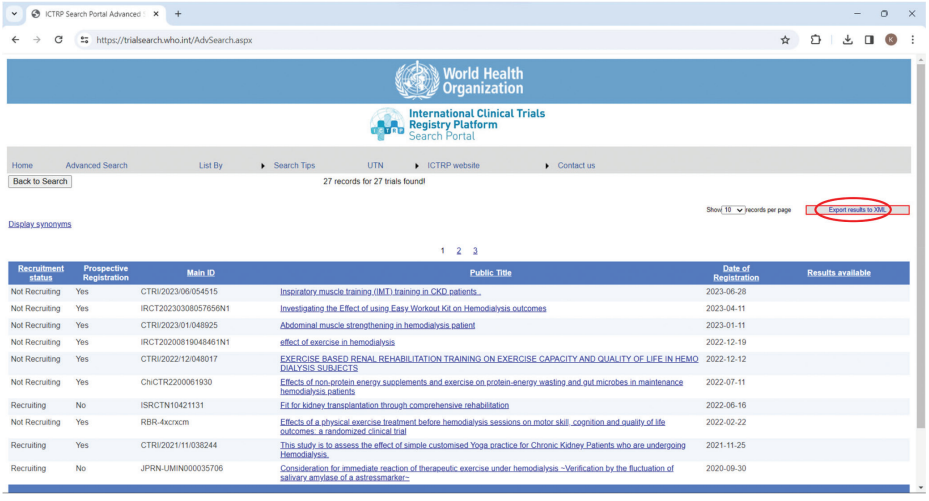
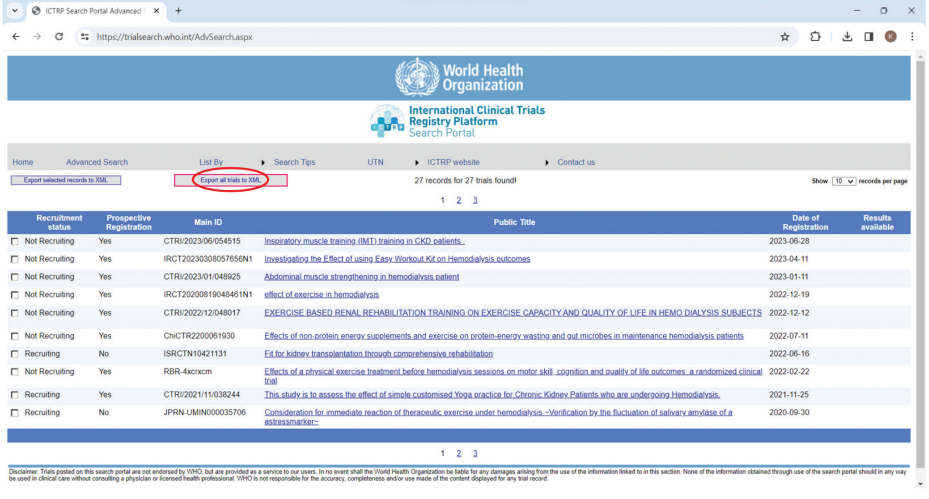


Figure 4.70 Running the searches in WHO ICTRP: exporting search results



Once you agree to the terms and conditions, the 'Export all trials to XML' button (red ellipse in Figure 4.70) appears. Click this to export the results.

Recording the searches in and from different databases

Most databases allow saving the search strategy: therefore, there is no need to type the searches during each search update. Saving the search strategy in the databases is optional. Record the date of search, the search terms used and the

number of references retrieved in each database. It is important to note this information since it is required in the report of the systematic review.

Step 5 Combine the results

The next step in the process of study selection is to combine the results from multiple databases. The downloads from different databases have different field headings. The details of what each field heading denotes are available from the downloads or from the help section of the databases. These can also be combined using the online software EQUAL-SR developed for helping with systematic review by the author of this book and available from https://sites.google.com/view/equal-group/home. This software accepts the files exported in the format described above, combines them and removes the duplicates (performs 'de-duplication') based on unique identifiers in databases. For further instructions of how to combine the search results and de-duplicate the references using the EQUAL-SR, please see the software website. Another way is to combine them using reference management software, which requires export of references specific to the reference management software used. EndNote, a reference management software package offers de-duplication.

Although electronic de-duplication of references removes many duplicates, you may still find some residual duplicates.

Step 6 Screen the titles and abstracts

The next step in the process of study selection is to screen the titles and abstracts to identify potentially eligible studies. It is recommended that two independent reviewers perform this to minimise the bias in the systematic reviews. [25] 'Independent reviewers' means that the selection is performed independently and not collectively. Any differences in study selection are usually resolved by discussion between the reviewers. This will remove any differences due to errors by one of the reviewers or any potential misunderstanding by the reviewers. You can use a third (senior) reviewer to arbitrate on any residual differences in study selection; however, it is usually more efficient to obtain the full text of the reference for which there is a difference between the reviewers in terms of eligibility. The independent reviewers may then make their judgement on selecting the study based on the full text of the reference.

The role of artificial intelligence in replacing humans to screen titles and abstracts is still being evaluated. [26] Some software has the potential to decrease the workload, [26] but the performance of such software at predefined thresholds for rejection in a wide range of systematic reviews requires further evaluation.

EQUAL-SR offers highlighting of keywords and ranking of references based on the presence or absence of the concepts (or sub-concepts) linked to the

keywords. You can also provide relative weights (including negative weights) to the different concepts to allow further refinement of ranking. For further instructions on how to perform this, please see the software website.

Step 7 Obtain full texts

The next step is to obtain full texts for the potentially eligible studies. While some full texts are available free, others need subscriptions. Usually, institutions have subscriptions to many journals. Some software, such as EndNote, can automatically retrieve the full text for some articles once you provide details of your institutional subscription and link, which should be available from the institutional library website. However, a considerable proportion of articles are not retrieved automatically and must be retrieved manually. Logging in through an institutional computer or using the institutional virtual private network (VPN) can make retrieving full texts more efficient, as many journals with institutional subscription provide full text access based on the IP (Internet Protocol) address.

If the full texts are not available either as full text or from the institutional library, you can request a copy from the author. If you are unable to obtain the full text by any of these means, you can request it through your library who can arrange inter-library loans through 'WorldCat' (https://search.worldcat.org/) or through other organisations that arrange inter-library loans. [8]

Step 8 Finalise study selection

It is recommended that at least two independent reviewers perform the final study selection to decrease the bias in the study selection process. [25] As for the title and abstract screening step, any differences in study selection are usually resolved by discussion between the independent reviewers, which will remove any differences due to errors by one of the reviewers or any potential misunderstanding. You can seek clarification from authors about any unresolved differences resulting from unclear information. If there are any residual differences, these can be resolved through arbitration by a third (senior) reviewer. If differences persist despite arbitration, a sensitivity analysis can be performed to assess the impact of including the study on the systematic review results.

Overview of resolving differences in study selection

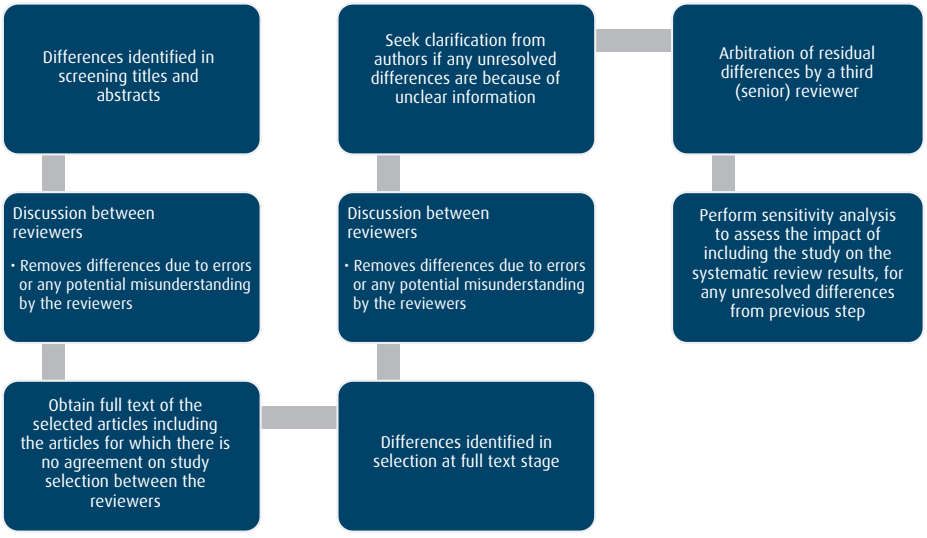
A suggested process for resolving differences is provided in Figure 4.71.

COMMON ERRORS AND DOUBTS IN STUDY SELECTION

There now follows a list of common errors that systematic reviewers commit and doubts they have during the process of study selection.

Figure 4.71 Suggested process for resolving differences in study selection

This figure shows an overview of the process of resolving differences.



Multiple reports of the same study

One of the common errors that systematic reviewers make is to exclude references based on these references being additional reports of included studies. This exclusion should be avoided: since we are interested in all the information from the study, it is important that all the available reports of the study are identified and collated. This includes any conference abstracts, preliminary reports and trial registry entries. During the stage of data extraction, maximum information should be sought from all the reports.

Multiple studies in the same report

Although much less common than multiple reports of the same study, multiple studies in the same report should be treated as different studies. If one or more of the studies meet the inclusion criteria for the systematic review, they should be included.

Studies in which some populations are eligible for inclusion in the systematic review

In general, when the primary research studies include populations which are not eligible for inclusion in the systematic review, such studies are excluded unless data on the subset of populations eligible for inclusion in the systematic review are available in the reports or can be obtained by writing to the primary research study author.

However, you should be aware that even in a randomised trial, the primary research study authors may not have used additional methods, such as stratification or minimisation, to ensure that the allocation of intervention or comparator is random in this subset of populations. Therefore, it is important to assess the contribution of the study to the systematic review results by performing a 'sensitivity analysis', that is, by excluding this study.

When data are sparse, different approaches can be followed, for example including the study when a certain proportion of populations (say 80% or 90%) are eligible for inclusion in the systematic review. However, it is possible that the effect of the intervention versus comparator is largely due to the effect of the intervention versus comparator on the subset of populations who are not eligible for inclusion in the systematic review. Therefore, an element of indirectness is introduced. This potential for obtaining a misleading effect of the intervention versus comparator (because of including populations who are not eligible for the systematic review) should be discussed as a limitation of the evidence and considered when assessing the uncertainty in the evidence.

An extreme situation is when there is no evidence available on the populations eligible for inclusion in the systematic review from RCTs or from non-randomised studies at low or moderate risk of bias. In such a situation, you may have to rely on evidence from populations not eligible for inclusion in the systematic review. However, this may result in a seriously misleading effect of the intervention versus comparator. The differences between populations eligible for inclusion in the systematic review and those not eligible for inclusion in the systematic review should be discussed in detail: the discussion should include information on how these differences may have an impact on the potential mechanisms of action of the intervention (or comparator). Lack of direct evidence on the populations eligible for inclusion in the systematic review should also be considered when assessing the uncertainty in the evidence.

Studies in which some interventions are eligible for inclusion in the systematic review

When primary research studies include some interventions that are eligible for inclusion in the systematic review and not others, such studies can be included. In Example 1, the research question was: *'Evaluate the effectiveness of school-based dental screening versus no screening on improving oral health in children aged 3–18 years by a systematic review and meta-analysis of randomised controlled trials'*. If you come across a study in which the primary study authors compared three groups, 'school-based dental screening' versus 'home-based dental screening' versus 'no screening', it is appropriate to include only the two groups from the study that are relevant for the systematic review, namely, 'school-based dental screening' and 'no screening'.

Studies in which some additional interventions are applied along with the eligible intervention and comparator of the systematic review

When additional interventions are applied along with the eligible intervention and comparator of the systematic review, they are called co-interventions. In general, you would include studies when the co-interventions were applied to both the intervention and comparator groups equally. In Example 2, the research question was: *'To assess the benefits and harms of laparoscopic distal pancreatectomy versus open distal pancreatectomy for people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma of the body or tail of the pancreas, or both'*. In this example, in some studies, the study populations in both groups (laparoscopic distal pancreatectomy and open distal pancreatectomy groups) may have received preoperative exercise and diet counselling routinely in addition to the surgery, while in other studies, the study populations may not have received preoperative exercise and diet counselling in either group. It is appropriate to include such studies where the co-interventions are planned equally in both groups. However, it is usually inappropriate to include studies where co-interventions are planned only in one group, since the differences in the effect between the groups may be due to the co-intervention and not because of the intervention.

However, if a co-intervention is part of the intervention – for example, a blood test to monitor complication of an intervention, which is not performed in the comparator group but is performed in all the studies eligible for the systematic review – such studies can be included, as such blood tests can be considered an integral part of the intervention, and the effect of the intervention versus comparator obtained in the systematic review is actually the effect of the intervention along with the blood test to monitor its complications (versus comparator). However, if this blood test is performed in only some studies, this indicates that the blood test is not an integral part of the intervention, and you should consider the impact of the additional visit for the blood test on the effect of the intervention versus comparator. The decision between using differential co-intervention as a criterion for rejection or considering it as a source of bias (but including the study) depends on the amount of evidence available with and without the co-intervention.

SUMMARY

In this chapter, we have provided some guidance on study selection for a systematic review.

PRACTICE QUESTION

1. Design a search strategy to identify RCTs of the role of chest physiotherapy for pneumonia.

A weblink for you to check your answer to the practice question can be found in the Appendix.

REFERENCES

1. Office of Public Affairs (US Department of Justice). GlaxoSmithKline to Plead Guilty and Pay $3 Billion to Resolve Fraud Allegations and Failure to Report Safety Data. www.justice.gov/opa/pr/glaxosmithkline-plead-guilty-and-pay-3-billion-resolve-fraud-allegations-and-failure-report (accessed on 27 December 2023). 2012.
2. Medicines and Healthcare Products Regulatory Agency (MHRA). MHRA Investigation into Glaxosmithkline/Seroxat. https://webarchive.nationalarchives.gov.uk/ukgwa/20141206221413/http:/www.mhra.gov.uk/home/groups/es-policy/documents/websiteresources/con014155.pdf (accessed on 27 December 2023). 2008.
3. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, 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(7): e1000100.
4. National Library of Medicine. PubMed Overview. https://pubmed.ncbi.nlm.nih.gov/about/ (accessed on 27 December 2023). 2023.
5. Marshall IJ, Marshall R, Wallace BC, Brassey J, Thomas J. Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. J Clin Epidemiol. 2019; 109:30–41.
6. Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, et al. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol. 2003; 56(10):943–55.
7. Cochrane Library. Cochrane Central Register of Controlled Trials (CENTRAL). www.cochranelibrary.com/central/about-central (accessed on 27 December 2023). 2023.
8. Xia J, Wright J, Adams CE. Five large Chinese biomedical bibliographic databases: accessibility and coverage. Health Information & Libraries Journal. 2008; 25(1):55–61.

9. Virtual Health Library. LILACS: Latin American and Caribbean Health Sciences Literature. https://lilacs.bvsalud.org/en/ (accessed on 27 December 2023). 2023.
10. International Committee of Medical Journal Editors. Clinical Trials. www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html (accessed on 29 December 2023). 2023.
11. Joury E, Bernabe E, Sabbah W, Nakhleh K, Gurusamy K. Systematic review and meta-analysis of randomised controlled trials on the effectiveness of school-based dental screening versus no screening on improving oral health in children. Dent. 2017; 58:1–10.
12. Riviere D, Gurusamy KS, Kooby DA, Vollmer CM, Besselink MG, Davidson BR, et al. Laparoscopic versus open distal pancreatectomy for pancreatic cancer. Cochrane Database Syst Rev. 2016; 4(4): Cd011391.
13. Svenonius E. Unanswered questions in the design of controlled vocabularies. Am Soc Inf Sci. 1986; 37(5):331–40.
14. Savoy J. Bibliographic database access using free-text and controlled vocabulary: an evaluation. Inf Process Manag. 2005; 41(4):873–90.
15. Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Marshall C, et al. Technical supplement to Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, Welch VA (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 6.3 (updated February 2022). Cochrane, 2022. Available from: www.training.cochrane.org/handbook (accessed on 28 December 2023).
16. SNOWMED International. SNOMED International SNOMED CT Browser. https://browser.ihtsdotools.org/? (acccessed on 10 September 2023). 2023.
17. World Health Organization. International Statistical Classification of Diseases and Related Health Problems 11th Revision. www.who.int/standards/classifications/classification-of-diseases (accessed on 31 March 2024). 2022.
18. National Library of Medicine. Unified Medical Language System (UMLS). www.nlm.nih.gov/research/umls/index.html?_gl=1*1krik6e*_ga*MTk4MjYzMTE5MS4xNjk0MzI2NTMy*_ga_P1FPTH9PL4*MTY5NDMyNjUzMS4xLjEuMTY5NDMyNjU2OC4wLjAuMA..*_ga_7147EPK006*MTY5NDMyNjUzMS4xLjEuMTY5NDMyNjU2OC4wLjAuMA (accessed on 10 September 2023). 2023.
19. National Library of Medicine. PubMed User Guide. https://pubmed.ncbi.nlm.nih.gov/help/ (accessed on 29 December 2023). 2023.
20. The InterTASC Information Specialists' Sub-Group. The ISSG Search Filter Resource. https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home (accessed on 31 March 2024). 2024.

21. Kamdar BB, Shah PA, Sakamuri S, Kamdar BS, Oh J. A novel search builder to expedite search strategies for systematic reviews. Int J Technol Assess Health Care. 2015; 31(1–2):51–3.
22. Wanner A, Baumann N. Design and implementation of a tool for conversion of search strategies between PubMed and Ovid MEDLINE. Res Synth Methods. 2019; 10(2):154–60.
23. Bramer WM, de Jonge GB. Improving efficiency and confidence in systematic review literature searching. https://clinicallibrarian.files.wordpress.com/2015/06/improving-efficiency-confidence-in-systematic-literature-searching-wb-gdej.pdf (accessed on 29 December 2023). 2015.
24. Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Metzendorf M, et al. Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, Welch VA (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 6.4 (updated October 2023). Cochrane, 2023. Available from: www.training.cochrane.org/handbook. (accessed on 28 December 2023).
25. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016; 69:225–34.
26. Yao X, Kumar MV, Su E, Flores Miranda A, Saha A, Sussman J. Evaluating the efficacy of artificial intelligence tools for the automation of systematic reviews in cancer research: a systematic review. Cancer Epidemiol. 2024; 88:102511.

5 Extract the data

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- *Extract the data necessary for analysis and interpretation of whether an intervention is better than its comparator*

OVERVIEW

In systematic reviews of intervention, the data extracted include data related to the source of information, PICO details, participant flow, follow-up, risk of bias, and conflicts of interests in the study.

Sufficient details about these items should be extracted to allow exploration of whether the differences in the study results could be explained by differences in one or more of the PICO details, follow-up, risk of bias, or conflicts of interests. In addition, sufficient details should be extracted to help with shared decision making. This includes extracting sufficient participant details to assess whether the results can be applied for an individual person in the specific clinical setting, and sufficient intervention and comparator details to establish whether these are choices that could be offered to the patient.

From an outcome data extraction perspective, the outcomes in systematic reviews of interventions are classified into binary outcomes, continuous outcomes, ordinal outcomes, count outcomes and time-to-event outcomes. The data extracted for these different types of outcomes vary.

In systematic reviews of intervention, the risk of bias in RCTs is assessed by the Risk of Bias in randomised trials (RoB 2) and that in non-randomised studies is assessed by the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) tool. Sufficient details should be extracted to allow these assessments.

A data extraction form should be used, to allow the systematic collection of data from the studies. Data extraction should be performed by at least two reviewers independently.

WHAT IS DATA EXTRACTION?

You may recollect from Chapter 2 that once the research question is determined, the next steps are to identify suitable methods for study selection, data extraction, analysis and reporting, and to complete protocol registration. In the previous chapter, we described the suitable methods for study selection and also how study selection works in practice.

The next step is to identify the suitable methods for data extraction and learn how data extraction works in practice. But what is data extraction?

Merriam-Webster Dictionary (online) defines data as factual information used as a basis for reasoning, discussion or calculation. [1] Data extraction, in the context of systematic reviews, simply refers to collecting data from the published or unpublished records of a study in a format that is suitable for subsequent analysis and interpretation.

WHY IS IT NECESSARY TO USE SUITABLE METHODS FOR DATA EXTRACTION?

The analysis and interpretation are based on the data extracted. Therefore, it is necessary to use suitable methods for data extraction. Incorrect data extraction due to the use of inappropriate methods for data extraction, insufficient data extraction or incorrect data extraction can lead to incorrect analysis, interpretation and conclusion.

OVERVIEW OF THE STEPS IN DATA EXTRACTION

An overview of the steps in data extraction is shown in Figure 5.1. Each of the steps in data extraction is discussed in detail next.

TYPES OF DATA FOR EXTRACTION

Before we go on to decide what data should be extracted, it is useful to understand some basic numerical concepts that can help with understanding what data should be extracted.

Classification

Data can be classified in many ways. A classification that we will use in the context of how data should be extracted and analysed for systematic reviews is shown in Figure 5.2.

Figure 5.1 Overview of steps in data extraction

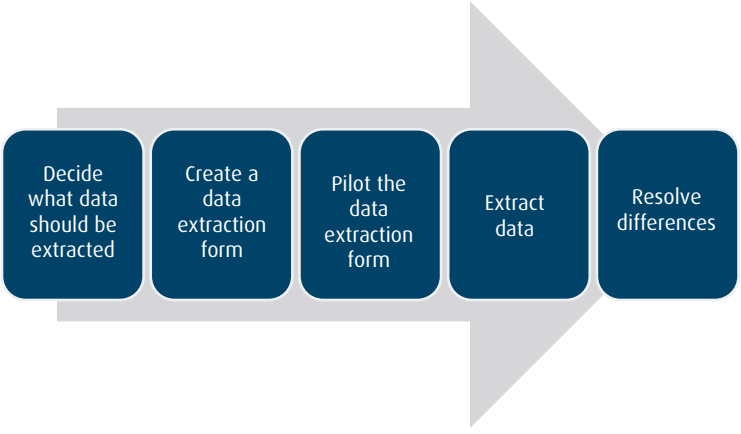
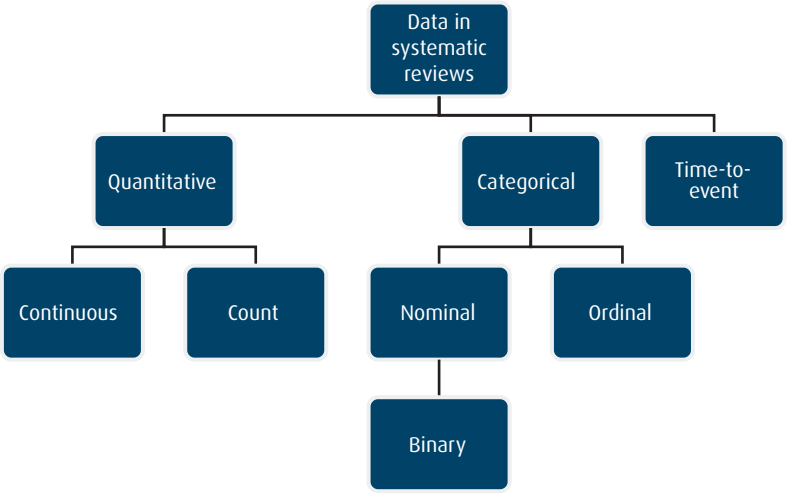


Figure 5.2 Classification of data in systematic reviews



In the context of systematic reviews, data can be broadly classified as quantitative and categorical data. [2]

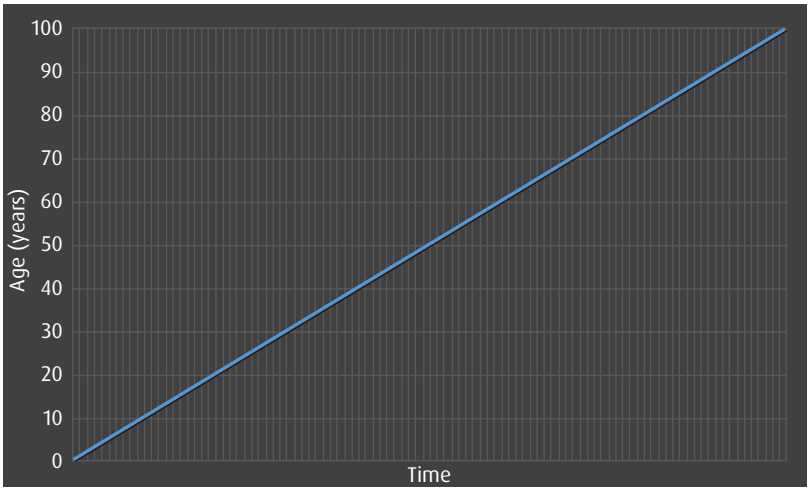
Quantitative data

Quantitative data are data that answer the question 'How much?' They can be further classified into 'continuous' data and 'counted' data (or simply 'count' data). [2]

Continuous data

Continuous data take any value within a given range. Examples include height, weight or age. The age of an individual over time is shown in Figure 5.3.

Figure 5.3 Age over time

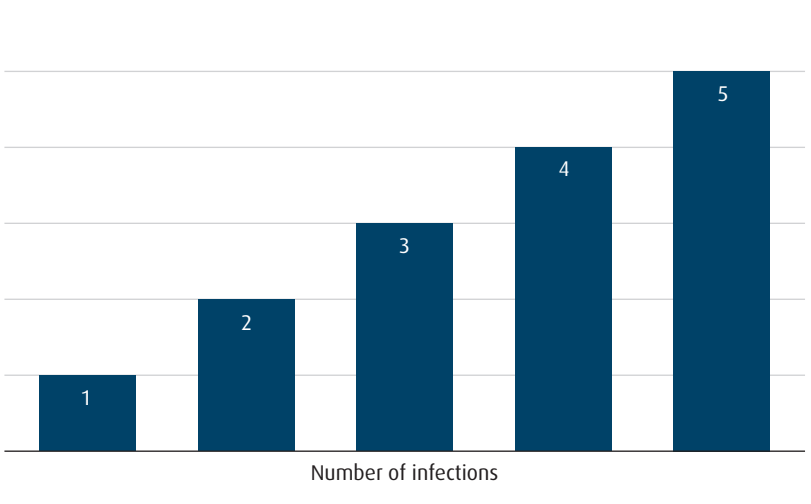


Continuous data can take negative values depending upon the measures used. For example, the temperature of water in degrees Celsius can be negative.

Count data

In the context of systematic reviews related to health and disease, count data are counts within a specified time or area. Examples include number of children, number of infections, number of heart attacks. The cumulative number of infections over time in an individual is shown in Figure 5.4.

Figure 5.4 Cumulative number of infections over time



Count data can never take negative values, for example, you cannot have '–1' infections. Therefore, the lower limit for the count data is 0.

Categorical data

Categorical data are data that answer the question 'What type?' They can be further classified into nominal and ordinal variables. [2]

Nominal data

The nominal data are categorical data that have no specific order. Examples include blood group, colour of people's eyes, colour of people's skin and gender.

The blood group of individuals is shown in Figure 5.5. There is no specific order for blood groups in terms of intelligence, although there may be an order in terms of frequency (how often they occur) in the population.

Binary data

Binary data are a special type of nominal categorical data with only two possibilities. Examples include alive or dead, presence or absence of disease. Whether the person was cured after treatment is shown in Figure 5.6. Another name for binary data is 'dichotomous data'.

Ordinal data

Ordinal data are categorical data that are ordered. Examples include grade of cancer, intensity of pain. The different intensities of pain in individuals are shown in Figure 5.7.

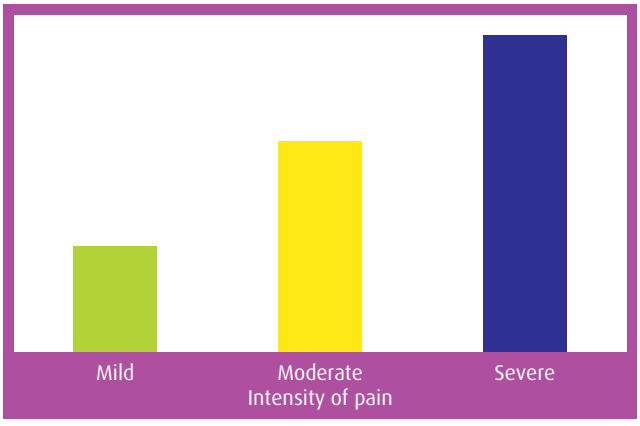
Nominal versus ordinal

It is easy to remember the definition for nominal and ordinal. Nominal means 'something in name or form only' [3] and ordinal means 'of a specified order or rank in a series'. [4]

Figure 5.5 Blood group of individuals

Figure 5.6 Was the disease cured after treatment?

Figure 5.7 Intensity of pain



Count data versus ordinal data

Sometimes, people get confused between count data and ordinal data. The difference between the two can be best explained by an example.

Let us say that a researcher wants to compare the reduction in the number of infection episodes by two treatments aimed at preventing throat infections – 'Treatment A' and 'Treatment B' – in people who are prone to develop throat infections. Prior to the start of the treatments, the number of infection episodes in the person who would later receive 'Treatment A' was 8 infection episodes in the last year; it was 6 infection episodes in the last year in the person who would later receive 'Treatment B'. After starting the treatment, the person who received 'Treatment A' developed 6 infection episodes in a year and the one who received 'Treatment B' developed 4 infection episodes in a year. Here pre-treatment versus post-treatment differences in the infection episodes are a reduction in 2 infection episodes in both scenarios, that is, the interval between 8 and 6 is the same as the interval between 6 and 4 in count data.

Let us say another researcher wants to compare two physiotherapy treatments, 'Treatment P' and 'Treatment Q'. One of the commonly used systems for grading muscle strength is the Medical Research Council Muscle Strength Grading. [5] This is shown in Table 5.1.

Prior to the start of the treatments, the muscle power in the person who would later receive 'Treatment P' was '0'; it was '2' in the person who would later receive 'Treatment Q'. After starting the treatment, the power improved to '1' in the person who received 'Treatment P' and to '3' in the person who received 'Treatment Q'. Here, even though the pre-treatment versus post-treatment differences in power are an improvement of power by 1, the amount of improvement in terms of

Table 5.1 Medical Research Council Muscle Strength Grading

0 = No muscle activation
1 = Trace muscle activation, such as a twitch, without achieving full range of motion
2 = Muscle activation with gravity eliminated, achieving full range of motion
3 = Muscle activation against gravity, full range of motion
4 = Muscle activation against some resistance, full range of motion
5 = Muscle activation against examiner's full resistance, full range of motion

activities that a person can do is clearly different in an improvement from '0' to '1' (where the difference is only additional twitching of muscles) compared to the improvement from '2' to '3' (where a person may be able to lift an arm).

Therefore, to differentiate count data and ordinal data, you must check whether a certain value for the difference (whether the difference indicates pre-treatment and post-treatment difference or a difference between two or more groups) means the same regardless of the absolute values (either the baseline or final value). If the difference means the same regardless of the absolute value, it is count data; if the difference means different things depending upon the baseline or final value, it is ordinal data.

It is important to differentiate count and ordinal data as the methods used for data analysis are different for count and ordinal data.

Time-to-event data

Most classification systems will categorise the data only as quantitative and categorical data based on the nature of the data. They do not specify 'time-to-event' data as a specific category. This is because 'time-to-event' data is a special type of data applicable only to outcomes and not to other types of variables such as prognostic factors. It is a combination of a categorical data, usually binary data, to indicate whether an individual has an outcome, and continuous data to indicate the time point at which the individual developed the outcome (if they developed the outcome), or the last follow-up at which the individual was reviewed and was found not to have the outcome (if they did not develop the outcome).

DATA DISTRIBUTION

After understanding the different types of data in the context of systematic reviews, it is important to understand what information must be extracted for

each type of data. However, the data to be extracted also depend upon data distribution. Therefore, we need to learn about data distribution too.

What is data distribution?

In simple terms, data distribution is a graphical representation that provides information on how often the data values (or intervals of data) occur. They can also be represented by a formula (or 'function').

The frequency of heads (number of times heads occurred) in each of 25 experiments of 10 coin-tosses is shown in Figure 5.8. This is a distribution, as it shows how often the heads occurred in each of the 25 experiments.

Figure 5.9 shows the age distribution in UK according to the 2011 census data. Note that the ages are grouped together in intervals of five to allow easier graphical representation compared to Figure 5.8, where each possible combination of heads and tails is represented by a bar.

Figure 5.8 Frequency of heads in 25 experiments of 10 coin-tosses in each experiment

H = Heads; T = Tails

Each possible combination of heads and tails is represented by a bar; the number inside the bar represents the number of experiments in which the combination was obtained. The height of the bars is proportional to the number of experiments in which the combination was obtained.

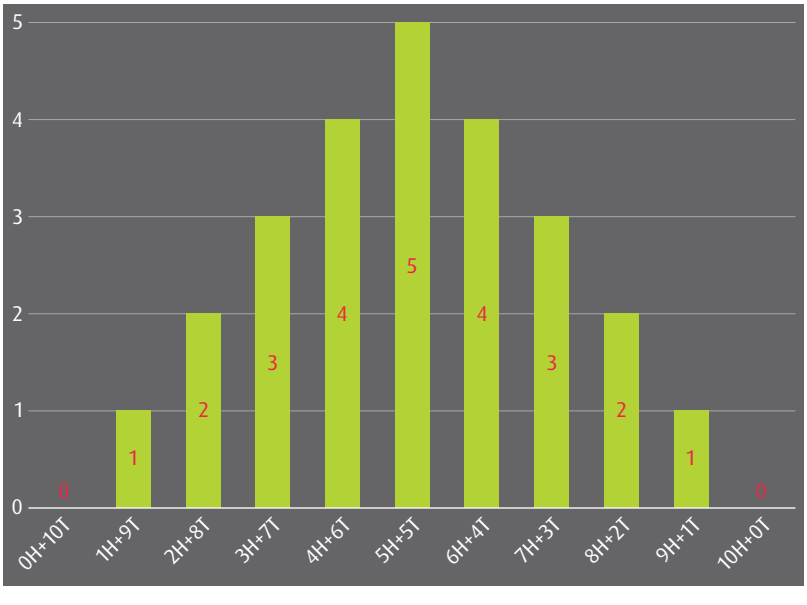
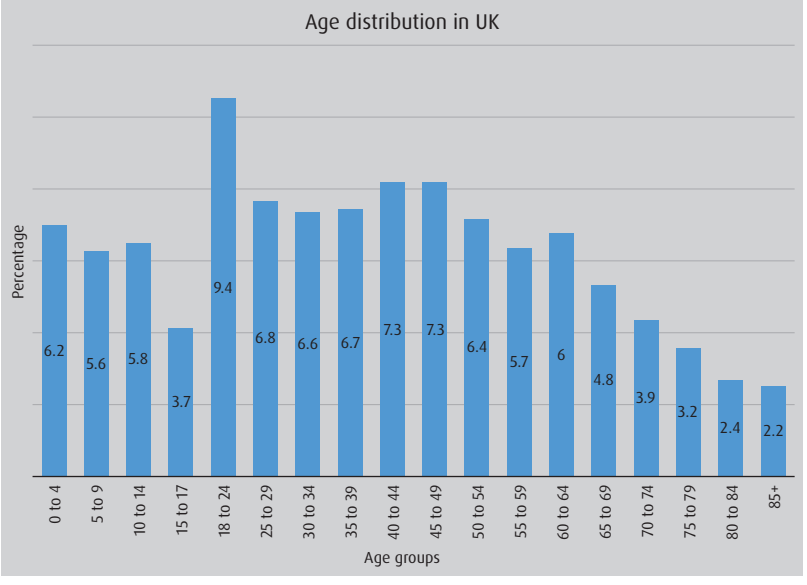


Figure 5.9 Age distribution in UK (2011 census data)

The bars represent each age group shown on the x-axis. The percentage of people in the UK who belonged to the age group is indicated inside the bar. The height of the bars is proportional to the percentage of people who belonged to the age group.



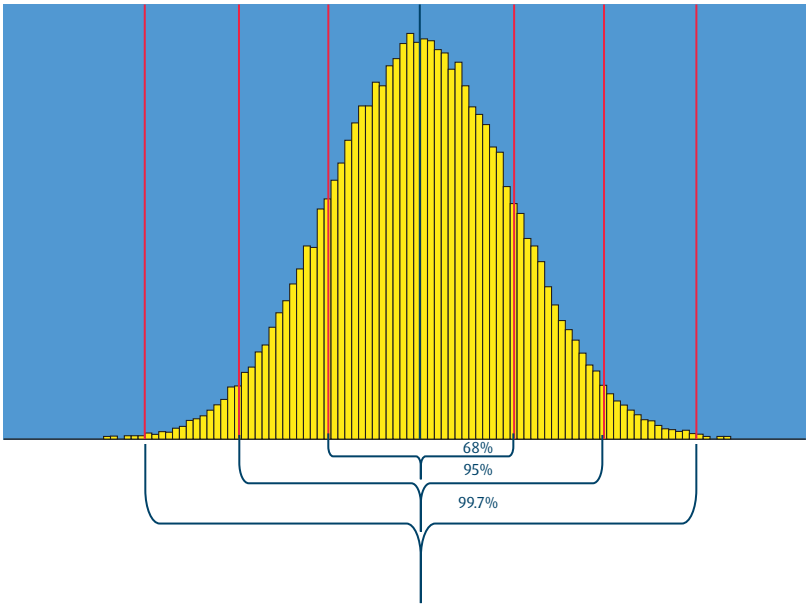
Normal versus non-normal distribution

There are many types of data distributions. [6] In the context of systematic reviews of intervention (comparing an intervention with a comparator), it is necessary to distinguish between normal and non-normal distribution.

Normal distribution (or 'Gaussian' distribution or 'parametric' distribution) is represented by a family of curves which are symmetrically bell-shaped, although just the fact that a curve is bell-shaped does not mean that it represents a normal distribution. This is because other distributions may have a similar sort of shape. [2] In addition to being symmetrical and bell-shaped, normal distribution has additional characteristics related to its summary measures, namely, mean and standard deviation, which are described later in this chapter. In normal distributions, mean±1 standard deviation covers 68% of the observations, mean±2 standard deviations cover 95% of the observations, and mean±3 standard deviations cover 99.7% of the observations. Here ± standard deviations indicate adding the multiples of standard deviation to the mean or subtracting the multiples of standard deviation from the mean. Figure 5.10 shows a normal distribution.

Figure 5.10 Normal distribution

The values are symmetrical around the blue line representing the mean. The red lines indicate multiples of standard deviation. Mean±1 standard deviation covers 68% of the observations, mean±2 standard deviations cover 95% of the observations, and mean±3 standard deviations cover 99.7% of the observations.



RELATION BETWEEN DATA TYPES, DATA DISTRIBUTION AND DATA TO BE EXTRACTED

The data to be extracted for different data types and distribution are shown in Table 5.2.

While information such as 'number of participants belonging to each category' is obvious, other terms such as median, quartiles, mean and standard deviation need further explanation. Median, quartiles, mean and standard deviation are types of summary statistics, that summarise the quantitative data. [2] Further descriptions of these terms follow.

Mean

Mean is otherwise known as average. To calculate the mean, we add up the observed values and divide by the number of observations. [2] 'Mean' indicates arithmetic mean and is the preferred measure to indicate the centre of a normal distribution.

Table 5.2 Summary measures in a single group for different types of data

Type of data	Data extracted[1]
Categorical data: binary data	Number of participants belonging to a specific category
Categorical data: nominal data (multiple categories)	Number of participants belonging to each category
Categorical data: ordinal data (few categories)	Number of participants belonging to each category
Categorical data: ordinal data (multiple categories)	Median and quartiles
Quantitative data: continuous data (normal distribution)	Mean and standard deviation
Quantitative data: continuous data (non-normal distribution)	Median and quartiles
Quantitative data: count data	Counts
Time-to-event data[2]	Hazard rate[3]

Notes

1. The number of participants should be extracted for all data types.
2. This is only applicable for outcome data.
3. This is not usually required because of some assumptions made while comparing the outcomes between the groups.

There are other forms of mean such as 'geometric mean' or 'harmonic mean', but these are calculated for non-normal distributions.

Standard deviation

Standard deviation is a measure of the variation between the samples, [2] and is the preferred measure of variation for normally distributed samples.

Median

Median is a measure of location of the data: that is, it indicates the middle of the data such that 50% of the observations lie above the median and 50% of the observations lie below the median. [2] Median is the preferred measure to indicate the centre of a non-normal distribution.

Quartiles

The lower and upper quartiles represent the 25th and 75th percentile of a set of observations. [2] This means that 25% of observations are below and 75% of observations are above the lower quartile; for the upper quartile this means

that 75% of observations are below and 25% of observations are above the upper quartile.

Quartiles are the preferred measures of variation for non-normally distributed samples.

WHAT DATA SHOULD BE EXTRACTED?

Having learnt the relationship between data types, data distribution and the data to be extracted from single groups, we will now go on to learn about the data to be extracted.

In systematic reviews of intervention, the data extracted include data related to the source of information, PICO details, participant flow, follow-up, risk of bias, and conflicts of interests in the study. (Note that in previous chapters, 'P' in PICO referred to population but in subsequent chapters, it will refer to participants; this is because the research question and selecting studies were at population level, while the data extraction and analysis are at trial participants level.) Participant flow and follow-up can be considered as part of the outcome, as these details are related to the effectiveness of an intervention versus comparator on the outcome. Therefore, these are considered under the outcome section.

Risk of bias might also be considered as part of the outcome, since it provides information on whether the estimate of the effectiveness of an intervention versus comparator on the outcome (that is, how well an intervention works compared to the comparator in terms of improving or worsening the outcome) is reliable. However, we have considered this separately because risk of bias is a separate topic in its own right.

Source of information (study details)

As we mentioned in Chapter 4, it is important to make the distinction between study and record (or report). Since we are interested in all the information from the study, it is important that all the available reports of the study are identified and collated.

This should include the following.

1. Name of the study: This is necessary so that the source of information is clear to the readers to allow independent verification of the results of the systematic review. This can be the acronym for the trial, trial ID, or last name of the first author followed by year of publication. This should be a unique ID for each study, for example: Doe 2021. If there was another trial with the same name, you can differentiate this by stating Doe 2021 (1), Doe 2021 (2), and so on. An alternative is to use numbers for studies, for

example, S001, S002, S003. However, for easy identification, even if you use S001, S002, S003, we recommend that you add the trial name, ID, or last name of first author followed by year of publication.

2. Trial registry number or name: This will help to identify multiple records of the same study and collate them. This will also help the healthcare professionals who might be more familiar with the acronym of the trial and can easily understand the source of information.
3. Record ID: This should be unique and refers to a record rather than a study. This will help in collating multiple records of a study.
4. Study author details (including contact details): This is useful to contact the study authors. There is currently no evidence that surface letters (printed and posted letters) are better than emails. On the contrary, a methodological review on methods to obtain unpublished data found that response rates were higher and quicker with emails than surface mails. [7] Therefore, we suggest that you extract the email address as the default contact detail and obtain the additional contact details when the email address is not available. Of course, while making the attempt to contact the authors, if the emails bounce, you might need the additional information to obtain alternate email or contact details. We suggest that information on additional contact details is obtained if necessary, rather than routinely, to avoid wasting resources unnecessarily.

PICO details

Participants

Why is it necessary to extract data on participants' details?

You can only extrapolate the results of a clinical study to the patients who are similar to the participants included in the studies. This has implications for the type of people in clinical practice for whom an intervention is recommended. Furthermore, the intervention may work differently in different groups of participants, who may all be eligible for the systematic review. Differences in participants may explain differences in the treatment effects (effectiveness of an intervention versus the comparator for an outcome) observed in the different studies. Therefore, information on participant characteristics will be helpful to find out if any differences in treatment effects between the studies are due to differences in the participant characteristics.

What should be extracted?

The inclusion and exclusion criteria in the trials should be recorded. However, when there is a long list of inclusion and exclusion criteria in the trials, it becomes difficult for both the reviewers and the readers to find out how similar or

different the participants were across different studies. Therefore, we recommend that the reviewers identify important participant characteristics that can influence how effective an intervention is (potential 'effect modifiers' or 'moderators') in a structured manner (for example, did the study include or exclude participants with certain characteristics) rather than recording a long list of inclusion and exclusion criteria in the trials verbatim from the studies. In addition to better understanding of the participants for whom the evidence is applicable, this method of capturing information can also help with subgroup analyses based on participant characteristics (that is, whether the treatment effects are different in different subsets of participants).

Information should also be captured on whether disadvantaged people were included in the study. The PROGRESS-PLUS framework can be used to identify disadvantaged people. [8] This will allow the reviewers to assess whether the treatment effects are similar in disadvantaged people (versus others) and whether the findings of the review are applicable in disadvantaged people. There may be other characteristics about the participants that the reviewer might want to record that could potentially explain the differences in treatment effects between the studies (if any), for example, symptom scores at baseline. The data extracted for PROGRESS-PLUS and other characteristics depends on whether the characteristic is of categorical data type or quantitative data type and can be guided by the information provided in Table 5.2.

Intervention and comparator

Why is it necessary to extract data on the details of the intervention and comparator?

You can only extrapolate the results of a clinical study to the intervention and comparator which are similar to those included in the studies. Therefore, sufficient details of intervention and comparator should be recorded to allow them to be implemented in clinical practice, that is, whether these are choices that could be offered to the patient as part of shared decision making.

Furthermore, the differences in the treatment effects across studies (if any) may be explained in differences in the intervention or the comparator.

What should be extracted?

The 'template for intervention description and replication' (TIDieR) checklist [9] can be used to capture sufficient details of the intervention and comparator to allow their replication in clinical practice. This will also help with exploring differences in treatment effects (across studies) due to differences in either the intervention or the comparator or both.

Outcome

Why is it necessary to extract data on outcomes?

In a systematic review of interventions, the main objective is usually to find out the effectiveness of an intervention versus comparator on an outcome. Therefore, it is necessary to extract outcome-related data.

In addition, differences in the way an outcome is measured might explain the differences in treatment effects across studies.

What should be extracted?

Participant flow

Participant flow is the term that is used in clinical trials and systematic reviews to describe the following.

1. Number of potential participants screened: Along with the number recruited, it provides a measure of the proportion of participants in clinical practice in whom the evidence can be applied.
2. Number recruited: Besides contributing to the calculation of the proportion of participants in clinical practice in whom the evidence can be applied, it can also be used to check whether the sum of the numbers allocated to each of the intervention and comparator groups equals the number recruited.
3. Reasons for non-recruitment: While it is reasonable to expect exclusion of participants who are ineligible and those who decline participation in the study, there may be other reasons for non-recruitment, for example, those who are unlikely to complete the intervention or comparator for various reasons, which might help in estimating the proportion of eligible patients who are likely to benefit from an effective intervention. In retrospective studies, if lack of data is a reason for non-recruitment, you must consider whether this can introduce bias. Further explanations related to bias will be provided later in this chapter.
4. Number allocated to each of the intervention and comparator groups: These form the basis for estimating the effect of starting an intervention versus the comparator.
5. Number receiving each of the intervention and comparator as planned: These form the basis for estimating the effect of starting and adhering to an intervention versus the comparator.
6. Number crossed-over to the other group, that is, intervention group receiving the comparator and vice versa: These can help in estimating whether the main reason for equivalence of an intervention versus the comparator was

because of cross-over. For example, consider waiting (intervention) versus surgery (comparator) for a health condition. If the study shows equivalence, but most participants in the waiting group had to undergo surgery because of worsening of symptoms in a short period of time, waiting cannot be considered equivalent to surgery for the health condition.

7. Reasons for cross-over: It is important to know the reasons for cross-over. For example, if this is because of lack of equipment, you might consider such cross-overs to estimate the proportion of participants in clinical practice in whom the evidence can be applied. On the other hand, a high proportion of participants crossing over to the comparator because of worsening of symptoms (as in the last example) would indicate that the intervention is ineffective.
8. Number of participants who had missing data, and the reasons: The number of participants who had missing data and the reasons for this are important in assessing the risk of bias and are covered later in this chapter.

Follow-up

The period of follow-up is integral to assessment of whether an intervention is effective versus a comparator on an outcome. For example, let us consider a comparison of a less invasive thermal ablation versus more invasive surgical treatment for cancer. The less invasive treatment may have fewer complications in the short term but may not provide equivalent cancer clearance to the more invasive treatment. Therefore, if the overall survival is measured at short term, say three or six months, the less invasive treatment may provide equivalent survival to the more invasive treatment, but if the overall survival is measured at long term, say 10 years, the less invasive treatment may be inferior to the more invasive treatment.

When all the participants are followed for a fixed period of time, say two years, the minimum, mean and median follow-up of participants will be about two years. However, when the participants are followed for different time periods (as in many clinical trials where participants are recruited during a visit to the healthcare professional), usually median follow-up is recorded. Mean follow-up or minimum follow-up may be other options if median follow-up is not reported.

Outcome data

Outcome data can be descriptive or numerical.

Descriptive outcome data

The definition or scale used to measure an outcome should be recorded. This is particularly important when there is no standard definition for the outcome or when multiple scales are available to measure an outcome.

Sometimes no numerical data are available from the studies and when numerical data are available, they are insufficient to calculate the treatment effects. In such situations, a narrative description by the study author related to the effectiveness of an intervention versus the comparator should be recorded.

Numerical outcome data from simple parallel RCTs

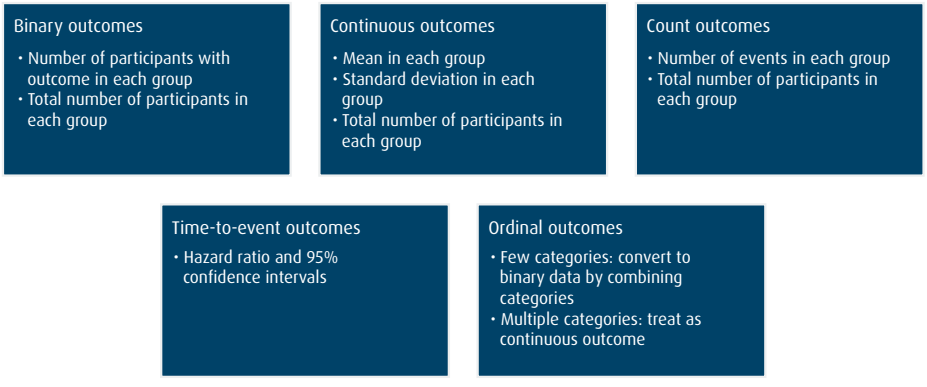
The numerical data extracted for outcomes depend upon the data type. An overview of the data to be extracted for different outcome types from simple parallel RCTs is shown in Figure 5.11.

While most of the terms used in Figure 5.11 were explained earlier in this chapter, the term confidence intervals (CI) is a new term and requires additional explanation. To understand the concept of confidence intervals, you need to understand the concept of population and sample. If you want to find the properties of a population, you can study the whole population. An example of such a study is the census, where the measurements are made on each individual of a population. However, this is not practical in every study because of the resources in terms of cost and time required to make these measurements. Therefore, a representative sample from the population is usually studied to make inferences about the population. This is possible because of a special statistical concept called the Central Limit Theorem. [10]

The 95% CI is the range within which 95% of the sample means from a population are expected to fall. [2] It also means that there is only a 5% probability that the 95% CI excludes the population mean. [2] The 95% CI can be calculated from a single sample using a formula, so there is no requirement to take repeated samples from a population to calculate the 95% CI.

In the context of clinical trials, you can consider that the participants included in the trial are a sample drawn from the patient population. The treatment effect

Figure 5.11 Numerical data extracted for outcomes from simple parallel RCTs



(of the intervention versus the comparator) observed in the trial is an estimation of the treatment effect in the patient population. The 95% CI of the treatment effect means that there is only a 5% probability that the 95% confidence interval excludes the average treatment effect in the patient population.

Final versus change scores for continuous outcomes

One of the questions frequently asked by researchers relates to which data to extract when the study authors report the change scores (change of an outcome from baseline observation) and final scores (final value of an outcome) for continuous outcomes. For example, in a systematic review on the impact of lifestyle modification on weight, some study authors might report the change in weight in each treatment group from baseline, while others might report the final weight in each treatment group. We recommend that the review authors extract whichever is reported fully. For example, some study authors might have reported the mean change scores along with the standard deviation but might have reported only the mean final scores; others might have reported the mean final scores along with the standard deviation, but not the change scores at all. While it is possible to calculate the mean final scores from the mean baseline and mean change scores, it is usually not possible to calculate the standard deviation of the final scores from the information available in the study reports. Similarly, it is possible to calculate the mean change scores from the mean baseline and mean final scores, but it is usually not possible to calculate the standard deviation of the change scores from the information available in the study reports. When both change scores and final scores are reported fully, we suggest that the review authors extract both, but only one of the scores might be used in the analysis depending on the analysis performed. Further details about analysing data when some study authors report the change scores and others report final scores are covered in Chapter 6.

Numerical outcome data when preferred data are not available from simple parallel RCTs

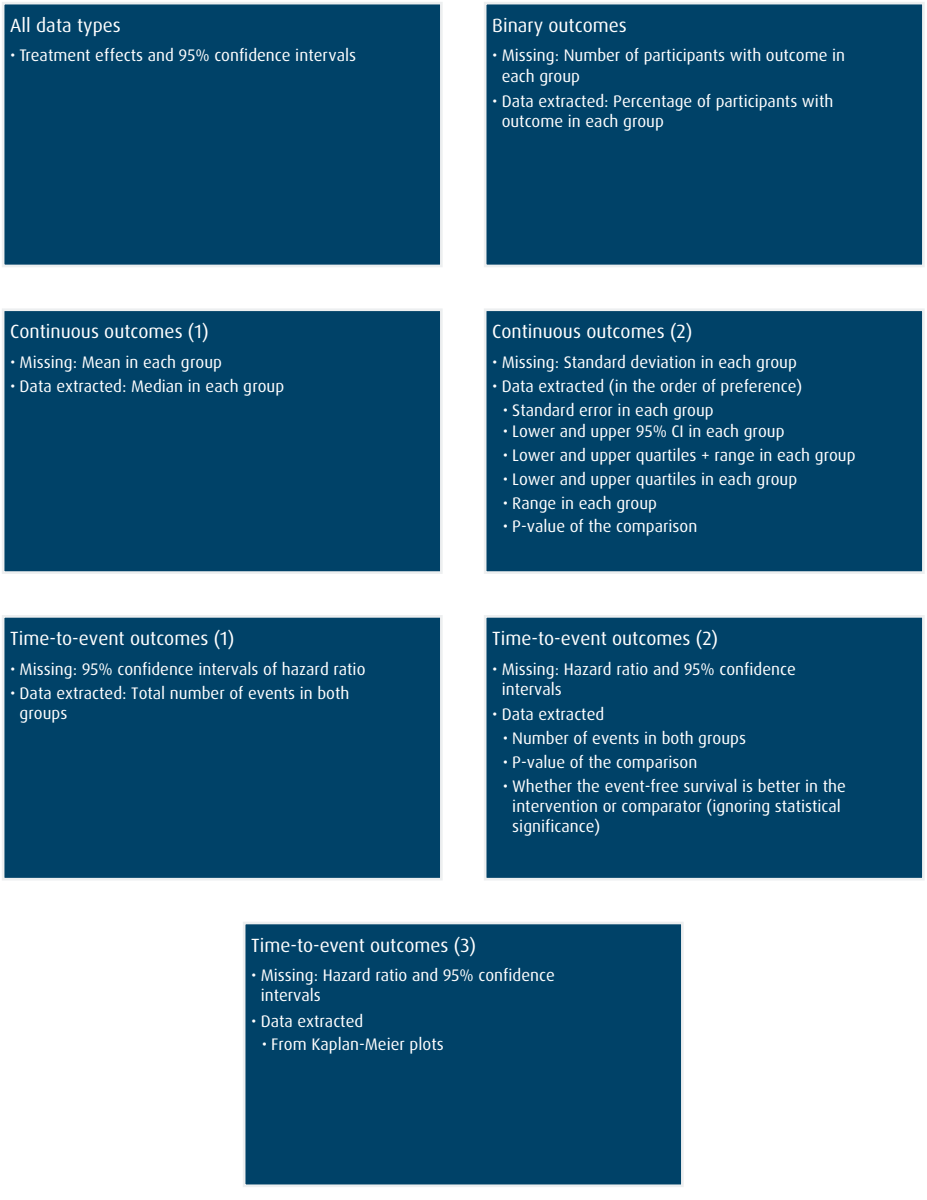
When the preferred outcome data are not available, you might be able to impute (make an educated guess) the missing information or calculate this from other available data. This is shown in Figure 5.12.

While most of the terms are self-explanatory or have been defined previously, some terms need additional explanation.

- Standard error: Standard error is the short form for 'standard error of means' and is a concept related to making inferences about a population from samples of a population (much like the CI; in fact, the 95% CI are calculated from the standard error, the sample mean and a constant).

Figure 5.12 Imputing or calculating missing information from available data in simple parallel RCTs

CI = confidence intervals



- Range: This is the difference between the lowest and highest observed values in a sample.
- *p*-value: *p*-value is a concept in testing a hypothesis. The American Statistical Association provides a description of *p*-value. [11] In the context

of comparing two groups in clinical trials and testing the assumption that the effect of both the treatment groups on an outcome is the same, the *p*-value indicates the probability that the actual difference between the two group is at least the observed difference (provided that there is no bias).
- Kaplan-Meier plots. These are plots that show the proportion of people without the events among those at risk of developing the event. Further information about extracting data from Kaplan-Meier plots is shown in the section 'Data extraction from plots'.

The methods for converting percentage, standard error, lower and upper quartiles, range and *p*-value to the format suitable for meta-analysis will be described in Chapter 6.

Numerical outcome data from cross-over or cluster RCTs

In cross-over or cluster RCTs, you should extract the treatment effects and 95% CI adjusted for the cross-over or clustering effect. When these are not available from the studies directly, you can extract the same data as for simple RCTs and, in addition, extract the following information.

- Cross-over RCTs: intraclass correlation coefficient.
- Cluster RCTs: number of clusters in each group and intraclass correlation coefficient.

Intraclass correlation coefficient is a measure of the similarity in the outcomes measured in the same individual (compared to other individuals) for cross-over RCTs, and the similarity in the outcomes in individuals belonging to a cluster (compared to those belonging to other clusters) for cluster RCTs.

Numerical outcome data from special situations

While it is relatively easy to identify cross-over or cluster RCTs, there are other scenarios where similar principles apply. It is important to recognise these special situations.

Randomisation at participant level but outcome is measured at body part level

This happens when the randomisation occurs at the participant level, but the outcome is measured at body part level. For example, participants may be randomised to receive prophylactic antibiotics or not receive prophylactic antibiotics before an operation which requires multiple incisions, but the surgical site infection is reported at the incision level. When the number of participants randomised to each group is available, you can treat this as count data. Regardless

of how the study authors analyse this, the review author can simply extract the number of surgical site infections in each group and the number of participants in each group in the usual way of extracting data for count outcomes. However, when the study authors do not report the number of participants randomised in each group but mention only the number of incisions in each group (although the randomisation was at the participant level), the situation is analogous to a cluster RCT. In this situation, each participant is equivalent to a cluster and each incision is equivalent to an individual. Therefore, you need the treatment effects and its 95% CI adjusted for clustering.

Randomisation at body part level

This happens when different parts of the body of an individual are randomised to different treatments. This is called split-body part design. For example, machine perfusion and cold storage are two methods of preserving organs. In cadaveric kidney donation, one kidney in each donor could be randomised to receive machine perfusion or cold storage. Treatment effects adjusted for the correlation of outcomes of recipients receiving kidneys from the same donor should be extracted if available. Alternatively, intraclass correlation coefficients should be extracted along with the outcomes in each group. Split-mouth design is an example of randomisation at body part level.

Numerical outcome data from non-randomised studies

For non-randomised studies, the treatment effects and 95% CI should be extracted. These could be adjusted for confounding variables (variables other than the treatment group which affect the outcome) or unadjusted for confounding variables. We recommend extracting both the unadjusted and adjusted treatment effects and 95% CI when available, to obtain a measure of the influence of the confounders on the treatment effects.

In the absence of treatment effects and 95% CI (particularly, the adjusted treatment effects which will be used for the interpretation), the data can be extracted as for RCTs.

Data extraction from plots

Any of the numerical data mentioned above can be obtained from plots if these are not available from the study. Caution is required when you are extracting standard deviation from error bars in plots. The error bars may represent standard deviation, standard error, or 95% CI. Unless the study authors specify what the error bars represent, you should not extract the data, as these measures differ considerably in the values and can impact the analysis results (and therefore, conclusions).

Data extraction from Kaplan-Meier plots require additional guidance as follows.

- The event rate within a time interval should be no more than 20% of that at the start of the time interval. [12,13]
- Time intervals which are consistent across studies can be chosen. [12]
- If the curve starts to level off, there is little value in extracting data from this area of a curve. [13]
- The final interval should not extend beyond the maximum follow-up. [12,13]

Once you have extracted the proportions without the event in each of the treatment groups at the specified time points chosen using the above guidance, you can obtain the data in the required format for meta-analysis using the Excel file provided by Tierney et al. [13]

EQUAL-SR software (available from https://sites.google.com/view/equal-group/home) can help with data extraction of plots and for calculating the hazard ratio and its 95% CI. Please see the instructions available from the software for further guidance.

Risk of bias

What is bias?

We have mentioned bias several time in this chapter. What is bias? In the context of clinical trials and systematic reviews, bias means deviation of the value of a statistical estimate from the quantity it estimates. [14]

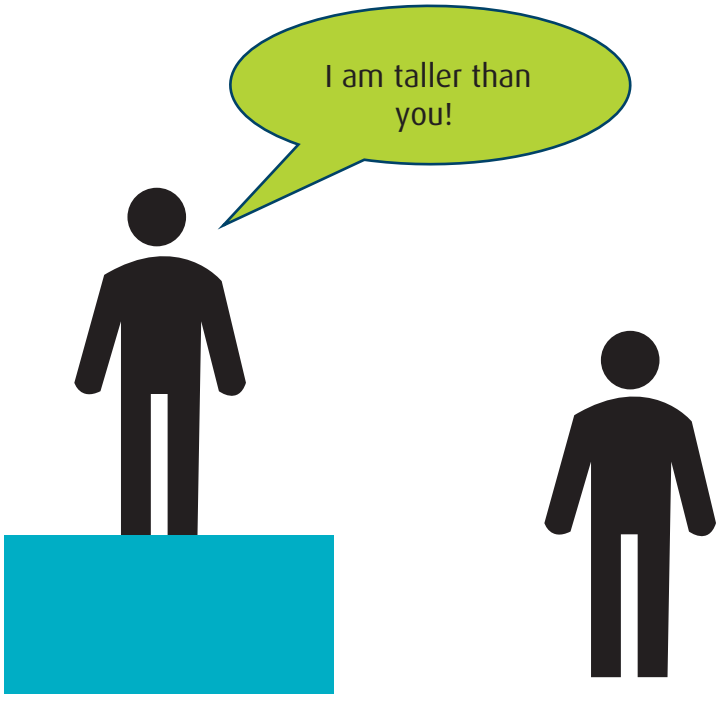
This is best explained by an analogy. Suppose we want to estimate the difference in heights between two people. To allow a fair estimation of the difference in heights between the two people, they should stand on the same level. What if one person stands on a platform while the other stands on the ground level while estimating the difference in heights? The difference estimated in this fashion will not be the actual difference in heights between the two people. This will be considered an unfair comparison. Figure 5.13 depicts this analogy in a graphical format.

How can we assess bias?

In the analogy (Figure 5.13), it is clear that one person is standing on a platform and the other person is standing on the ground level. However, in clinical trials, there is no 'visible' object to understand whether the comparison between the intervention and the comparator was unfair. If that is the case, how do you know if the treatment effect observed in the trial is the accurate treatment effect in the sample studied? You can use a risk of bias assessment tool, which is essentially a

Figure 5.13 Analogy for bias

In this analogy, the two people represent the intervention and comparator; difference in heights represents the treatment effect in a clinical trial; and the platform represents one or more aspects of clinical trial design, conduct or analysis that result in an unfair comparison between the intervention and comparator in terms of an outcome in the trial (the observed effect in the trial is not the true effect). This means that the treatment effect in the patient population is also wrongly estimated and wrong treatments may be given to the patient.



set of assessments developed by methodological experts that allow us to decide whether the study provides an unbiased estimate of the treatment effect in the sample. While different tools have different terminologies, the broad classifications include low risk of bias, moderate risk of bias or high risk of bias to provide an indication of whether the treatment effect observed in the trial is an accurate portrayal of the treatment effect in the sample studied, and is estimating the treatment effect in the population correctly.

In systematic reviews of intervention, the risk of bias in RCTs is assessed by RoB 2, [15,16] and that in non-randomised studies is assessed by ROBINS-I. [17,18] Sufficient details should be extracted to allow the assessment of risk of bias in the studies.

Risk of bias in randomised trials (RoB 2)

The RoB 2 tool has five domains or aspects of assessment.

- Bias arising from the randomisation process.
- Bias due to deviations from intended interventions.
- Bias due to missing outcome data.
- Bias in measurement of the outcome.
- Bias in selection of the reported result.

Each of the domains has multiple signalling questions, that is, questions that can be answered as 'Yes', 'Probably Yes', 'No Information', 'Probably No', and 'No'. Some of the questions can also be answered as 'Not applicable' depending on the answers provided to earlier questions. The answers to the signalling questions allow you to classify the risk of bias in each of the domains as 'Low risk', 'Some concerns' or 'High risk'. From this perspective, the answers 'Yes' and 'Probably Yes' for the signalling questions are treated in the same way; the answers 'No' and 'Probably No' are treated in the same way. While the assessment of the first domain is at the study level, the remaining domains are assessed at the outcome level. Such outcome-level assessments are used to interpret the evidence. The overall risk of bias for the outcome can be calculated from the risk of bias classifications for each domain. Broadly, a study can be considered to be at 'high risk of bias' (for an outcome) if any of the domains are classified as 'High risk' or multiple domains are classified as 'Some concerns' in a way that substantially lowers the confidence in the results; it can be considered to be at 'low risk of bias' (for the outcome) if all the domains are classified as 'Low risk'; the remaining combinations of classifications result in a study being classified as 'Some concerns' (for the outcome). [15]

Mechanisms of how each aspect leads to risk of bias, and detailed guidance for the risk of bias classifications are available at [16]. We provide a short description and some practical guidance on how to answer the signalling questions 'Yes' (or 'Probably Yes') and 'No' (or 'Probably No') in this chapter. Any signalling questions which are dependent on the answers provided for earlier questions can be identified from the wording of the signalling question. Therefore, we have not specified when the question should be answered 'Not Applicable'. The signalling questions for which there is insufficient information to answer as 'Yes' (or 'Probably Yes'), 'No' (or 'Probably No'), or 'Not Applicable' are answered as 'No Information'. Therefore, we have not provided guidance on when the signalling question should be answered 'No Information'.

The algorithm to calculate the risk of bias for each domain from signalling questions is available from the publication and guidance documents of the RoB 2 team. [15,16] Automated calculation of the risk of bias is also possible using the

Excel file provided by the RoB 2 team. [16] The risk of bias for each domain and the overall risk of bias can also be calculated automatically using the EQUAL-SR software.

Bias arising from the randomisation process

In an RCT, a participant is allocated to receive the intervention or the comparator randomly. This is to prevent bias due to confounding which arises when the participants who received the intervention and the comparator differ by one or more important prognostic factors that determine the outcome. [16] The signalling questions related to this domain and how they should be answered are as follows. [15,16]

1.1 Was the allocation sequence random?

'Yes' or 'Probably Yes': Allocation was performed using computer-generated random numbers, minimisation with a random element, a sequence from a random number table, or a random sequence generated by coin tossing, shuffling cards or envelopes, throwing dice, or drawing lots.

'No' or 'Probably No': Allocation was performed without a random element or sequence was predictable, for example, alternation, birth date, admission date, patient record numbers, allocation by clinicians or participants, allocation based on the availability of the intervention or the comparator.

1.2 Was the allocation sequence concealed until participants were enrolled and assigned to interventions?

'Yes' or 'Probably Yes': Allocation was concealed using centrally administered methods such as web-based or telephone-based randomisation service; a third party not involved in enrolment of participants; sealed envelopes that were opaque, sequentially numbered, sealed with a tamper-proof seal, and opened only after the envelope had been irreversibly assigned to the participant; drug containers that were sequentially numbered, of identical appearance, and dispensed or administered only after they had been irreversibly assigned to the participant.

'No' or 'Probably No': There is suspicion that the enrolling investigator or the participant had knowledge of the next allocation.

Some researchers are confused between allocation concealment and blinding (a concept explained in the next domain). The key difference between the two is whether the researcher and patient are unaware of the assignment before recruiting into the study (which is always possible in an RCT) or after recruiting into the study (which may not be possible in many situations).

1.3 Did baseline differences between intervention groups suggest a problem with the randomisation process?

'Yes' or 'Probably Yes': One or more of the following are observed in the study.

- Substantial differences between sample sizes of the intervention versus comparator compared to the planned allocation ratio. If the planned allocation ratio was 1:1, the sample sizes of the intervention and comparator groups should be approximately the same; if the planned allocation ratio was 2:1, the sample size of the intervention group should be approximately twice that of the comparator group, and so on.
- A substantial excess in statistically significant differences in the participant characteristics or prognostic factors between the intervention and comparator, beyond what is expected by chance. For example, if the sample size was large and the number of baseline characteristics being compared between the groups is also large, you can expect approximately 5% of the comparisons to show statistically significant differences at a p-value of 0.05 even if there were no real differences.

No' or 'Probably No': No imbalances are reported or if reported, they might have arisen by chance (please also see the previous point about when you can consider that the differences might have arisen by chance).

Bias due to deviations from intended interventions

Before assessing this domain, you need to understand that two effects of the intervention versus comparator may be of interest to the healthcare professionals and the patients. The first effect is the effect of starting the intervention (versus comparator). This is called the intention-to-treat effect. The second effect is the effect of adhering to the intervention as planned (versus comparator). This is called the per-protocol effect.

In the context of shared decision making, the main effect of interest is the intention-to-treat effect, as it has to be decided whether the treatment should be the intervention or the comparator. Per-protocol effect may be of interest in the context of a treatment which has to be repeated ('sustained treatment'), for example, medicines or other treatments that have to be repeated periodically. This is because we will be able to provide information to the healthcare professionals and patients as to why adherence to the treatment as planned is necessary to obtain the desired effect by comparing the intention-to-treat and per-protocol effects. However, per-protocol effect is likely to have a limited role in the context of shared decision making for one-off treatments delivered by

healthcare professionals who are able to deliver the treatment as planned unless some findings during the treatment necessitate a change in treatment.

The signalling questions for the intention-to-treat effect and per-protocol effect are different. However, to answer these signalling questions correctly, you also need to understand the concept of intention-to-treat analysis, modified intention-to-treat analysis, per-protocol analysis and as-treated analysis. Intention-to-treat analysis refers to an analysis that includes all the trial participants according to the group to which they were randomised. [19] A modified intention-to-treat analysis refers to an analysis that includes only the trial participants in whom the outcome could be measured, but the participants are included under the groups to which they were randomised. [16] A naive per-protocol analysis refers to an analysis that includes only trial participants who received the treatments as per the protocol. [16] An as-treated analysis refers to an analysis in which the participants are analysed according to the treatment they received rather than to the group to which they were randomised. [16] An appropriate per-protocol analysis should include adjustments for prognostic factors that resulted in deviations from the planned treatment. Advanced statistical methods, such as instrumental variable analysis for one-off treatments and inverse probability weighting for sustained treatments, are necessary to calculate the correct per-protocol effect. The description of these advanced statistical methods is beyond the scope of this book; interested readers can read an introduction to these methods available from the literature. [20,21]

Another concept that requires explanation is the concept of 'blinding'. Blinding refers to one or more of the healthcare providers, outcome assessor or the participant being unaware of the treatment that a participant received. The main reason for blinding is to prevent bias due to deviations from the planned treatment and differences in the assessment of outcomes in the intervention versus comparator. [15,16]

Blinding is usually achieved by 'dummy treatment' or placebo. While the use of placebo may be feasible for many pharmacological interventions, it may not be possible in non-pharmacological interventions. For example, it is not possible to blind the healthcare provider to the participant group if the healthcare provider has to perform different procedures in the intervention versus comparator, say surgery versus no treatment. It may not be possible to blind the participants in some trials, for example, patient education versus no intervention in adults. The use of placebo may be unethical in some situations. For example, placebos for surgical interventions are only recommended when the study question cannot be answered with a different design and the risks of performing a sham operation are outweighed by the importance of the knowledge to be gained. [22]

The signalling questions related to this domain and how they should be answered are as follows. [15,16]

Intention-to-treat effect

2.1 Were participants aware of their assigned intervention during the trial?

'Yes' or 'Probably Yes': The study was reported as 'open' or 'open label', lack of blinding was mentioned or discussed as a limitation, blinding is difficult to achieve ethically and there is no description about blinding, or the scars or side effects are unique to a treatment, allowing participants to know the group to which they were allocated, based on scars or side effects.

'No' or 'Probably No': Blinding of participants was achieved.

2.2 Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?

'Yes' or 'Probably Yes': Blinding is impossible to achieve, the study was stated as 'open label', lack of blinding was mentioned or discussed as a limitation, the scars or side effects are unique to a treatment, allowing carers and people delivering the interventions to know the group to which participants were allocated based on scars or side effects.

'No' or 'Probably No': Blinding of healthcare professionals was achieved.

2.3 If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention that arose because of the trial context?

'Yes' or 'Probably Yes': Cross-overs to the other treatment because of the perceived concern about lack of effect (initiated by the participant or the healthcare professional) were reported, or the follow-up in the trial participants is different from that in standard clinical practice.

'No' or 'Probably No': Any changes to the treatment are changes that would be normally expected in clinical practice (for example, reduction in dosage of a drug, complete cessation of a drug because of side effects, treatment of a complication).

2.4 If Y/PY to 2.3: Were these deviations likely to have affected the outcome?

'Yes' or 'Probably Yes': Changes to the treatment are likely to affect the outcome.

'No' or 'Probably No': Changes to the treatment are not likely to affect the outcome.

2.5 If Y/PY/NI to 2.4: Were these deviations from intended intervention balanced between groups?

'Yes' or 'Probably Yes': Changes to the treatment were balanced between the groups (that is, changes were in similar proportions to the intervention and comparator).

'No' or 'Probably No': Changes to the treatment were unbalanced between the groups (that is, changes were not in similar proportions to the intervention and comparator).

2.6 Was an appropriate analysis used to estimate the effect of assignment to intervention?

'Yes' or 'Probably Yes': Intention-to-treat analysis or modified intention-to-treat analysis was used, or the only exclusions were those of ineligible participants when the treatment does not influence the eligibility.

'No' or 'Probably No': Naive per-protocol analysis or as-treated analysis was used, or eligible participants were excluded from the analysis for reasons such as toxicity or lack of efficacy.

If sufficient information is available from the studies, it might be possible to re-analyse the data using appropriate analysis and include this in the systematic review. If that is the case, the answers to this question should be based on the re-analysed data by the systematic reviewer and not based on the inappropriate analysis by the study author.

2.7 If N/PN/NI to 2.6: Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomised?

The answers to this question can only be determined by a sensitivity analysis (analyses performed with different assumptions). Therefore, we recommend that the systematic review authors answer this question after performing the analysis. There is no fixed proportion of exclusions that can be used to guide the answer to this signalling question.

'Yes' or 'Probably Yes': Sensitivity analysis showed an impact on the results.

'No' or 'Probably No': Sensitivity analysis did not show an impact on the results.

Per-protocol effect

2.1 Were participants aware of their assigned intervention during the trial?

Same as for intention-to-treat effect.

2.2 Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?

Same as for intention-to-treat effect.

2.3 [If applicable] If Y/PY/NI to 2.1 or 2.2: Were important non-protocol interventions balanced across intervention groups?

Important non-protocol interventions are unplanned interventions that are given along with the treatment or after the treatment is started and have the potential to influence the outcome.

'Yes' or 'Probably Yes': Important non-protocol interventions were balanced between the treatment groups.

'No' or 'Probably No': Important non-protocol interventions were not balanced between the treatment groups.

2.4 [If applicable] Were there failures in implementing the intervention that could have affected the outcome?

This usually refers to one-off treatments by healthcare professionals.

'Yes' or 'Probably Yes': Failure of implementing the treatment as planned could affect the outcome.

'No' or 'Probably No': The treatment was implemented as planned for most participants.

2.5 [If applicable] Was there non-adherence to the assigned intervention regimen that could have affected participants' outcomes?

This usually refers to sustained treatments.

'Yes' or 'Probably Yes': Non-adherence to the treatment, for example, stopping the treatment, cross-over to the other treatment, another active treatment (not being compared in the trial) is high.

'No' or 'Probably No': Most participants adhered to the planned treatment.

2.6 If N/PN/NI to 2.3, or Y/PY/NI to 2.4 or 2.5: Was an appropriate analysis used to estimate the effect of adhering to the intervention?

'Yes' or 'Probably Yes': Appropriate per-protocol analysis (instrumental variable analysis for one-off treatments and inverse probability weighting for sustained treatments) was used.

'No' or 'Probably No': Intention-to-treat analysis, naive per-protocol analysis or as-treated analysis was used, or all participants in one group received an important non-protocol treatment.

Bias due to missing outcome data

3.1 Were data for this outcome available for all, or nearly all, participants randomised?

In the context of this question, imputed data should be considered missing data.

'Yes' or 'Probably Yes': outcome data are available for all participants and an intention-to-treat analysis is used when the desired effect is intention-to-treat effect. For continuous outcomes, data are available for 95% or more participants. For binary outcomes, this depends upon the sensitivity analysis. Therefore, if participants were missing, we recommend answering this question based on the sensitivity analysis.

'No' or 'Probably No': When data are not available for all or nearly all randomised participants.

3.2 If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data?

'Yes' or 'Probably Yes': Sensitivity analysis demonstrating that the results do not change with different assumptions about missing data was

available or appropriate analysis methods that correct for missing data were used. The sensitivity analysis can be performed by the study author or can be performed by the review author.

'No' or 'Probably No': Sensitivity analysis demonstrating that the results change with different assumptions about missing data was not available and there is insufficient data for performing the sensitivity analysis.

3.3 If N/PN to 3.2: Could missingness in the outcome depend on its true value?

'Yes' or 'Probably Yes': Missingness was unrelated to the outcome, for example, interruption to data collection.

'No' or 'Probably No': Missingness was related to the outcome, for example, loss to follow-up or withdrawal was related to lack of improvement of symptoms.

3.4 If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?

'Yes' or 'Probably Yes': Missingness is likely to depend on its true value if any of the following conditions are met.

- Differences in the proportions of missing outcome data in the treatment groups.
- Reported reasons for missing outcome data provide evidence that missingness in the outcome depends on its true value, for example, patients with serious adverse events were excluded from the analysis of HRQoL or mortality.
- Reported reasons for missing outcome data differ between the intervention groups.
- The circumstances of the trial make it likely that missingness in the outcome depends on its true value. For example, in trials evaluating chronic conditions, missingness may be related to the failure of improvement of symptoms.
- For time-to-event outcomes, participants' follow-up is censored when they stop or change their assigned treatment rather than at the last follow-up of patients.

'No' or 'Probably No': Analysis accounted for participant characteristics that are likely to explain the relationship between missingness and outcome.

Bias in measurement of the outcome

4.1 Was the method of measuring the outcome inappropriate?

'Yes' or 'Probably Yes': Method of measurement of outcome was inappropriate, for example, the instrument used to measure the outcome had poor validity.

'No' or 'Probably No': Method of measurement of outcome was appropriate.

4.2 Could measurement or ascertainment of the outcome have differed between intervention groups?

'Yes' or 'Probably Yes': The follow-up was not similar in the treatment groups or different definitions or methods were used to measure the outcome in the two treatment groups.

'No' or 'Probably No': The follow-up was similar, and similar definitions and methods were used to measure the outcome in both the treatment groups.

4.3 If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants?

'Yes' or 'Probably Yes': Blinding of outcome assessors was not achieved.

'No' or 'Probably No': Blinding of outcome assessors was achieved.

In participant-reported outcome measures, the outcome assessor is the participant, even if a researcher records the outcome after interviewing the participant.

4.4 If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received?

'Yes' or 'Probably Yes': The outcome is subjective, for example, patient-reported outcome measures.

'No' or 'Probably No': The outcome is objective, for example, all-cause mortality or laboratory tests results using an automated analyser.

4.5 If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?

'Yes' or 'Probably Yes': The assessment of the outcome was influenced by knowledge of the treatment received, for example, strong belief about the effectiveness of a treatment or when the treatment provider is also the outcome assessor.

'No' or 'Probably No': The assessment of the outcome was not influenced by knowledge of the treatment received.

Bias in selection of the reported result

5.1 Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalised before unblinded outcome data were available for analysis?

'Yes' or 'Probably Yes': The analysis plan was finalised before unblinded outcome data were available and the data were analysed according to the pre-specified analysis plan.

'No' or 'Probably No': The data were not analysed according to the pre-specified analysis plan.

5.2 Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible outcome measurements (for example, scales, definitions, time points) within the outcome domain?

'Yes' or 'Probably Yes': There are multiple ways of measuring the outcome and the study author does not report all the outcome measures mentioned in the pre-specified analysis plan.

'No' or 'Probably No': There is only one way of measuring the outcome, the study author reported the outcome measures according to the pre-specified analysis plan, or the study author reported the preferred outcome measure mentioned in the pre-specified systematic review protocol.

5.3 Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible analyses of the data?

'Yes' or 'Probably Yes': There are multiple ways of analysing the outcome data and the study author did not report all the analyses mentioned in the pre-specified analysis plan.

'No' or 'Probably No': There is only one way of analysing the outcome data, the study author reported the outcome data according to the pre-specified analysis plan, or the study author reported the preferred analysis mentioned in the pre-specified systematic review protocol.

Risk of bias in cross-over and cluster RCTs

The risk of bias in cross-over and cluster RCTs can be assessed using the RoB 2 tool for cross-over RCTs [23] and the RoB 2 tool for cluster RCTs. [24] These are largely based on RoB 2 with additional considerations in relation to the study design. Interested systematic reviewers should read the additional guidance available from the RoB 2 team. [23,24]

Risk Of Bias In Non-randomised Studies – of Interventions (ROBINS-I)

The ROBINS-I tool has seven domains or aspects of assessment.

- Bias due to confounding.
- Bias in selection of participants into the study.
- Bias in classification of interventions.
- Bias due to deviations from intended interventions.
- Bias due to missing data.
- Bias in measurement of outcomes.
- Bias in selection of the reported result.

Of these, the last four domains are conceptually similar to those of RoB 2. There are other similarities too between ROBINS-I and RoB 2. Similar to RoB 2,

ROBINS-I has signalling questions for each domain which help in the classification of risk of bias of that domain. [17,18] While the risk of bias for each domain is different for ROBINS-I (the risk of bias is classified as 'Low risk of bias', 'Moderate risk of bias', 'Serious risk of bias', 'Critical risk of bias' or 'No information'), the choices for the signalling questions are the same as for RoB 2, except for signalling question 1.1 which does not have 'No information' as one of the choices. As for RoB 2, the answers 'Yes' and 'Probably Yes' for the signalling questions are treated in the same way; the answers 'No' and 'Probably No' are treated in the same way.

The assessment of the first three domains in ROBINS-I are at the study level, while the remaining domains are assessed at the outcome level. The overall risk of bias for the outcome can be calculated from the worst risk of bias classifications among the domain classifications. For example, if at least one of the domains for the outcome was classified as 'Critical risk of bias', the overall risk of bias classification for the outcome is 'Critical risk of bias'; if the worst risk of bias for an outcome was 'Moderate risk of bias', then the overall risk of bias classification for the outcome is 'Moderate risk of bias'. When it is not possible to classify the risk of bias between 'Serious risk of bias' and 'Critical risk of bias' because of the classification of 'No information' for one of the domains, then the classification of the overall risk of bias for the outcome can be 'No information'.

Mechanisms of how each aspect leads to risk of bias and detailed guidance for the risk of bias classifications are available at [17]. We provide a short description of some practical guidance on how to answer the signalling questions 'Yes' (or 'Probably Yes') and 'No' (or 'Probably No') in this chapter for the same reasons as described under RoB 2.

The algorithm to calculate the risk of bias for each domain from signalling questions is available from [17,18]. Automated calculation of the risk of bias is also possible if you used the EQUAL-SR software for data extraction.

Bias due to confounding

Confounding in the context of non-randomised studies of intervention arises when one or more important prognostic factors that determine the outcome also predict the treatment that an individual receives. [17] Some confounders do not vary over time, for example, birth sex, while others may vary over time, for example, comorbidities which are related to an outcome, say, all-cause mortality. In sustained treatments, sometimes, changes between treatments ('switches') are possible, for example, switches between two anti-diabetic agents are possible in people with type 2 diabetes, and this might be related to the change in comorbidities. When changes in the confounders determine the changes in treatment, this is called 'time-varying confounding'. To adjust for 'time-varying confounding', you need to use special statistical methods such as inverse

probability weighting, or marginal structural models are required. [17] Description of such advanced models is beyond the scope of this book.

1.1 Is there potential for confounding of the effect of intervention in this study?

'Yes' or 'Probably Yes': There is potential for confounding.

'No' or 'Probably No': There is no potential for confounding. This classification is rare as it is unlikely that all confounders are known and that these are distributed equally in the absence of randomisation.

1.2 If Y/PY to 1.1: Was the analysis based on splitting participants' follow-up time according to intervention received?

'Yes' or 'Probably Yes': Participants can switch between the treatments. For example, in a chronic condition, the participants may be able to switch between two painkillers.

'No' or 'Probably No': The treatment is one-off, such as surgery.

1.3 If Y/PY to 1.1: Were intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome?

'Yes' or 'Probably Yes': Change in confounders determines the treatment.

'No' or 'Probably No': Change in confounders do not determine the treatment.

1.4 If Y/PY to 1.1: Did the authors use an appropriate analysis method that controlled for all the important confounding domains?

'Yes' or 'Probably Yes': Study authors used an appropriate analysis such as stratification, regression, matching, standardisation and inverse probability weighting for controlling the important confounding domains.

'No' or 'Probably No': An appropriate analysis was not used.

1.5 If Y/PY to 1.1 and 1.4: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?

'Yes' or 'Probably Yes': Valid and reliable measures of the confounding variable were used for confounding.

'No' or 'Probably No': Valid and reliable measures of the confounding variable were not used for confounding, for example, the measures chosen by the study authors were subjective.

1.6 If Y/PY to 1.1: Did the authors control for any postintervention variables that could have been affected by the intervention?

'Yes' or 'Probably Yes': The authors controlled for postintervention variables that could have been affected by the treatment.

'No' or 'Probably No': The authors did not control for postintervention variables that could have been affected by the treatment.

1.7 If Y/PY to 1.1: Did the authors use an appropriate analysis method that controlled for all the important confounding domains and for time-varying confounding?

'Yes' or 'Probably Yes': If switches were related to time-varying confounders, an appropriate method such as inverse probability weighting was used to control for the confounders.

'No' or 'Probably No': An inappropriate analysis was used to control for the confounders.

1.8 If Y/PY to 1.1 and 1.7: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?

Same as for signalling question 1.5.

Bias in selection of participants into the study

Bias in selection of participants into the study relates to whether some eligible participants were selectively excluded from the study prior to study recruitment because of the method of identifying participants for inclusion into the study. For example, some eligible participants may be excluded from the study because the study included only prevalent users of the treatment, which would have excluded participants who stopped the treatment because of side effects. [17]

Some researchers get confused between bias in selection of participants into the study and generalisability. When some participants are excluded from the study because they do not meet the eligibility criteria (the exclusions were before the treatment was started), this is an issue with generalisability of the findings—that is, the findings cannot be generalised to patients who are similar to the excluded participants—rather than a bias in selection of participants into the study. Others get confused between bias in selection of participants into the study and bias due to missing data. When the initial follow-up of participants is missing, this relates to bias in the selection of participants. For example, this can be due to only prevalent users being identified or the exclusion of those who developed the outcomes soon after the treatment was started. When the later follow-up is missing, for example, due to loss to follow-up or death, this results in bias due to missing data.

2.1 Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention?

'Yes' or 'Probably Yes': Some participants are excluded from the study (and effectively from the analysis) based on participant characteristics observed after the start of intervention.

'No' or 'Probably No': Participants are not excluded from the study based on participant characteristics observed after the start of intervention.

2.2 If Y/PY to 2.1: Were the postintervention variables that influenced selection likely to be associated with intervention?

'Yes' or 'Probably Yes': The postintervention variables that influenced selection are likely to be associated with the treatment.

'No' or 'Probably No': The postintervention variables that influenced selection are not associated with the treatment.

2.3 If Y/PY to 2.1 and 2.2: Were the postintervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome?

'Yes' or 'Probably Yes': The postintervention variables that influenced selection can lead to the outcome or can result from the outcome.

'No' or 'Probably No': The postintervention variables that influenced selection are unrelated to the outcome.

2.4 Do start of follow-up and start of intervention coincide for most participants?

'Yes' or 'Probably Yes': Participants were included and followed up from the start of the treatment.

'No' or 'Probably No': Participants who developed the outcome soon after the treatment was started were excluded from the analysis.

2.5 If Y/PY to 2.2 and 2.3, or N/PN to 2.4: Were adjustment techniques used that are likely to correct for the presence of selection biases?

'Yes' or 'Probably Yes': Inverse probability weighting methods may correct for selection biases.

'No' or 'Probably No': Appropriate methods to correct for the presence of selection biases were not used.

Bias in classification of interventions

While the intervention and comparator are classified correctly in randomised controlled trials and prospective cohort studies as the treatment status of the participants are recorded at the time of initiation of the treatment, misclassification of interventions can happen in retrospective studies. The misclassification may or may not be related to the outcome. For example, knowledge of whether a participant developed the outcome at the time of recording the treatment status (i.e., whether a participant received the intervention or comparator) can result in differential misclassification.

3.1 Were intervention groups clearly defined?

'Yes' or 'Probably Yes': The treatment groups are clearly defined to avoid confusion as to whether a participant received the intervention or the comparator.

'No' or 'Probably No': The treatment groups are not clearly defined.

3.2 Was the information used to define intervention groups recorded at the start of the intervention?

'Yes' or 'Probably Yes': The information regarding whether a participant received the intervention or comparator was recorded at the start of the treatment.

'No' or 'Probably No': The information regarding whether a participant received the intervention or comparator was not recorded at the start of the treatment.

3.3 Could classification of intervention status have been affected by knowledge of the outcome or risk of the outcome?

'Yes' or 'Probably Yes': The outcome status was available at the time of classification of the treatment status of a participant.

'No' or 'Probably No': The information regarding whether a participant received the intervention or comparator was recorded at the start of the treatment or the treatment status was classified without the knowledge of the outcome status of the participant.

Bias due to deviations from intended interventions

Intention-to-treat effect

4.1 Were there deviations from the intended intervention beyond what would be expected in usual practice?

'Yes' or 'Probably Yes': There were deviations in the planned treatments beyond usual clinical practice.

'No' or 'Probably No': There were no deviations in the planned treatments, or the deviations were those expected in usual clinical practice, for example, cessation of treatment for toxicity.

4.2 If Y/PY to 4.1: Were these deviations from intended intervention unbalanced between groups and likely to have affected the outcome?

'Yes' or 'Probably Yes': The deviations were not balanced between the groups and would have affected the outcome.

'No' or 'Probably No': The deviations would not have affected the outcome, or the deviations were balanced between the groups.

Per-protocol effect

4.3 Were important co-interventions balanced across intervention groups?

'Yes' or 'Probably Yes': Co-interventions would not have affected the outcome, or the co-interventions were balanced between the groups.

'No' or 'Probably No': Co-interventions were not balanced between the groups and would have affected the outcome.

4.4 Was the intervention implemented successfully for most participants? (This usually refers to one-off treatments implemented by healthcare professionals.)

'Yes' or 'Probably Yes': The intervention was implemented successfully by healthcare professionals for most participants.

'No' or 'Probably No': The intervention was not implemented successfully by healthcare professionals for most participants.

4.5 Did study participants adhere to the assigned intervention regimen? (This usually refers to sustained treatments by the study participants.)

'Yes' or 'Probably Yes': The study participants adhered to the assigned treatment.

'No' or 'Probably No': The study participants did not adhere to the assigned treatment, for example, stopped treatment, crossed-over, or took another treatment not being compared.

4.6 If N/PN to 4.3, 4.4 or 4.5: Was an appropriate analysis used to estimate the effect of starting and adhering to the intervention?

'Yes' or 'Probably Yes': An appropriate analysis such as inverse probability weighting method or instrumental variable method was performed.

'No' or 'Probably No': An appropriate analysis was not performed.

Bias due to missing data

5.1 Were outcome data available for all, or nearly all, participants?

'Yes' or 'Probably Yes': Outcome data were available for all or nearly all participants. We recommend similar strategy as for RoB 2, that is, use sensitivity analysis to answer this question for binary outcomes.

'No' or 'Probably No': Outcome data were not available for all or nearly all participants.

5.2 Were participants excluded due to missing data on intervention status?

'Yes' or 'Probably Yes': Participants were excluded because the treatment status was missing.

'No' or 'Probably No': Treatment status was available for all participants.

5.3 Were participants excluded due to missing data on other variables needed for the analysis?

'Yes' or 'Probably Yes': Participants were excluded because data on the confounders were missing.

'No' or 'Probably No': Participants were not excluded because data on the confounders were missing.

5.4 If PN/N to 5.1, or Y/PY to 5.2 or 5.3: Are the proportion of participants and reasons for missing data similar across interventions?

'Yes' or 'Probably Yes': Similar proportions of participants were excluded from both the treatment groups and the reasons for exclusion are similar in the two groups.

'No' or 'Probably No': The proportion of participants missing was higher in one of the treatment groups or the reasons for being missing were different between the two groups.

5.5 If PN/N to 5.1, or Y/PY to 5.2 or 5.3: Is there evidence that results were robust to the presence of missing data?

'Yes' or 'Probably Yes': A sensitivity analysis shows that the results are robust to different assumptions of treatments, confounders, or outcomes of participants with missing data.

'No' or 'Probably No': No sensitivity analysis was performed, or the sensitivity analysis shows that the results changed depending upon different assumptions of treatments, confounders or outcomes of participants with missing data.

Bias in measurement of outcomes

6.1 Could the outcome measure have been influenced by knowledge of the intervention received?

'Yes' or 'Probably Yes': Subjective outcomes such as HRQoL or pain.

'No' or 'Probably No': Objective outcomes such as all-cause mortality or laboratory test results using an automated analyser.

6.2 Were outcome assessors aware of the intervention received by study participants?

'Yes' or 'Probably Yes': Blinding of outcome assessors was not achieved.

'No' or 'Probably No': Blinding of outcome assessors was achieved.

In participant-reported outcome measures, the outcome assessor is the participant, even if a researcher records the outcome after interviewing the participant.

6.3 Were the methods of outcome assessment comparable across intervention groups?

'Yes' or 'Probably Yes': The follow-up was similar, similar definitions and methods were used to measure the outcome in both the treatment groups.

'No' or 'Probably No': The follow-up was not similar in the treatment groups or different definitions or methods were used to measure the outcome in the two treatment groups.

6.4 Were any systematic errors in measurement of the outcome related to intervention received?

'Yes' or 'Probably Yes': There were systematic errors in measurement of the outcome, and this was related to the intervention.

'No' or 'Probably No': There were no systematic errors in measurement of the outcome, or the systematic errors were not related to the intervention.

Bias in selection of the reported result

7.1 Is the reported effect estimate likely to be selected, on the basis of the results, from multiple outcome measurements within the outcome domain?

'Yes' or 'Probably Yes': There are multiple ways of measuring the outcome and the study author did not report all the outcome measures mentioned in the pre-specified analysis plan.

'No' or 'Probably No': There is only one way of measuring the outcome, the study author reported the outcome measures according to the pre-specified analysis plan, or the study author reported the preferred outcome measure mentioned in the pre-specified systematic review protocol.

7.2 Is the reported effect estimate likely to be selected, on the basis of the results, from multiple analyses of the intervention–outcome relationship?

'Yes' or 'Probably Yes': There are multiple ways of analysing the relationship between the treatment and the outcome, and the study author did not report all the analyses mentioned in the pre-specified analysis plan.

'No' or 'Probably No': There is only one way of analysing the relationship between the treatment and the outcome, the study author reported the outcome data according to the pre-specified analysis plan, or the study author reported the preferred analysis mentioned in the pre-specified systematic review protocol.

7.3 Is the reported effect estimate likely to be selected, on the basis of the results, from different subgroups?

'Yes' or 'Probably Yes': The study author did not report all the subgroup analyses mentioned in the pre-specified analysis plan.

'No' or 'Probably No': The study author reports all the subgroup analyses mentioned in the pre-specified analysis plan.

Conflicts of interest

Explanation of related terms

Before we explain what conflict of interest is, we need to understand several terms.

Quality assurance: planned and systematic actions to ensure that the trial is conducted and the data are collected, recorded and reported in compliance with good clinical practice (GCP) and relevant regulatory requirements. [25] An

example of quality assurance is source data verification in 20% of case report forms, that is, comparing 20% of trial participants' records in the trial database versus original medical records or certified copies.

Quality control systems: operational techniques and activities performed within the quality assurance system to verify that the requirements for quality of the trial-related activities have been fulfilled. In the example for quality assurance, an example of a quality control system might be to collect the number of trial participants and the number of trial participants in whom source data verification was performed (in a reliable way) to ensure that source data verification has been performed in 20% of case report forms.

Sponsor: an individual, company, institution or organisation which takes responsibility for the initiation, management and/or financing of a clinical trial. Key responsibilities include implementing and maintaining quality assurance. [25]

Funder: the funder of a trial provides the funding to perform a trial. The same individual, company, institution or organisation can be the funder and sponsor (this is quite common in industry-funded trials), but the funder and sponsor may also be different (this is quite common in government or charity-funded trials, where the funder is the government or charity and the sponsor is the institution to which the lead researcher belongs).

What is a conflict of interest?

Having understood the similarities and differences between the sponsor and funder, the next step is to understand what a conflict of interest is. From a clinical trial perspective, a conflict of interest can be considered as a risk of influence by an interest other than conducting an unbiased investigation. For example, a funder or a researcher involved in how a trial is designed, conducted or reported stands to gain financially, depending on the published results of the trial. [26]

Why is it necessary to extract data on conflicts of interest?

There is currently evidence that sponsorship of drug and device studies by the manufacturer leads to more favourable results and conclusions than sponsorship by other sources, which cannot be explained by the standard risk of bias assessments. [27] Therefore, it is useful to explore whether the results in trials sponsored by the drug and device studies are different compared to those from trials sponsored by other sources.

How do you deal with multi-arm studies?

An issue closely related to what data should be extracted is how one deals with studies that include more treatment arms than those being compared. In such situations, you can include only the arms relevant for the review. Sometimes, you may find that multiple arms are eligible for a single comparison in the review.

For example, in a comparison of platelet inhibitors versus no treatment for preventing heart attacks in healthy population, you may find a trial that compares aspirin versus clopidogrel versus no treatment in healthy population. In this situation, you must extract data from all the arms in the trial as they are all relevant for the systematic review. Sometimes you may find factorial trials in which the intervention is compared with the comparator in the presence or absence of a co-intervention, for example, in a trial of '*aspirin*' versus '*no treatment*' with and without lifestyle advice. In such situations, we recommend extracting data for the comparison of '*aspirin*' versus '*no treatment*' with lifestyle advice and the comparison of '*aspirin*' versus '*no treatment*' without lifestyle advice separately.

How do you deal with studies that do not report the outcome results?

When the studies do not report any outcome results, the minimum information required are details on the participants, intervention, comparator, outcomes measured but not reported adequately to perform a meta-analysis, risk of bias assessment, and the number of participants included in the intervention and comparator groups when available, or at least the total number of participants in the study (when the number of participants included in each group is not available).

CREATE A DATA EXTRACTION FORM

We recommend creating a data extraction form to allow the systematic collection of data from the studies. Various formats of data extraction forms are available. [28] The advantages and disadvantages of different formats of data extraction forms are discussed in detail by Elamin et al. [28] Currently, electronic forms using spreadsheet software are the most popular format for extracting data in systematic reviews. [29] This is probably because of the almost universal access to spreadsheet software at no extra costs, familiarity with the use of spreadsheets, no requirement for internet access, and the relative ease of customising the form without the requirement for a specialist programmer. These advantages probably outweigh the disadvantages of a spreadsheet-based data extraction form for all but very large and complex reviews.

One row per study or multiple rows per study

In the context of a spreadsheet-based data extraction form, an additional decision to be made is whether to use one row per study or multiple rows per study. In a simple review with only a single comparison and very few outcomes which are well-defined, probably a single row per study approach is sufficient. This allows the compilation of data for comparisons between two reviewers (please see

section 'Resolve differences') and analysis of data with minimal programming skills. However, the number of columns becomes unmanageable with the current risk of bias tools where most of the domains are at the outcome level (although it is likely that there is a considerable overlap of answers among the outcomes within the study). Therefore, a single outcome per row (that is, multiple rows per study) may be preferable for moderate size systematic reviews with multiple outcomes. However, this approach needs more programming skills to compare data since the studies from the two data extractors are not in the same row.

EQUAL-SR software

EQUAL-SR software (available from https://sites.google.com/view/equal-group/home) creates an Excel-based data extraction form which can be customised to a review quickly. In the resulting data extraction form, each outcome has to be extracted in a row. However, once the data extraction is completed, the algorithm allows comparison of data between the reviewers (please see section 'Resolve differences').

PILOT THE DATA EXTRACTION FORM

We recommend that the data extraction form is piloted in two or three studies to see if the required data can be adequately captured.

EXTRACT DATA

We have described the data that should be extracted in the section 'What data should be extracted?' The data are collected in the pre-piloted data extraction form. If per-protocol effect is of relevance to patients, then per-protocol effect should be extracted in the same manner as for intention-to-treat effect. The number of participants included in the analysis will be different when the per-protocol effects are calculated; this may lead to differences in the data extracted for per-protocol effect versus intention-to-treat effect.

RESOLVE DIFFERENCES

How many reviewers should extract the data?

Data extraction should be performed by at least two people independently, as this appears to decrease the errors compared to single data extraction or data extraction by a single reviewer verified by a second reviewer. [30] However, if

resources are limited, it is probably more useful to ensure that double data extraction is performed for at least the data critical for making the conclusions, for example, the primary outcome data compared to participant characteristics, intervention details or other less important outcomes. [30]

Resolving differences between the reviewers

As for study selection, the differences in data extractions should be resolved between the reviewers by discussion between the reviewers and arbitration by a third (senior) reviewer. If the differences persist, we recommend that you record the differences and perform a sensitivity analysis. In other words, perform two analyses, one using the data (pertaining to the unresolved differences) from the first reviewer and another using the corresponding data (pertaining to the unresolved differences) from the second reviewer to assess the impact of the unresolved differences in data.

Comparing data in spreadsheets

If you have used a spreadsheet-based data extraction form where you have extracted one row per study, you can compare the data from the two reviewers and resolve the differences efficiently using formulae. This is best explained by an example. Let us say that the data extracted by the first reviewer are in DataExtractionForm_First sheet and the data extracted by the second reviewer are in DataExtractionForm_Second sheet of a Microsoft Excel file. Assuming that the data extraction forms for the two reviewers are identical (a reasonable assumption), the studies are extracted as one row per study, and the data from different studies are extracted in the corresponding rows by the two reviewers (that is, data from study S001 is extracted in the second row, study S002 is extracted in the third row, and so on by both the reviewers), you can create another sheet, say DataExtractionForm_Comparison to compare the data automatically using simple formulae. Let us say that the cells A1 to Z1 are column headings and the data have been extracted in cells A2 to Z11 by the two reviewers. This means that there are 10 rows of data (that is, data have been extracted from 10 studies) and 26 columns of data (data related to 26 fields have been extracted from each study). You can insert the formula DataExtractionForm_First!A2 = DataExtractionForm_Second!A2 in cell A2 of the DataExtractionForm_Comparison sheet to compare the data in cells A2 of the two forms. If the data in the two forms DataExtractionForm_First and DataExtractionForm_Second) are identical, the value 'TRUE' is shown in cell A2 of the DataExtractionForm_Comparison sheet. Otherwise, the value 'FALSE' is shown in this cell. This formula can be copied and pasted across the entire range to be compared (that is, cells A2 to Z11) of the DataExtractionForm_Comparison sheet. This will allow the comparison of data from each cell in DataExtractionForm_First with the corresponding cell in DataExtractionForm_Second. There is

no need to change the formula in each cell as Microsoft Excel and Google Sheets do this automatically. However, you should note that the comparison is based on the two cells being exactly the same. Even minor differences, such as an extra space or punctuation, can be identified as 'FALSE'. More complex formulae are needed to ignore extra spaces or punctuations.

More processing is required to compare the data from two reviewers when data are extracted as multiple rows per study. If you use the EQUAL-SR software to generate a Microsoft Excel-based data extraction form, each study can have multiple rows corresponding to one row per outcome. Provided that the same study IDs are used by the two reviewers (which can be achieved by harmonisation of study names prior to uploading into EQUAL-SR software), the differences in data extraction between the two reviewers are provided in a table. These differences should be resolved before data analysis.

SOME PRACTICAL GUIDANCE ON CONTACTING AUTHORS

While the major focus of this chapter has been about extracting data from reports, contacting authors is currently recommended to obtain missing data or clarify unclear data. [31] There is paucity of high-quality research in how best to contact authors. Some issues to consider and practical recommendations are included below.

Email versus surface mail (printed and posted letters)

A methodological review on methods to obtain unpublished data from study authors found that response rates were higher and quicker with emails than surface mails. [7] Given the resources required to send out printed letters, we recommend that emails be used as the first contact option. Since repeated use of the same method of contacting study authors did not increase the response rates, [7] contacting study authors by surface mail can be an option if there is no response, depending on the resources and time available to complete the systematic review.

Single contact versus active follow-up

There is currently no evidence that active follow-up of non-responding study authors results in better response rates than a single contact. [7] However, we recommend an active follow-up with at least one more attempt when the study results are likely to change the conclusions of the systematic review conclusions, for example, a low risk of bias study which constitutes a considerable proportion of the participants included in the systematic review with missing information on a critical outcome.

How long should one wait for a response?

The mean time for response from the study authors to emails was about three days and for surface letters was about one month. [7] We suggest waiting at least two weeks for emails and six weeks for surface letters to allow for some variability in response. When the study authors have an out of office reply or have forwarded the response to another researcher, it is reasonable to wait some more time, but unless the study results are likely to change the conclusions of the systematic review conclusions, you should carry on with the analysis.

SUMMARY

In this chapter, we have provided guidance on extracting data for a systematic review.

PRACTICE QUESTIONS

1. Classify the following as different types of outcomes (binary, continuous, ordinal, count, time-to-event outcomes).
   a. Number of patients who died within one year.
   b. Number of tonsillitis episodes in six months.
2. Extract the data in the appropriate format for the specified binary outcomes from the following publications.
   a. www.ncbi.nlm.nih.gov/pubmed/29079379 (deaths, serious adverse events, any adverse events all at maximal follow-up available).
   b. www.ncbi.nlm.nih.gov/pubmed/28898281 (deaths, serious adverse events, any adverse events all at maximal follow-up available).
3. Extract the data in the appropriate format for the specified continuous outcomes from the following publications.
   a. www.ncbi.nlm.nih.gov/pubmed/29419692 (length of hospital stay, visual analogue scale (VAS) scores at 12 months, health-related quality of life (HRQoL) using a validated scale at 12 months).
   b. www.ncbi.nlm.nih.gov/pubmed/29079379 (HRQoL, change in HBA1C%, change in plasma fasting glucose, all at 24 weeks).
4. Extract the data in the appropriate format for the specified count data outcomes from the following publication.
   a. www.ncbi.nlm.nih.gov/pubmed/29183844 (total number of adverse events, total number of serious adverse events, total number of hypoglycaemic episodes).

5. Extract the data in the appropriate format for the specified time-to-event outcomes from the following publication.
    a. www.ncbi.nlm.nih.gov/pubmed/29367198 (time-to-death, time-to-any postoperative pulmonary complication).
6. Assess the risk of bias for the specified outcome in the following two trials. Both the effect of starting the treatment and the effect of adhering to the treatment are of interest. State the sections in the text that helped make this judgement and give the main reason for your judgement.
    a. https://pubmed.ncbi.nlm.nih.gov/30988825/ (HRQoL).
    b. https://pubmed.ncbi.nlm.nih.gov/30912414/ (overall survival).

Answers to the practice questions can be found in the Appendix.

REFERENCES

1. Merriam-Webster Dictionary. Data. www.merriam-webster.com/dictionary/data (accessed on 7 March 2024). 2024.
2. Campbell M. Statistics at Square One. 12th ed. Wiley-Blackwell. 2021.
3. Merriam-Webster Dictionary. Nominal. www.merriam-webster.com/dictionary/nominal (accessed on 11 March 2024). 2024.
4. Merriam-Webster Dictionary. Ordinal. www.merriam-webster.com/dictionary/ordinal (accessed on 11 March 2024). 2024.
5. Naqvi U, Sherman A. Muscle Strength Grading. StatPearls Publishing: Treasure Island (FL). www.ncbi.nlm.nih.gov/books/NBK436008/ (accessed on 11 March 2024). 2023.
6. Swat MJ, Grenon P, Wimalaratne S. ProbOnto. https://sites.google.com/site/probonto/home (accessed on 11 March 2024). 2018.
7. Young T, Hopewell S. Methods for obtaining unpublished data. Cochrane Database Syst Rev. 2011; 2011(11): Mr000027.
8. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. J Clin Epidemiol. 2014; 67(1):56–64.
9. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014; 348: g1687.

10. NIST/SEMATECH e-Handbook of Statistical Methods. www.itl.nist.gov/div898/handbook/ (accessed on 13 March 2024). 2012.
11. Wasserstein RL, Lazar NA. The ASA Statement on *p*-values: context, process, and purpose. The American Statistician. 2016; 70(2):129–33.
12. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med. 1998; 17(24):2815–34.
13. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials. 2007; 8(1):16.
14. Merriam-Webster Dictionary. Bias. www.merriam-webster.com/dictionary/bias (accessed on 14 March 2024). 2023.
15. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019; 366: l4898.
16. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool/current-version-of-rob-2 (accessed on 14 March 2024). 2019.
17. Sterne J, Higgins J, Elbers R, Reeves B, Development group for ROBINS-I. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance. www.riskofbias.info (accessed on 10 September 2023). 2016.
18. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016; 355: i4919.
19. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999; 319(7211):670–4.
20. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Stat Med. 2014; 33(13):2297–340.
21. Mansournia MA, Altman DG. Inverse probability weighting. BMJ. 2016; 352: i189.
22. Beard DJ, Campbell MK, Blazeby JM, Carr AJ, Weijer C, Cuthbertson BH, et al. Placebo comparator group selection and use in surgical trials: the ASPIRE project including expert workshop. Health Technol Assess. 2021; 25(53):1–52.
23. Higgins JP, Li T, Sterne J, RoB 2 working group on crossover trials. Revised Cochrane risk of bias tool for randomized trials (RoB 2). Additional considerations for crossover trials. https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool/rob-2-for-crossover-trials (accessed on 17 March 2024). 2021.

24. Eldridge S, Campbell MK, Campbell MJ, Drahota AK, Giraudeau B, Reeves BC, et al. Revised Cochrane risk of bias tool for randomized trials (RoB 2): additional considerations for cluster-randomized trials (RoB 2 CRT). https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool/rob-2-for-cluster-randomized-trials (accessed on 17 March 2024). 2021.

25. ICH. Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2). https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf (accessed on 9 September 2023). 2016.

26. Østengaard L, Lundh A, Tjørnhøj-Thomsen T, Abdi S, Gelle MHA, Stewart LA, et al. Influence and management of conflicts of interest in randomised clinical trials: qualitative interview study. BMJ. 2020; 371: m3764.

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

28. Elamin MB, Flynn DN, Bassler D, Briel M, Alonso-Coello P, Karanicolas PJ, et al. Choice of data extraction tools for systematic reviews depends on resources and review complexity. J Clin Epidemiol. 2009; 62(5):506–10.

29. Büchter RB, Rombey T, Mathes T, Khalil H, Lunny C, Pollock D, et al. Systematic reviewers used various approaches to data extraction and expressed several research needs: a survey. J Clin Epidemiol. 2023; 159:214–24.

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

31. Higgins J, Lasserson T, Chandler J, Tovey D, Churchill R. Methodological Expectations of Cochrane Intervention Reviews (MECIR). Cochrane: London. https://community.cochrane.org/sites/default/files/uploads/MECIR%20PRINTED%20BOOKLET%20FINAL%20v1.01.pdf (accessed on 7 September 2023). 2016.

6 Analyse the data

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- *Perform a meta-analysis of the data comparing an intervention versus comparator*
- *Summarise data when meta-analysis is not appropriate or possible*

OVERVIEW

In systematic reviews of intervention, meta-analysis can be performed when there are at least two studies with reasonably similar participants, intervention, comparator, and they report one or more similar outcomes.

It is extremely rare for the different studies in a systematic review to have identical criteria for inclusion of participants and for the interventions and comparators to be performed or delivered identically. Therefore, some differences in these aspects can be expected in the different studies. Whether these differences are minor enough to consider them reasonably similar or sufficiently major for them to be considered different is a clinical decision and not a statistical decision (unless advanced statistical methods are used).

Effect measure is a measure used to indicate whether the intervention is beneficial or harmful compared to the comparator for the specified outcome. The effect measures used, and the meta-analysis methods, depend upon the type of outcome data, that is, binary outcomes, continuous outcomes, ordinal outcomes, count outcomes or time-to-event outcomes. The major effect measures used in the meta-analysis are risk ratio, odds ratio and risk difference for binary outcomes, mean difference or standardised mean difference for continuous outcomes, rate ratio (and less commonly rate differences) for count outcomes, and hazard ratio for time-to-event outcomes. From a meta-analysis perspective, ordinal outcomes are converted to binary outcomes or treated as continuous outcomes depending upon the number of levels. Treatment effect is a related term

and indicates whether the intervention is beneficial or harmful compared to the comparator for an outcome using the appropriate effect measure.

The treatment effects from different studies can be meta-analysed using different assumptions about what the studies estimate ('fixed-effect model' and 'random-effects model'). The meta-analysis results are usually presented graphically using 'forest plots'.

Differences in the treatment effects between the studies can be explored by subgroup analysis and metaregression. Any assumptions made in the extraction and analysis should be tested by sensitivity analysis. Reporting biases should also be explored.

When formal meta-analysis of treatment effects is not possible or not appropriate, the information can be summarised using alternative summary measures, such as median and quartiles.

WHAT IS META-ANALYSIS?

You may recollect from Chapter 2 that once the research question is determined, the next steps are to identify suitable methods for study selection, data extraction, analysis and reporting, and to complete protocol registration. In the previous chapter, we described the suitable methods for data extraction.

The next step is to identify the suitable methods for meta-analysis and learn how meta-analysis works in practice.

But what is meta-analysis? As mentioned in Chapter 2, meta-analysis is the use of statistical techniques to combine the results of included studies.

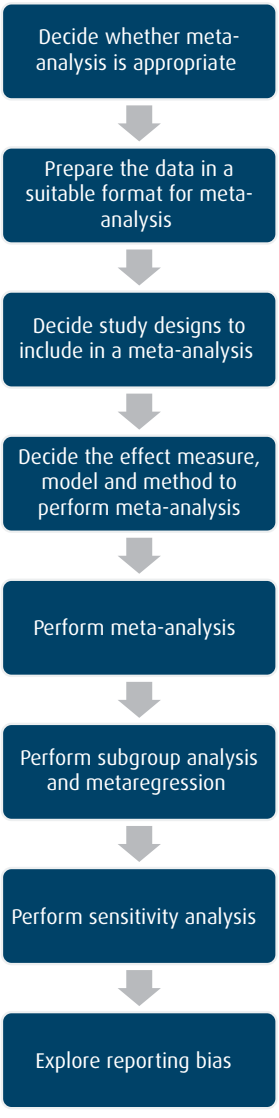
WHAT IS THE ADVANTAGE IN PERFORMING META-ANALYSIS?

By combining information from all relevant studies, meta-analysis can provide more precise estimates of the effects of healthcare than those derived from the individual studies included within a review. [1] By more precise estimates, we mean narrower CI, a concept discussed in Chapter 5. By providing more precise estimates, we get a narrower range within which the population average lies. This decreases the uncertainty in the effectiveness of an intervention versus comparator.

OVERVIEW OF THE STEPS IN META-ANALYSIS

An overview of the steps in meta-analysis is shown in Figure 6.1. Each of the steps is discussed in detail next.

Figure 6.1 Overview of steps in meta-analysis



EFFECT MEASURE AND TREATMENT EFFECT

Before we describe the steps in meta-analysis, it is useful to understand some related terms.

Effect measure is a measure used to indicate whether the intervention is beneficial or harmful compared to the comparator for the specified outcome. The effect measures used, and the meta-analysis methods, depend upon the type of

outcome data, that is, binary outcomes, continuous outcomes, ordinal outcomes, count outcomes and time-to-event outcomes. The major effect measures used in the meta-analysis are risk ratio, odds ratio and risk difference for binary outcomes, mean difference or standardised mean difference for continuous outcomes, rate ratio (and less commonly rate difference) for count outcomes, and hazard ratio for time-to-event outcomes. From a meta-analysis perspective, ordinal outcomes are converted to binary outcomes or treated as continuous outcomes depending upon the number of levels.

Treatment effect is a related term and indicates whether the intervention is beneficial or harmful compared to the comparator for an outcome using the appropriate effect measure.

DECIDE WHETHER META-ANALYSIS IS APPROPRIATE

Before performing meta-analysis, it is necessary to decide whether meta-analysis is appropriate. Meta-analysis is appropriate only when there are at least two studies with reasonably similar participants, intervention and comparator, and which report one or more similar outcomes. But what do we mean by similar in this context?

It is extremely rare for the different studies in a systematic review to have identical criteria for inclusion of participants and for the interventions and comparators to be performed or delivered identically. Therefore, some differences in these aspects can be expected in the different studies. Whether these differences are minor enough to consider them reasonably similar or sufficiently major for them to be considered different is a clinical decision and not a statistical decision (unless advanced methods such as hierarchical meta-analysis [2] beyond the scope of this book are used).

Example 7

For example, let us consider the research question *'Is eating unprocessed fruits better than eating ultra-processed food in decreasing heart attack-related deaths in people at high risk of heart attack?'*.

The PICO in this research question are '*people at high risk of heart attack*', '*eating unprocessed fruits*', '*eating ultra-processed food*' and '*heart attack*', respectively. Let us say we find several studies addressing this research question, some of which used '*apples*' as the intervention and some '*oranges*', and compared these with some ultra-processed food. To answer this research question, it is appropriate to combine the apples and oranges (as interventions) in the same meta-analysis. On the other hand, if the systematic review is performed to answer the research question, *'Which unprocessed fruit is better than eating ultra-processed food in decreasing heart*

attack-related deaths in people at high risk of heart attack?', it is inappropriate to combine the apples and oranges (as interventions) in the same meta-analysis. While it is obvious to everyone that apples and oranges are two different fruits, a difference between interventions may not be obvious in many situations.

Example 8

Let us say that a researcher wants to answer the research question *'Are regular platelet inhibitors beneficial in the elderly population without previous heart attacks or stroke in terms of decreasing all-cause mortality?'* through a systematic review. The PICO in this research question are '*elderly population without previous heart attacks or stroke*', '*platelet inhibitors*', '*no intervention*' (although not explicitly stated in the research question) and '*all-cause mortality*', respectively. Let us say that the review authors found 15 studies of which six used aspirin 75 mg daily, five used clopidogrel 75 mg daily, and four used aspirin 75 mg and clopidogrel 75 mg in combination daily. Can you combine the three sets of studies in a single meta-analysis? It is reasonable to look for previous mechanistic, preclinical and clinical studies to make this decision. For example, if it is known that aspirin 75 mg daily but not clopidogrel 75 mg daily can cause major gastrointestinal bleeding (resulting in increased mortality), it will be unreasonable to combine the studies using aspirin 75 mg daily, clopidogrel 75 mg daily and the combination in a single meta-analysis. However, it is reasonable to perform a meta-analysis of studies using aspirin 75 mg daily, another one of studies using clopidogrel 75 mg daily, and one more of studies using the combination. Therefore, the decision about whether the differences in PICO are large is a clinical decision (supported by previous research when available) and not a statistical decision. Advanced methods such as hierarchical meta-analysis [2] to deal with variations in the interventions or comparators (by comparing the model fit with different assumptions and allowing you to make the decision whether it is reasonable to combine the variations) are beyond the scope of this book. However, it should be noted that even if you used hierarchical meta-analysis to make this decision, it can only provide information on whether the variations have the same or different effects for that outcome.

There are other situations when meta-analysis is not appropriate. When continuous outcomes are not normally distributed, it is not appropriate to perform a meta-analysis unless the data can be 'transformed' to normal distributions by some mathematical functions. However, it should be noted that such 'transformations' to normal distributions are possible only for some data. Furthermore, since only the summary results rather than individual participant data are available from the studies, you have to rely on the study authors to have used the appropriate transformations. Furthermore, to perform the meta-analysis of the transformed data, all the study authors should have used the same transformation.

This is done rarely. Therefore, meta-analysis of non-normally distributed continuous data is usually not appropriate (except when all the study authors have used the same appropriate transformation).

Why is meta-analysis not appropriate in other situations?

Since meta-analysis involves statistical techniques to combine the results of included studies, there must be at least two studies to combine. Therefore, it is not appropriate or possible to perform a meta-analysis of a single study.

The main reason for meta-analysis is to provide more precise estimates of the treatment effects. However, if there are large differences in one or more of PICO, which leads to different results in different studies, and the data are combined, the resulting treatment effect is invariably different from that in the studies. We will not know whether the results obtained from the meta-analysis represent the 'intervention' as described in the first study, the 'intervention' as described in the second study, the 'intervention' as described in either study, or the 'intervention' as described in neither study (for example, average of the doses used in the first and second studies). A similar principle applies to large differences in other aspects of PICO. Therefore, it is meaningless to combine studies which have large differences in one or more of PICO, particularly when the results are different in different studies.

There are currently no appropriate meta-analysis methods to analyse non-normally distributed continuous data. Therefore, when you perform a meta-analysis of continuous outcomes, the implicit assumption is that the data are normally distributed. However, the data (or transformed data) may not be normally distributed. If you perform a meta-analysis in such a situation, the results obtained will not be correct. Therefore, meta-analysis should be avoided in such situations. The distribution (whether it is normal or non-normal) may not be reported in the study. However, if the majority of the studies report the medians rather than the means, this gives a clue that the data are non-normally distributed.

Study-level versus outcome-level decision to perform meta-analysis

The decision to meta-analyse the studies can be at the study level or outcome level. For example, if there is only one study in the systematic review or if the decision not to perform meta-analysis is because of large differences in the participants, intervention or comparator, the decision (not to meta-analyse) is made at the study level. On the other hand, if the decision not to perform meta-analysis is because of differences in an outcome across studies with reasonably similar participants, intervention and comparator or because an outcome is non-normally distributed, the decision (not to meta-analyse) is at the outcome level. For example, there may be differences in the way adverse events were described in the different studies, making meta-analysis inappropriate for the outcome

'*adverse events*', but the studies may have used the same definition of cure of the disease, making meta-analysis appropriate for the outcome '*cure of the disease*'. Another example is where a meta-analysis of intensive care unit stay after major surgery may not be appropriate because it was non-normally distributed but the meta-analysis of the length of hospital stay may be appropriate as the length of hospital stay was normally distributed.

PREPARE THE DATA IN A SUITABLE FORMAT FOR META-ANALYSIS

The data that were extracted in Chapter 5 may have to be processed further before meta-analysis can be performed.

If the data were extracted in the preferred format for binary outcomes and continuous outcomes for RCTs, there is no further processing necessary; the data extracted can be used directly in meta-analysis. However, some processing of data is required before meta-analysis in the following situations.

1. Preferred data were not available for binary and continuous outcomes in RCTs.
2. Data were extracted in the preferred format for count outcomes in RCTs.
3. Data were extracted in the preferred or non-preferred format for time-to-event outcomes.
4. Data were extracted in the preferred format for non-randomised trials for any type of outcome.

The descriptions below are of how to prepare the data in a suitable format that is accepted by most software. Some of these may not be necessary in some statistical software. Since all of the necessary processing can be done by EQUAL-SR software (available from https://sites.google.com/view/equal-group/home), we have only provided details of simple formulae and refer the reader to the source for advanced methods. The details of how to process the data to a format suitable for meta-analysis using EQUAL-SR software is available from instructions in the software.

Binary outcomes

Preferred data were not available for RCTs

When the number of participants with outcome in each group is missing, it can be calculated from the percentage of participants with outcome in each group and the number of participants in each group using the relationship in Equation 1.

Equation 1 Number of participants from percentage

$$\begin{aligned} &\textit{number of participants with outcome} \\ &\quad = \textit{percentage of participants with outcome} \times \textit{number of participants} \end{aligned}$$

This has to be done for each group.

Continuous outcomes

Preferred data were not available for RCTs

Mean is missing

When mean is missing and median is available, the median is accepted as the mean without needing any further processing. This is because median and mean are identical in normal distributions, theoretically. [3] In practice, some minor differences can be expected. However, these differences are generally not considered too large. As meta-analysis of continuous outcomes is appropriate only when the data are normally distributed, the question of median not being representative of the mean does not arise in the context of meta-analysis of continuous outcomes.

Standard deviation is missing

When standard deviation is missing, it can be calculated from various other measures available from the reports.

Standard error in each group

Standard deviation in each group can be calculated from the standard error in each group using the relationship in Equation 2. [4]

Equation 2 Standard deviation from standard error

$$\textit{standard deviation} = \textit{standard error} \times \textit{number of participants}^{0.5}$$

Lower and upper 95% CI in each group

Standard deviation in each group can be calculated from the lower and upper 95% CI in each group using the relationship in Equation 3. [4] While this relationship is correct for sample sizes above 100 participants (in each group), for sample sizes below 60 participants, 3.92 in Equation 3 should be replaced with a different number involving calculation of *t*-statistic from sample size. For sample sizes between 60 and 100, unless information is available from the study regarding the method used for calculating the CI (which is usually not available from the studies), methods used for sample sizes below 60 participants should be used.

Equation 3 Standard deviation from 95% CI

$$Standard\ deviation = \frac{upper\ CI - lower\ CI}{3.92} \times number\ of\ participants^{0.5}$$

Lower quartile, upper quartiles and range in each group

Standard deviation in each group can also be calculated from the lower quartile, upper quartile, range and number of participants using the complex relationship available from Wan et al. [5] Wan et al. also provide the formulae to use when only the lower and upper quartiles are available or when only the range is available. They have also provided an Excel-based calculator to calculate the standard deviation from these measures. [5]

p-value of the comparison

Standard deviation can be calculated from the means in each group and *p*-value of the comparison. [4] Several steps are necessary to perform the calculations: reviewers wanting to create their own Excel calculator can read the detailed instructions available from Higgins et al. [4]

Please note that this is the standard deviation for each of the groups. The standard deviation is identical for the two groups as the calculations are based on the assumption that the standard deviation is identical for the two groups. [4]

Calculations of standard deviation using EQUAL-SR software

All these methods are available from EQUAL-SR software. For details of how to prepare the data in a suitable format for EQUAL-SR software, please see the software.

None of the above measures are available

When none of the above measures are available from the reports, the options are as follows. [4]

1. Exclude the study from meta-analysis.
2. Impute the standard deviation from studies in the same meta-analysis.
3. Impute the standard deviation from studies in a different meta-analysis.

From a practical perspective, excluding the study from meta-analysis solely for the reason of missing standard deviation means that we are excluding some available information. Therefore, we recommend against this. It is easier to impute the standard deviation from studies in the same meta-analysis as there is no need to search for additional information or to decide the appropriateness of the standard deviation from a different meta-analysis. When multiple standard

deviations are available – for example, there are multiple other studies available in the meta-analysis which have different standard deviations – there are various options available. [4] We recommend using the highest standard deviation from the remaining studies, since that approach will decrease the contribution of the study because of the way meta-analysis calculations work. This decrease in the contribution of the study with missing standard deviation appears fair, as the lack of reporting of important basic information may be an indicator of how well the study was conducted. However, when the effect measure used is standardised mean difference, this results in decrease in treatment effect. [4] Therefore, the impact of imputing standard deviation should be assessed as discussed in the section 'Perform sensitivity analysis'.

Count outcomes

For count outcomes, you need to calculate the natural logarithm (ln) and its standard error from the number of events and total number of participants in each group. These can be calculated using the formulae in Equation 4 and Equation 5 based on the formulae from Higgins et al. [4] In RCTs where the intervention and comparators are followed up in the same way, it is reasonable to assume that follow-up is the same in both groups; therefore, ln rate ratio can be calculated even when the mean follow-up is missing by using a simpler formula, shown in Equation 6.

EQUATION 4 ln rate ratio

$$\text{ln rate ratio} = \text{natural logarithm}\left(\frac{\text{number of events}_{Int} \,/\, (\text{number of participants}_{Int} \times \text{mean follow up}_{Int})}{\text{number of events}_{Com} \,/\, (\text{number of participants}_{Com} \times \text{mean follow up}_{Com})}\right)$$

EQUATION 5 Standard error of ln rate ratio

$$\text{standard error of ln rate ratio} = \left(\frac{1}{\text{number of events}_{Int}} + \frac{1}{\text{number of events}_{Com}}\right)^{0.5}$$

EQUATION 6 ln rate ratio (RCTs)

$$\text{ln rate ratio} = \text{natural logarithm}\left(\frac{\text{number of events}_{Int} \,/\, \text{number of participants}_{Int}}{\text{number of events}_{Com} \,/\, \text{number of participants}_{Com}}\right)$$

In Equations 4, 5 and 6, the suffixes 'Int' and 'Com' indicate intervention and comparator.

Time-to-event outcomes

For time-to-event outcomes, you need ln hazard ratio and its standard error for performing the meta-analysis.

Preferred data in RCTs

When hazard ratio and its 95% CI are reported, ln hazard ratio and its standard error can be calculated using the formulae in Equations 7 and 8. [6]

Equation 7 ln hazard ratio from hazard ratio

$$ln\ hazard\ ratio = natural\ logarithm\ (hazard\ ratio)$$

Equation 8 Standard error of ln hazard ratio from 95% CI of hazard ratio

$$standard\ error\ of\ ln\ hazard\ ratio = \frac{\ln\left(Upper\ CI\right) - \ln\left(lower\ CI\right)}{3.92}$$

Preferred data were not available for RCTs

When the 95% CI of the hazard ratio are not reported in the study, the standard error of ln hazard ratio can be calculated from the total number of events in both groups, using the formula in Equation 9, based on the formulae by Parmar et al. (6) and Tierney et al. [7]

Equation 9 Standard error of ln hazard ratio from total number of events in both groups

$$standard\ error\ of\ ln\ hazard\ ratio = \frac{number\ of\ participants_{Int} + number\ of\ participants_{Com}}{\left(\begin{array}{c} total\ number\ of\ observed\ events \times number\ of\ participants_{Int} \\ \times number\ of\ participants_{Com} \end{array}\right)^{0.5}}$$

When the hazard ratio and its 95% CI are not reported but the total number of events in both groups, log-rank *p*-value, and information on whether the proportion without the event was higher in the intervention or comparator (ignoring the statistical significance) are available, it is possible to estimate the hazard ratio and its standard error using complex calculations available from Tierney et al. [7]

Non-randomised studies

Whether the treatment effects are adjusted or unadjusted, they need to be processed further.

Ratios

The common effect measures which are ratios include odds ratio, risk ratio, rate ratio, hazard ratio. Any ratios and 95% CI have to be converted to their corresponding natural logarithms and standard error by the formulae in Equations 10 and 11, based on the formulae from Higgins et al. [4]

Equation 10 ln odds, risk, rate or hazard ratio from odds, risk, rate or hazard ratio

$$ln\ odds, risk, rate, or\ hazard\ ratio = natural\ logarithm\ (odds, risk, rate, or\ hazard\ ratio)$$

Equation 11 Standard error of ln odds, risk, rate or hazard ratio from 95% CI of ln odds, risk, rate or hazard ratio

$$standard\ error\ of\ ln\ odds, risk, rate, or\ hazard\ ratio = \frac{\ln(Upper\ CI) - \ln(lower\ CI)}{3.92}$$

You might notice the similarity between Equation 10 and Equation 7 and that between Equation 11 and Equation 8. This is because of the way 95% CI are calculated for ratios.

Differences

The common effect measures which are differences include mean difference and standardised mean difference. The 95% CI have to be converted to standard error before meta-analysis. This can be done using the formula in Equation 12. [4]

Equation 12 Standard error of mean or standardised mean difference from 95% CI of mean or standardised mean difference

$$standard\ error\ of\ mean\ or\ standaridsed\ mean\ difference = \frac{upper\ CI - lower\ CI}{3.92}$$

Note that there is no logarithmic transformation of the CI for differences as there is for ratios.

Conversion of risk ratios to odds ratios

Sometimes, for binary outcomes, some studies may report odds ratios and others risk ratios. The majority of adjusted treatment effects are likely to be odds ratios,

as logistic regression is the method usually used for adjusting for confounders. Therefore, only the occasional risk ratios may need conversion to odds ratios. This can be done using the relationship between odds ratio, proportion of participants with the event in the comparator, and risk ratio. [8]

Additional processing of continuous outcomes with different units or different scales when cross-walking information is available

Regardless of the data extracted for continuous outcomes, additional work is required for processing of continuous outcomes in which different studies use different units to report the outcome and a conversion factor between units is available.

For example, in a comparison of two surgical treatments, the operation time was reported in minutes in some studies while other studies reported the operation time in hours. Prior to meta-analysis, the units should be converted to the same units (either minutes or hours). Both the mean and standard deviation should be converted using the same conversion factor.

Table 6.1 shows hypothetical data on operation time after the intervention and comparator reported in minutes or hours.

Table 6.2 shows the data after processing so that all the data are in minutes. There is no change in the data for Study_1 as Study_1 reported the data in minutes; however, the data for Study_2 is changed by multiplying the mean and standard deviation in each group by 60.

You can also convert both the studies to hours by dividing the means and standard deviations of Study_1 by 60 if it is desirable to report the operation time in hours in systematic reviews. In such a scenario, you should convert only the means and standard deviations of Study_1, since Study_2 already reports the operation time in hours.

'Mapping' and 'cross-walking' are terms used to convert different measurement scales into a single scale. [9,10] When cross-walk tables are available, you can use the conversion factors to convert the means and standard deviations of each group from one scale to another.

Additional processing of continuous outcomes with different scales

Regardless of the data extracted for continuous outcomes, additional work is required for processing of continuous outcomes in which multiple scales are used to measure an outcome domain. For example, anxiety can be measured in different ways. The Spence Children's Anxiety Scale: Parent Version [11] and Parent Global Impressions-2 anxiety subscale [12] are two scales used for measuring a child's anxiety by a parent. For the Spence Children's Anxiety Scale: Parent Version, lower scores indicate lesser anxiety [11] while for the Parent Global Impressions-2 anxiety subscale, lower scores indicate greater

Table 6.1 Hypothetical data on operation time (as extracted)

Study name	Mean (intervention)	Standard deviation (intervention)	Total number of participants (intervention)	Mean (comparator)	Standard deviation (comparator)	Total number of participants (comparator)	Units
Study_1	75	20	50	60	18	52	Minutes
Study_2	1.5	0.5	40	1.2	0.4	38	Hours

Table 6.2 Hypothetical data on operation time (after processing)

Study name	Mean (intervention)	Standard deviation (intervention)	Total number of participants (intervention)	Mean (comparator)	Standard deviation (comparator)	Total number of participants (comparator)	Units
Study_1	75	20	50	60	18	52	Minutes
Study_2	**90**	**30**	40	**72**	**24**	38	Minutes

anxiety. [12] As will be mentioned in the section 'Effect measure', standardised mean difference can be used to combine multiple scales of a continuous outcome. However, for performing a meta-analysis, the scales should all be in the 'same direction'. In other words, in all studies lower scores indicate better health of participants, or in all studies higher scores indicate better health of participants. There is no definitive rule around what lower scores should indicate and what higher scores should indicate, but we recommend that the review authors analyse and report the results in a way that is intuitive to interpret. For example, it is intuitive to interpret if lower scores of a measure of anxiety indicate better health of participants. For the Spence Children's Anxiety Scale: Parent Version, this is already the case; however, for the Parent Global Impressions-2 anxiety subscale, this is not the case. To convert the Parent Global Impressions-2 anxiety subscale to the 'same direction' as the Spence Children's Anxiety Scale: Parent Version (that is, lower scores indicate lesser anxiety), you have to multiply only the means of Parent Global Impressions-2 anxiety subscale by '–1' in each group. Note that the standard deviations should not be multiplied by '–1'.

Table 6.3 shows hypothetical data on change in anxiety after the intervention and comparator measured using the Spence Children's Anxiety Scale: Parent Version and the Parent Global Impressions-2 anxiety subscale, as extracted.

Table 6.4 shows the data after processing so that lower scores indicate better outcomes for participants for all scales in the analysis. Note that only the means of Study_2 are multiplied by '–1'. This is because Study_1 already uses a scale in which lower scores indicate better health of participants, while Study_2 uses a scale in which higher scores indicate better health of participants.

Additional processing for multi-arm studies

When you find studies with more than two arms and all studies are eligible for inclusion in the same meta-analysis, additional processing is necessary for all types of outcomes regardless of whether the information is from RCTs or from non-randomised studies.

Binary outcomes

The processing for binary outcomes is best explained by an example. Let us say for the systematic review in Example 8, you find a three-armed study which compares aspirin 75 mg, clopidogrel 75 mg, and no treatment. Let us say that 100 participants are included in each group of whom 20 in the aspirin group, 18 in the clopidogrel group and 16 in the 'no treatment' group died after 10 years. Since aspirin and clopidogrel are both platelet inhibitors, both the interventions are eligible for inclusion in the meta-analysis of platelet inhibitors. However, if you include aspirin 75 mg (100 participants) versus no treatment (100 participants) and clopidogrel 75 mg (100 participants) versus no

Table 6.3 Hypothetical data on anxiety (as extracted)

Study name	Mean (intervention)	Standard deviation (intervention)	Total number of participants (intervention)	Mean (comparator)	Standard deviation (comparator)	Total number of participants (comparator)	Scale
Study_1	–4.2	2	60	–0.2	1.2	62	Spence Children's Anxiety Scale: Parent Version
Study_2	1.02	0.6	30	0.84	1.1	25	Parent Global Impressions-2 anxiety subscale

Table 6.4 Hypothetical data on anxiety (after processing)

Study name	Mean (intervention)	Standard deviation (intervention)	Total number of participants (intervention)	Mean (comparator)	Standard deviation (comparator)	Total number of participants (comparator)	Scale
Study_1	–4.2	2	60	–0.2	1.2	62	Spence Children's Anxiety Scale: Parent Version
Study_2	**–1.02**	0.6	30	**–0.84**	1.1	25	Parent Global Impressions-2 anxiety subscale (converted so that lower scores indicate better health of participants)

treatment (100 participants) in the same meta-analysis, there will be 400 participants included in the meta-analysis under this study, which is clearly wrong. On the other hand, if you include only aspirin 75 mg (100 participants) versus no treatment (100 participants) or clopidogrel 75 mg (100 participants) versus no treatment (100 participants) in the meta-analysis, there will be only 200 participants included in the meta-analysis under this study, which is also clearly wrong. Furthermore, you have to choose whether to include aspirin versus no treatment or clopidogrel versus no treatment in the analysis arbitrarily or based on better treatment effect, which again will lead to wrong results of the treatment effect of platelet inhibitors across studies. Therefore, neither of these options (including some participants twice in the same meta-analysis or excluding some participants from the meta-analysis) is acceptable. One acceptable option is to combine the results of both aspirin 75 mg and clopidogrel 75 mg (38 deaths in 200 participants) and compare this with no treatment (16 deaths in 100 participants). This approach is easier to implement than the second acceptable option (described next) and is appropriate if aspirin 75 mg and clopidogrel 75 mg have similar effects on the outcome. The second acceptable option is to split the 'no treatment' into roughly equal parts and include each part in a comparison—aspirin 75 mg (20 deaths in 100 participants) versus no treatment (8 deaths in 50 participants) and clopidogrel 75 mg (18 deaths in 100 participants) versus no treatment (8 deaths in 50 participants)—and include both comparisons in the same meta-analysis. This approach does not assume that aspirin 75 mg and clopidogrel 75 mg have similar effects. In the example provided, the 'no treatment' group has an even number of participants, the number of deaths were also even, and these had to be split into two equal halves. But what do you do when either the number of participants with events or the total number of participants cannot be split equally? Most statistical software allows fractions to be included in the number of participants with events or total number of participants. Therefore, the same approach (split the participants into roughly equal groups) can be followed when the number of participants cannot be split equally between the comparisons. However, some software may not allow fractions to be entered in the number of participants with events or total number of participants. In such cases, you can assign the remaining participants, after splitting the participants into equal parts, to one or more of the comparisons arbitrarily.

Continuous outcomes

For continuous outcomes, when two (or more) groups are combined, you have to calculate the weighted average and the combined standard deviations using the formula available from Higgins et al. [4] When you split the intervention or comparator group into roughly equal parts, the mean is the same. However, the

standard deviation in each of the roughly equal parts may require statistical software that can use iterative methods to calculate values. EQUAL-SR software has this functionality.

Count outcomes

A similar approach as for binary outcomes can be followed.

Time-to-event outcomes

For time-to-event outcomes, usually the hazard ratio and 95% CI are available for each of the arms versus the comparator group. But it is likely that the 95% CI of the hazard ratios for each comparison are calculated based on all the participants in the comparator group. This means that a form of double-counting participants described above (and considered inappropriate) is likely if you calculate the standard error from the 95% CI. Combining the two intervention groups effectively means that you will need to perform a meta-analysis which will suffer from the problem of double-counting participants. Therefore, splitting the groups appears to be the more appropriate option. For splitting the comparator group, you need the hazard ratio (when reported for each comparison) and the number of participants with events in each group. The natural logarithm of the hazard ratio can be calculated from the reported hazard ratio of each intervention group versus comparator group using the formula in Equation 7. However, the standard error of the natural logarithm of hazard ratio of each intervention group versus comparator group has to be calculated by first splitting the events and number of participants in the comparator group into roughly equal parts and then using the formula in Equation 9. When the hazard ratio for each comparison or the number of participants with events in each group is not available but Kaplan-Meier curves are available, you use the methods described by Parmar et al. and Tierney et al. for calculating the ln hazard ratio and its standard error [6,7] with a modification of the number of participants in the group to be split: the number of participants in the group to be split should be split into roughly equal parts.

DECIDE STUDY DESIGNS TO INCLUDE IN A META-ANALYSIS

Once the data are prepared in a suitable format for meta-analysis, you need to decide the study designs to include. Advanced methods for analysis are required to combine evidence from RCTs and non-randomised studies in a single meta-analysis. [13] Therefore, we recommend that meta-analysis of RCTs and non-randomised studies are performed separately (one meta-analysis including only RCTs and another including non-randomised studies).

DECIDE THE EFFECT MEASURE, MODEL AND METHOD TO PERFORM META-ANALYSIS

Effect measure

The next step in the analysis is to decide the effect measure. We described in the section 'Effect measure and treatment effect' that effect measure is a measure used to indicate whether the intervention is beneficial or harmful compared to the comparator for the specified outcome. A summary of the common effect measures and what they mean is shown in Table 6.5.

Table 6.5 Effect measures

Type of outcome	Effect measure	Meaning
Binary	Odds ratio	What are the odds[1] of a participant who received an intervention developing an outcome compared to another receiving the comparator?
Binary	Risk ratio	What is the risk[2] (probability) of a participant who received an intervention developing an outcome compared to another receiving the comparator?
Binary	Risk difference	What is the difference in the risk (probability) of a participant who received an intervention developing an outcome compared to another receiving the comparator?
Continuous	Mean difference	What is the difference in the outcome between a participant who received an intervention developing an outcome compared to another receiving the comparator?
Continuous	Standardised mean difference	This is simply a measure of how well an intervention works versus the comparator. This is difficult to interpret and, therefore, is converted to another measure, such as mean difference, to be meaningful.
Count	Rate ratio	What is the rate[3] of a participant who received an intervention developing an outcome compared to another receiving the comparator?
Count	Rate difference	What is the difference in the rate of a participant who received an intervention developing an outcome compared to another receiving the comparator?
Time-to-event outcome	Hazard ratio	What is the hazard[4] of a participant who received an intervention developing an outcome compared to another receiving the comparator?

Notes:

1. Odds: probability that a particular event will occur against the probability it will not occur. [8]
2. Risk: probability that an event will occur. [8]
3. Rate: probability that an event will occur per unit time.
4. Hazard: probability that an event will occur at a certain instant having not occurred prior to the event. [14]

Why are there so many effect measures and which one should you use? Since the different types of outcomes measure different things, it is reasonable to have different ways of measuring what happened to the outcome in a participant when they received an intervention or comparator. However, there are multiple effect measures for binary, continuous and count outcomes, so further explanations are necessary to understand why there are so many effect measures and which effect measures you should use.

Binary outcomes

There are multiple effect measures for binary outcomes because of issues around consistency across studies, the ease of interpretation and the mathematical properties of the effect measures. [15] In general, relative measures (ratios such as risk ratio or odds ratio) are more consistent across studies, [16] but more difficult to interpret than absolute measures (differences such as risk difference). [15] Even among relative measures, there is lack of consensus around whether odds ratios or risk ratios are better because of the mathematical properties of the two measures and the difficulty in interpreting odds ratios compared to risk ratios. [15] When the events are rare, say less than 1%, logistic regression is one of the methods that provides the least biased estimate of the treatment effect. [15] Logistic regression provides odds ratios. While risk ratios can be calculated using other types of regression, [17] these are not used routinely.

As a practical guidance, we recommend that either risk ratio or odds ratio is chosen when only RCTs are included in the systematic review, and then re-expressed as absolute differences to allow easier interpretation. However, if the events are rare, we recommend logistic regression. Since logistic regression provides odds ratios for rare outcomes, you might consider using odds ratios for all the outcomes when one or more outcomes are rare, to allow uniform reporting.

When the systematic review includes non-randomised studies, odds ratios are the commonly reported measures. Therefore, we recommend using odds ratios as the effect measure and re-expressing the results as absolute differences to allow easier interpretation.

Continuous outcomes

For continuous outcomes, mean difference allows meta-analysis of change from baseline and final scores in the same meta-analysis [15] and is easier to interpret. Therefore, it is the preferred effect measure when the different studies use the same measurement scales or units or can be converted to the same measurement scales or units. However, when the studies use different measurement scales (and it is not possible to convert one scale to another,

which is usually the case), the effect measure used should be standardised mean difference. [15]

Count outcomes

There are no studies comparing the properties of rate ratio versus rate difference for count outcomes. We recommend rate ratio as the preferred measure for count outcomes, as rate ratio is the measure available by Poisson regression used to analyse count outcomes, and similar considerations for binary outcomes may also be applicable for count outcomes.

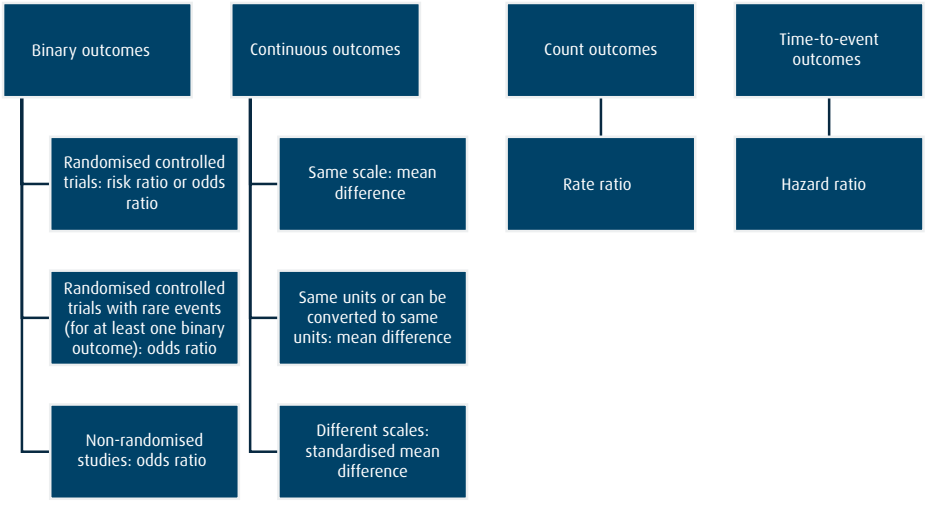
A summary of which effect measures to use for different types of outcomes is available in Figure 6.2.

Conversion between effect measures across different types of outcomes

Sometimes, you may find that the same outcome is reported as a binary outcome in some studies and continuous outcome in others. There are approximate ways of converting between different effect measures even across different types of outcomes. Odds ratios can be converted to standardised mean difference using the relationship described by Chinn. [18] Therefore, to combine binary and continuous outcomes, you can use this approach.

Hazard ratios can be converted to risk ratios using the relationship described by VanderWeele. [19] Odds ratios can also be converted to risk ratios as described by Sinclair et al. [8] Therefore, to combine binary and time-to-event outcomes, you can use this approach.

Figure 6.2 Algorithm for the choice of effect measures to use for different types of outcomes



Model

Having learnt how to choose the effect measure for the different outcomes, the next decision is whether to use the fixed-effect model or random-effects model for meta-analysis.

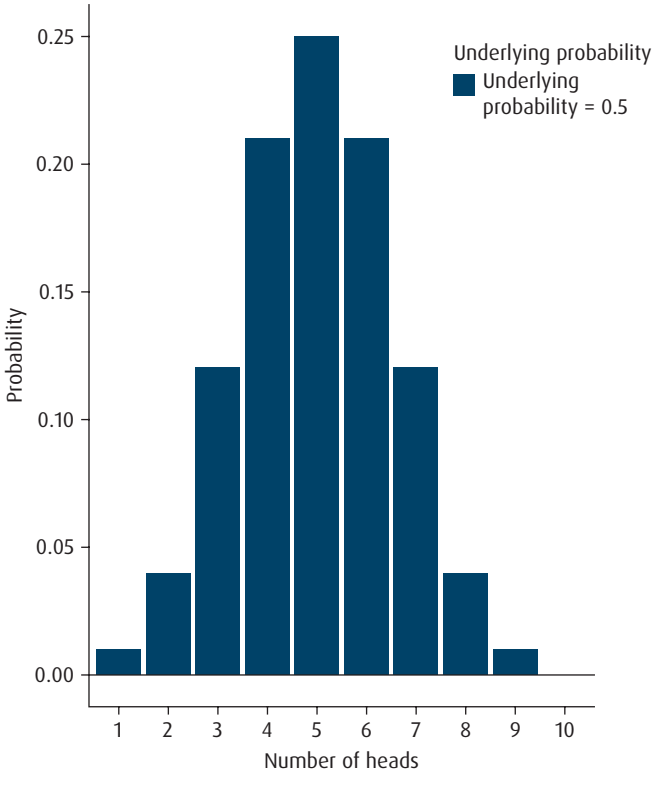
What do we mean by fixed-effect or random-effects model? It is easier to explain the concept with the analogy of a coin toss.

Fixed-effect model

Let us say that you perform 25 experiments of 10 coin-tosses (in each experiment). The probability of obtaining 1 head, 2 heads, 3 heads, and so on, in each of the experiments using a fair coin (a coin in which the probability of obtaining heads and tails is equal—you can expect 5 heads and 5 tails in 10 coin-tosses) using the concept of distributions is shown in Figure 6.3. However, you can see that even though you expect 5 heads in 10 coin-tosses, this happens in only about

Figure 6.3 Probability of obtaining heads in a fair coin-toss experiment

The probability of obtaining heads in experiments of 10 coin-tosses involving a fair coin is shown.



25% of the experiments. In the remaining experiments, you can get a number other than 5 heads. Thus, even though the underlying probability of heads is 0.5 (with an expected number of heads being 5), some experiments can give a value different from 5 because of chance.

Consider that each experiment is equivalent to a clinical study, each coin toss is equivalent to a participant, and 'heads' is equivalent to an event. Applying similar principles as the coin-toss experiment to a clinical study, if the underlying probability of an event in a participant is 0.5 and the clinical study was conducted without bias, you can expect 5 of 10 participants to develop the event. However, you can have clinical studies in which a number different from 5 participants (of 10 participants) develop the event by chance, even in the absence of bias. While the analogy shows the probability of an event in a single group, similar principles apply for differences in an outcome between the intervention and comparator (that is, treatment effects) in clinical studies.

The fixed-effect model assumes that there is a single underlying treatment effect; any differences observed between the different studies is simply by chance. A graphical representation of the fixed-effect model is shown in Figure 6.4. The meta-analysis results represent the single treatment effect that the different studies estimate.

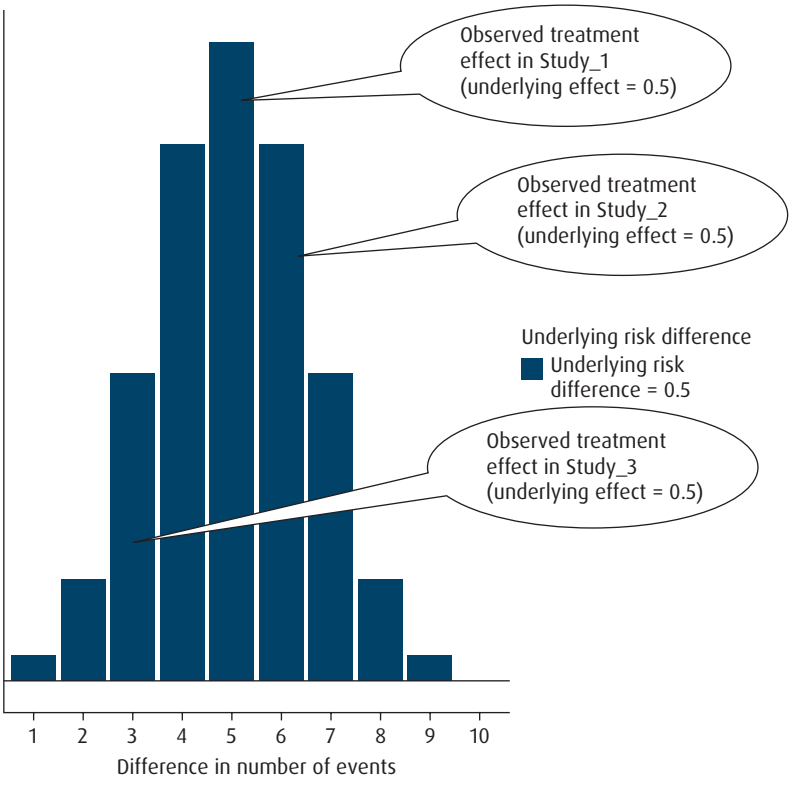
Random-effects model

Let us consider the same scenario (25 experiments of 10 coin-tosses each) but let us toss three different coins. One is a fair coin as earlier, but the other two are biased coins (the probability of obtaining heads and tails is not equal). Let us say that the underlying probabilities of obtaining heads in these biased coins were 0.4 and 0.6. The probability of obtaining 1 head, 2 heads, 3 heads, and so on, in each of the experiments involving the three different coins with underlying probabilities of 0.4, 0.5 and 0.6 using the concept of distributions is shown in Figure 6.5. As in Figure 6.3, you can see that even though you expect 4, 5 and 6 heads in 10 coin-tosses (involving the three different coins), this happens in only about 25% of the experiments. In the remaining experiments, you can get a number other than the expected number of heads.

Using the same analogy as before, each experiment is equivalent to a clinical study, each coin toss equivalent to a participant, and 'heads' equivalent to an event. Applying similar principles as the coin-toss experiment to a clinical study and extrapolating the concept to treatment effects, if the underlying risk difference was 0.4, 0.5, 0.6 (which may be due to differences in one or more of participants, intervention, comparator, outcomes or study design) and the clinical study was conducted without bias, you could expect to see a difference in the number of participants who develop the event of 4, 5 and 6 in 10 participants receiving the intervention rather than the comparator.

Figure 6.4 Fixed-effect model

Each bar represents some clinical studies. The difference in number of events if 10 participants received the intervention rather than the comparator when the underlying risk difference is 0.5 is shown. The meta-analysis results represent the single treatment effect that the different studies estimate.



The random-effects model assumes that the observed treatment differences in the studies are because the different studies are estimating different treatment differences and that these treatment differences follow a normal distribution. A graphical representation of the random-effects model is shown in Figure 6.6. The meta-analysis results represent the average of the treatment effects.

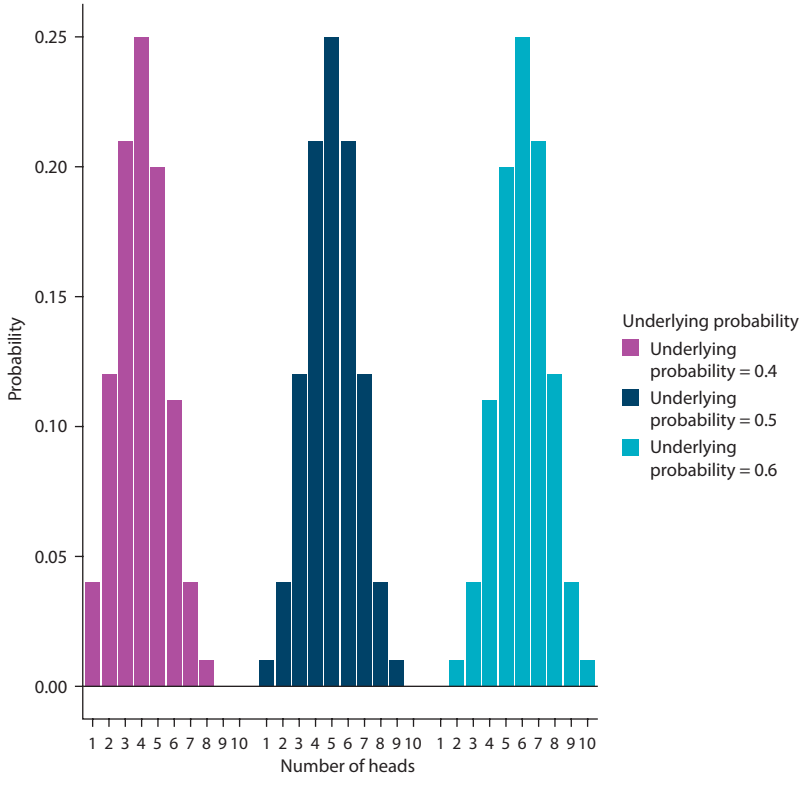
Which model should you choose?

There is no universally agreed way of deciding which model you should choose. [15]

Invariably, there are some differences in the PICO between studies. These differences in PICO may lead to differences in treatment effect between the studies. Differences in treatment effects between studies may also arise due to differences

Figure 6.5 Probability of obtaining heads in three different types of coins

The probability of obtaining heads in experiments of ten coin-tosses involving three different coins with different underlying probabilities of obtaining heads is shown.

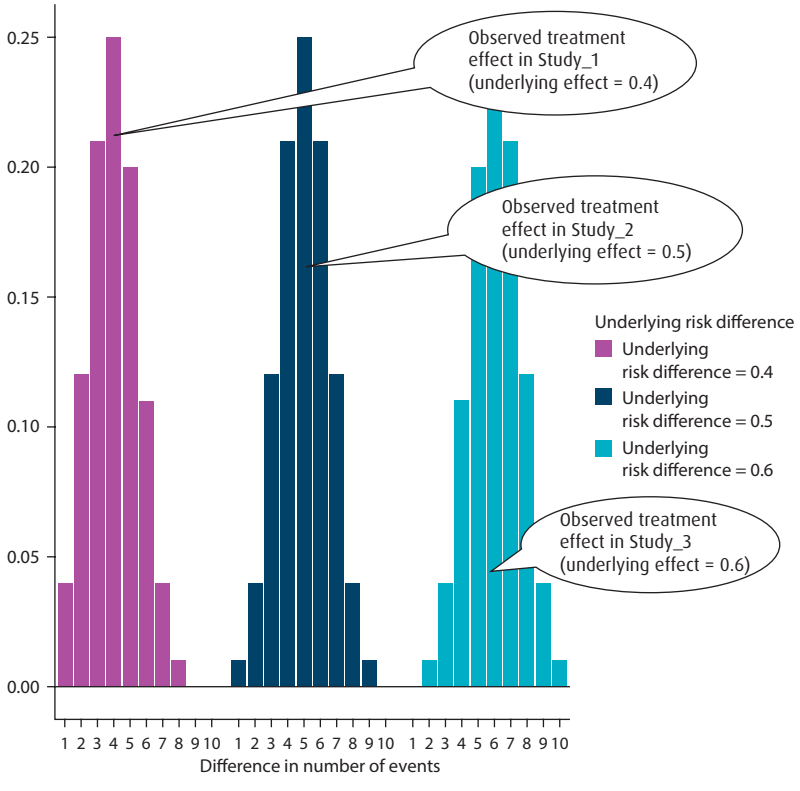


in risk of bias or other variations in study design. It is not appropriate to ignore these differences. In the absence of differences in treatment effects between the studies, the results of the random-effects model are like those obtained by the fixed-effect model. In the presence of differences in treatment effects between the studies, the random-effects model is usually conservative (that is, it results in wider CI). All these points favour using the random-effects model. On the other hand, because of the way the studies are combined to provide a single meta-analysis result (for an outcome), the random-effects model gives greater weight to smaller studies compared to the fixed-effect model: this can be problematic if smaller studies are at higher risk of bias than larger studies and have greater treatment effects ('small-study effects'). [20]

A pragmatic approach could be using the random-effects model as the default model for meta-analysis of studies comparing two treatments along with a

Figure 6.6 Random-effects model

Each bar represents some clinical studies. The difference in number of events if 10 participants received the intervention rather than the comparator when the underlying risk differences 0.4, 0.5 or 0.6 is shown. If the studies are conducted without bias, each study has a different underlying treatment effect; the treatment effects that the studies estimate follow a normal distribution. The meta-analysis results represent the average of the treatment effects.



sensitivity analysis checking whether the results change with the fixed-effect model. You can also check whether there are small-study effects using the funnel plot asymmetry (please see later section 'Explore reporting bias'). In the presence of small-study effects, you can perform a sensitivity analysis excluding small studies at high risk of bias to check whether the meta-analysis results change considerably.

Method

There are many methods for combining the studies. A detailed discussion of the advantages and disadvantages of the different methods is beyond the scope of

this book. Therefore, we limit our discussion to practical guidance of the methods you should use for combining the studies.

Binary outcomes

For binary outcomes, the Mantel-Haenszel method (MH) can be used as default, but if data are sparse (zero event in one or more trials), the Generalised Linear Mixed Model (GLMM) can be used. [21,22]

Other types of outcomes

For continuous, count and time-to-events outcomes, the inverse variance method is used.

PERFORM META-ANALYSIS

The key aspects in the analysis of data are deciding whether to perform a meta-analysis, preparing the data in a format suitable for meta-analysis, and deciding the measure, model and method for meta-analysis, and the study designs to include in the meta-analysis. Once these key decisions are made, the actual step of performing the meta-analysis, that is, combining the treatment effects from the different studies to obtain the treatment effect across studies, is fairly straightforward. If per-protocol effect is of importance to patients, these should be meta-analysed separately from the intention-to-treat effects (one meta-analysis for intention-to-treat effect and another for per-protocol effect).

We recommend statistical software to perform the meta-analysis, although it is technically feasible to perform it manually or using spreadsheets.

The data formats vary between software packages, but if the data is prepared in the formats described in the section 'Prepare the data in a suitable format for meta-analysis', it should be possible to perform the meta-analysis using major software. The guidance for different software is different. The method of performing meta-analysis using the R-based EQUAL-SR software is available from https://sites.google.com/view/equal-group/home.

When the meta-analysis is performed, the results of meta-analysis from most software are available in two formats: treatment effect across studies and forest plots.

Meta-analytical treatment effect

The results are available as the 'summary effect estimate' or simply 'summary estimate' which represents the mean treatment effect (across studies) and the 95% CI which provide information on the 95% probability that the mean treatment effect lies within those intervals. Since we are estimating the treatment

effect across studies based on the treatment effects from individual studies, the meta-analysis results from an intervention review can be referred to as summary effect estimates and 95% CI.

Some software, such as R, also provides the 95% prediction intervals, which give a measure of the differences in treatment effect. [15]

Forest plots

Forest plots are graphical representations of the meta-analysis results. They show the treatment effects of individual studies and the treatment effects across studies (which are provided as default by most statistical software). In general, only one outcome is presented in each forest plot. While it is possible to present multiple outcomes in forest plots, different effect measures and scales for different types of outcomes mean that only one outcome is presented in each forest plot. An example of a forest plot (based on hypothetical data) is shown in Figure 6.7.

What the different aspects of the forest plot represent, and the interpretation of forest plots will be described in Chapter 7.

PERFORM SUBGROUP ANALYSIS AND METAREGRESSION

Heterogeneity

In the context of intervention reviews, 'statistical heterogeneity' or simply 'heterogeneity' is a term used to describe differences in treatment effects observed between the studies. [15] These differences in treatment effect between studies may arise because of differences in PICO (clinical diversity or clinical heterogeneity) and the study design (methodological diversity or methodological heterogeneity). [15] Subgroup analysis and metaregression can help us determine whether the differences in treatment effects are due to differences in PICO or study design.

An overview of the steps required to perform subgroup analysis and metaregression is provided in Figure 6.8.

The first step in both subgroup analysis and metaregression is to identify characteristics (PICO or study design) that could potentially affect the treatment effects. The next step is to split the characteristics into categorical data and quantitative data. Subgroup analysis is possible only for categorical data or quantitative data that can be converted to categorical data. For example, age of the participants is quantitative data but can be converted to categorical data by splitting the age into different categories, for example 18 years or less, 19 to 64 years and 65 years or above. Metaregression is possible for both categorical and quantitative characteristics.

Figure 6.7 Forest plot

Kidney failure

Study	Novel Events	Novel Total	Standard Events	Standard Total	Weight (fixed)	Weight (random)	Odds ratio MH, Fixed + Random, 95% CI
Avgerinos 1994	5	269	5	298	13.5%	14.7%	1.11 [0.32, 3.88]
Choi 2012a	4	189	2	208	5.4%	7.9%	2.23 [0.40, 12.30]
Choi 2012b	2	188	2	132	6.8%	5.9%	0.70 [0.10, 5.03]
Durand 2017	3	209	4	198	11.8%	10.1%	0.71 [0.16, 3.20]
Garcia-Tsao 2003	5	348	3	352	8.5%	11.1%	1.70 [0.40, 7.15]
Inamdar 2016	4	383	8	381	23.1%	15.7%	0.49 [0.15, 1.65]
Manns 2016	6	327	7	360	19.0%	19.0%	0.94 [0.31, 2.83]
Mullen 2002	8	397	4	361	11.9%	15.7%	1.84 [0.55, 6.15]
Total (fixed effect, 95% CI)		**2310**		**2290**	**100.0%**	**—**	**1.06 [0.66, 1.69]**
Total (random effect, 95% CI)					**—**	**100.0%**	**1.05 [0.59, 1.88]**
Prediction interval							**[0.58, 1.92]**

Heterogeneity: $Tau^2 = 0$; $Chi^2 = 3.97$, df = 7 (P = .78); $I^2 = 0\%$
Test for overall effect (fixed effect): Z = 0.24 (P = .81)
Test for overall effect (random effects): $t_7 = 0.22$ (P = .83)

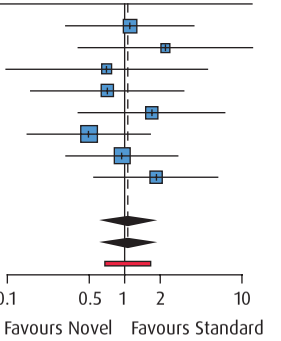
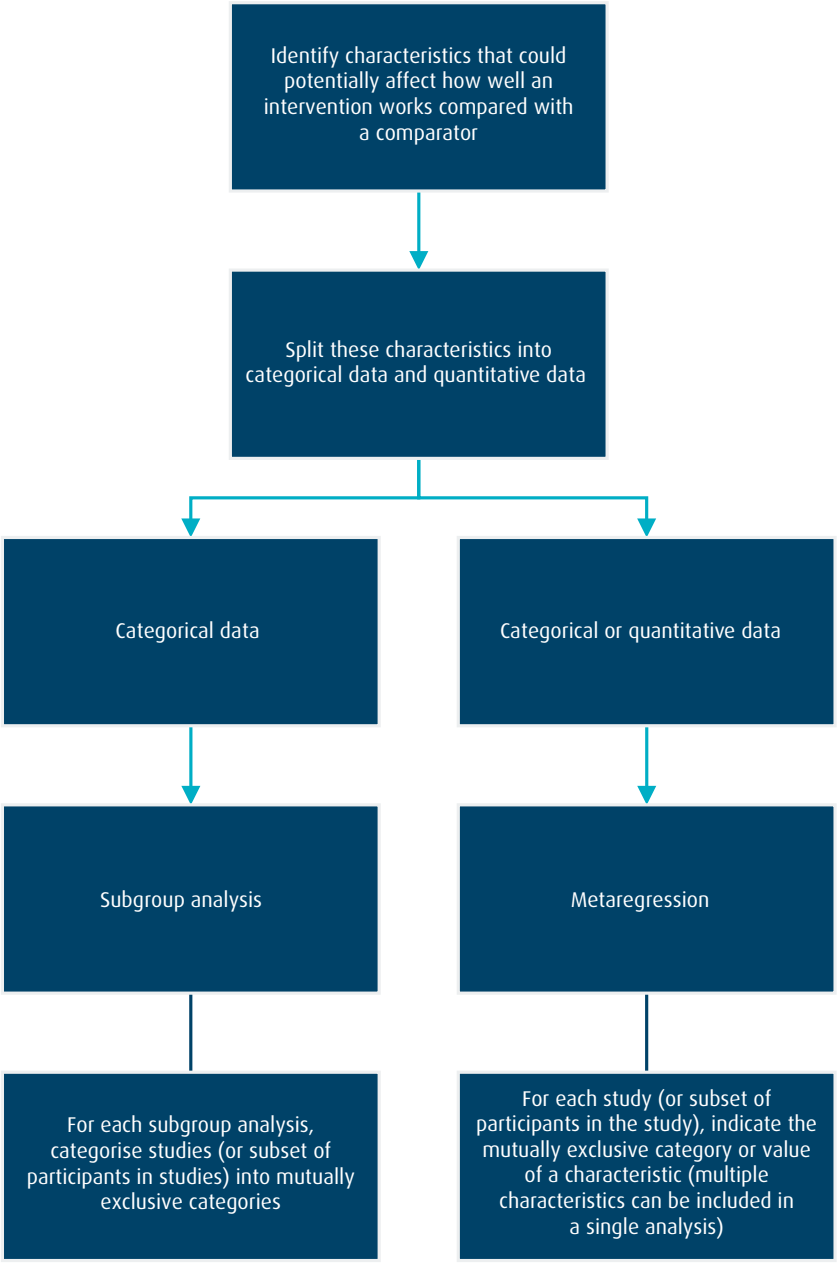


Figure 6.8 Overview of subgroup analysis and metaregression



Subgroup analysis

Subgroup analysis provides the meta-analytical treatment effects of a subset of studies which are similar in terms of a particular characteristic. Usually, in a systematic review, multiple subgroup analysis is planned. This is usually based on PICO (depending on the focus of the research question) and study design. In Example 8, you might want to consider subgroup analyses based on different definitions of elderly population, birth sex and different platelet inhibitors. You might also want to consider subgroup analyses based on the risk of bias in the studies. Therefore, multiple subgroup analyses are possible in a systematic review. A study may be similar to another study in one characteristic and therefore be included in the same subgroup as the second study for the first characteristic, but may be different in another characteristic and therefore belong to a different subgroup as the second study in the second characteristic. For subgroup analysis, it is important to ensure that the categories are mutually exclusive, that is, a study can belong to only one subgroup. Occasionally, you might find a study that reports outcome data for different subgroups separately, that is, the study authors report outcome data on a subset of participants with a certain characteristic, for example birth sex. In such situations, the study can be split into two, one for males and one for females, and can be included in the two subgroups. It is important not to include the same data in two or more subgroups.

There is no guidance available on the number of studies required to perform subgroup analysis. A pragmatic approach will be to perform subgroup analysis when there are at least two studies in each of the categories to allow some estimation of whether the treatment effects are similar within each category but different between the categories.

When the subgroup analysis is performed, you usually obtain the treatment effect across the subset of studies in each subgroup and the treatment effect of all the studies. Usually, you also get a *p*-value of whether the treatment effect is different in the different subgroups.

Metaregression

Metaregression provides information of whether the treatment effects are affected by a characteristic (which could be categorical or quantitative). For quantitative characteristics, the mean value of a normally distributed characteristic and median value of a non-normally distributed characteristic are usually recorded. In Example 8, you might want to consider metaregression based on the average age of the participants included in the study. While a subgroup analysis is not possible (unless the average age can be grouped into categories), the influence of average age of the participants on the treatment effects can be found using metaregression. Theoretically, multiple characteristics that could explain

the treatment effects could be included in the same metaregression, but in practice, there are usually insufficient studies to perform this. Therefore, usually a single factor at a time is included. Metaregression involving less than 10 studies may be misleading. [15]

The data preparation for each software package varies, but in general, along with the usual data required for meta-analysis (in a suitable format), each characteristic is recorded in one column, and you would specify the category or value of the characteristic for each study (or subset of participants in a study for whom outcome data are available separately) in the column corresponding to a characteristic.

When metaregression is performed, a measure of the influence of a characteristic on the treatment effects is obtained along with a *p*-value to indicate whether the influence of the characteristic on the treatment effects was statistically significant. You might also get a *p*-value for the omnibus test which indicates whether at least one of the characteristics influences the treatment effects.

PERFORM SENSITIVITY ANALYSIS

What is sensitivity analysis?

Sensitivity analysis is simply checking whether the results of meta-analysis are altered considerably based on some assumptions made.

Common sensitivity analyses in systematic reviews of intervention

We suggest some common types of sensitivity analyses that review authors can plan in advance.

1. Random-effects meta-analysis versus fixed-effect meta-analysis.
2. For binary outcomes, you can assess the impact of missing data by making different assumptions about the missing data. [23]
3. For continuous outcomes, all calculations and imputations other than the calculations of standard deviation from standard error involve some assumptions. Therefore, it is important to assess the impact of such assumptions by performing a sensitivity analysis.
4. Different correlation coefficients: when you do not find the adjusted analyses but have to adjust the data for clustering or cross-over or similar scenarios, various correlation coefficients may be used to adjust the analysis. It is necessary to check whether the results of meta-analysis are altered considerably based on the correlation coefficient used.
5. In non-randomised studies, you can assess the impact of confounding on the treatment effect by comparing the adjusted versus unadjusted analyses.

Practical implementation of sensitivity analysis

From a practical perspective, one repeats the meta-analysis under different scenarios. The best way to perform this software will depend on the software. In EQUAL-SR software, random-effects versus fixed-effect meta-analysis and sensitivity analyses based on missing data for binary outcomes, and excluding studies based on calculation or imputation of standard deviation (other than calculation from standard error), are done routinely as part of meta-analysis. A separate function is available to test the impact of using different correlation coefficients in studies that do not provide the treatment effects adjusted for clustering or cross-over. The remaining sensitivity analyses are carried out by performing additional analyses with altered data.

Output from sensitivity analyses

The output simply gives you the altered treatment effects across studies using different assumptions.

EXPLORE REPORTING BIAS

What is reporting bias?

Reporting bias arises when reporting of studies or outcomes within studies is based on the results. [24–30] When some studies or outcomes are not available for meta-analysis because the results are not in favour of an intervention, the meta-analysis will not include all the results about an intervention versus comparator and this can lead to meta-analysis resulting in incorrect summary estimates and 95% CI. In other words, you can no longer trust the results.

Various types of reporting bias have been proposed. The most commonly recognised among them are publication bias, selective reporting bias and language bias.

Publication bias is considered to be the publication or non-publication of research findings, depending on the nature and direction of the results. [29] Selective reporting bias can be considered as the selective reporting of some outcomes or analyses but not others, depending on the nature and direction of the results. [29] Language bias can be considered as the publication of research findings in a particular language, depending on the nature and direction of the results. [29]

There is considerable evidence for the existence of publication bias and selective reporting bias. [24–28] There is some uncertainty around whether language bias exists because of inconsistent empirical evidence. [31,32]

How do you assess reporting bias?

There are various ways to assess reporting bias. Each has its own advantages and disadvantages. We describe only the ones that are commonly used and can be practically implemented with reasonable resources.

Searching trial registries

Since registration in a public trials registry became a prerequisite for publication in the International Committee Journal of Medical Editors (ICJME) for trials that commenced recruitment on or after 1 July 2005, [33] you can expect most RCTs in the last decade to be registered in a public trials registry. In Chapter 4, we recommended that trial registries are searched routinely. We also recommended that studies are not excluded solely on the basis of reporting of outcomes. Therefore, at the end of the study selection and data extraction, you can expect to have a list of studies that were conducted on the participants, intervention and comparator relevant to the research question but did not report the outcomes. The publicly available registration entry usually has the outcomes listed. This allows the assessment of whether the outcomes were measured in a study, allowing you to assess whether the predefined outcomes have been reported and seek further information from the study authors if they were not available in the report. However, you need to check the initial document submitted for protocol registration prior to participant enrolment to be sure that an outcome was not measured, as the study authors may have updated the outcomes after results from some participants became available or could be easily predicted.

The disadvantage of this approach is that trial registration is not mandatory for non-randomised studies and protocols of non-randomised studies may not be available. Therefore, it is difficult to assess whether there were reporting biases using this approach.

Searching protocols published in journals

Protocols published in journals may contain more details on the background and methods, including details about outcomes. These are likely to have been identified by the formal searches. However, the same disadvantages as searching trial registries to assess reporting bias exist; in addition, journal publications may not contain the date of enrolment for the participants, making it difficult to assess whether the outcomes were chosen after observing them or could be predicted from other outcomes in some trial participants.

Searching methods sections in the publication of trial reports

Unless the study authors clearly list the deviations from the protocol, it is difficult to assess whether outcomes were not reported because of the nature of the results. Therefore, you cannot rule out that the outcomes changed from the original plan. However, when the authors do not report the outcomes mentioned in the methods, or report them in a manner that cannot be included in the meta-analysis, you can suspect reporting bias strongly.

Funnel plot asymmetry

This involves plotting treatment effects against sample size or a measure of variability in the study. [34–36] Along with visualisation of the funnel plot, various methods are available to test funnel plot asymmetry. [37] For binary outcomes, when the between-study variance (a measure of the variability of treatment effects between studies) is less than 0.1, the Harbord test is recommended; [37] otherwise the Thompson test after arcsine transformation can be used. Egger's test can be used for other outcomes.

Tests for funnel plot asymmetry provide *p*-values to indicate whether there was statistically significant funnel plot asymmetry and are generally recommended only if there are 10 or more trials included in the meta-analysis. [37]

Small-study effects

The main ways of assessing small-study effects are by funnel plot asymmetry and by comparing the results of the random-effects model versus the fixed-effect model, both of which have been discussed previously.

Risk of Bias due to Missing Evidence in a meta-analysis (ROB-ME) tool

You can use the ROB-ME tool to assess the risk of bias due to reporting bias. [38,39] This tool relies on searching trial registries, protocols and methods sections in reports, but also on the interpretation of funnel plots and identification of small-study effects. Therefore, we have provided further details of the ROB-ME tool in Chapter 7.

DEALING WITH SITUATIONS WHEN META-ANALYSIS IS NOT POSSIBLE

Earlier in the chapter, we mentioned that meta-analysis requires a minimum of two studies with reasonably similar participants, intervention and comparator, and which report one or more similar outcomes. What if meta-analysis is not appropriate or possible for one or more reasons? We discuss the different situations and provide some practical guidance on the use of forest plots in these scenarios. Throughout this guidance, when we refer to forest plots, we mean one forest plot for each outcome in the study.

Only one study meets the inclusion criteria for the systematic review

When only one study meets the inclusion criteria for the systematic review, we recommend that you prepare and re-analyse the data using similar effect measures as those you would use if meta-analysis was performed, for example,

present the treatment effect of the intervention versus comparator of a continuous outcome as mean difference and 95% CI. While some methodologists recommend against presenting forest plots when there is only one study, representing the information visually can help with the interpretation even for a single study. Of course, you should not include treatment effects across studies in such forest plots, as the estimate is only from a single study.

Multiple studies are included in the systematic review, but meta-analysis is not possible

When multiple studies are included in the systematic review, meta-analysis may not be possible for the following reasons.

- The differences in PICO are considered too large to perform a meta-analysis.
- Multiple studies are available for reasonably similar PICO, but data are not available in a format suitable for meta-analysis.

When multiple studies are included in the systematic review, but meta-analysis is not appropriate because of large differences in one or more of PICO, and numerical data are available, each subset of studies with similar PICO can be combined in a meta-analysis (as described in the earlier example about combining different platelet inhibitors). This will result in multiple meta-analyses for each outcome. Each comparison (resulting from the large differences in one or more of PICO) can be presented either in separate forest plots or the same forest plot. The advantage of presenting all the data across the different comparisons (whether these are meta-analyses or single study comparisons or a combination of meta-analyses and single study comparisons) in the same forest plot is that the reader can understand the effects of each study and summary effects from each subset of studies included in the comparison from a single plot. In practical terms, while it is not a subgroup analysis, the software is 'tricked' into considering the different subsets of studies as different subgroups. However, by default, most software also presents the treatment effect (combining all the studies) which is inappropriate because of the large differences in one or more of PICO. This can be removed from the forest plot by selecting an option that does not display the summary estimates. You should also provide a label as appropriate. For example, if the large differences are in participants or doses or route of administration or outcomes, you can continue to use the specific intervention and comparator as labels in the forest plot. However, if the large differences are because of different treatments, then a generic name such as 'intervention' or 'comparator' should be used. When presenting multiple comparisons in the same forest plot, we recommend the following order.

1. All comparisons in a certain type of participants are presented first, another type of participants is presented next, and so on.
2. All comparisons within a certain type of participants where intervention is the same but the comparators are different are presented next to each other.
3. Risk of bias in the studies.

In Example 8, let us hypothetically say that among the 15 studies that the review authors found, eight studies included only elderly people at high risk of developing heart attack or stroke, for example because of smoking, hypertension, diabetes and imbalance in lipid profile, and the remaining seven studies included only elderly people at low risk of developing heart attack or stroke. It is reasonable to expect the beneficial effect of platelet inhibitors to be different in people with different risk of developing heart attack or stroke. In this situation, you would report the comparisons in the following order.

1. High-risk population
    a. Aspirin 75 mg daily
    b. Clopidogrel 75 mg daily
    c. Combination of aspirin 75 mg and clopidogrel 75 mg daily
2. Low-risk population
    a. Aspirin 75 mg daily
    b. Clopidogrel 75 mg daily
    c. Combination of aspirin 75 mg and clopidogrel 75 mg daily

Of course, the order of presentation can be different. For example, the combination can be presented first, and the individual drugs presented next. Any logical order can be followed, provided the order is consistent across the different categories of participants. If there were no studies in a particular comparison, say that there were no studies in the combination of aspirin 75 mg daily and clopidogrel 75 mg daily in the low-risk population, this category can be omitted.

When meta-analysis is appropriate for one or more comparisons, but continuous outcome data are not available in the format required for meta-analysis, we recommend the following.

Medians are available

This is usually because study authors consider the distribution of the continuous outcome as non-normal. Nevertheless, some study authors may report the mean and standard deviation in each group, sufficient to perform the meta-analysis. If there is strong evidence that a continuous outcome is non-normally distributed,

it is inappropriate to perform a meta-analysis even when sufficient data are available from some studies.

Therefore, when only medians are available, or when a mixture of medians and means is available and the data are non-normally distributed, the review authors should tabulate the medians (and means, if only means were reported in some studies) as a minimum. In terms of the order of comparisons in a table, the same principles as described for the order of comparisons in forest plots can be followed.

In Example 8, let us hypothetically say that, of the eight studies in the high-risk population, the risk of bias in studies that used aspirin 75 mg daily as the platelet inhibitor was 'low risk' in one study (study_1), 'some concerns' in two studies (study_2 and study_3), and 'high risk' in one study (study_4); the risk of bias in studies that used clopidogrel 75 mg daily as the platelet inhibitor was 'low risk' in two studies (study_5 and study_6), and 'high risk' in one study (study_7); and the risk of bias in the only study that used the combination (study_8) was 'low risk'. Let us also say that of the seven studies in the low-risk population, the risk of bias in studies that used aspirin 75 mg daily as the platelet inhibitor was 'low risk' in three studies (study_9, study_10 and study_11) and 'high risk' in two studies (study_12 and study_13); the risk of bias in studies that used clopidogrel 75 mg daily as the platelet inhibitor was 'low risk' in both studies in this category (study_14 and study_15); and there were no studies that used the combination of drugs in the low-risk population.

The order in the table will be as follows.

1. High-risk population
- **a.** Aspirin 75 mg daily
  - **i.** Low risk
    - study_1
  - **ii.** Some concerns
    - study_2
    - study_3
  - **iii.** High risk
    - study_4
- **b.** Clopidogrel 75 mg daily
  - **i.** Low risk
    - study_5
    - study_6
  - **ii.** High risk
    - study_7
- **c.** Combination of aspirin 75 mg and clopidogrel 75 mg daily
  - **i.** Low risk
    - study_8

2. Low-risk population
    a. Aspirin 75 mg daily
        i. Low risk
            - study_9
            - study_10
            - study_11
        ii. High risk
            - study_12
            - study_13
    b. Clopidogrel 75 mg daily
        i. Low risk
            - study_14
            - study_15

In addition, you might use a box-whisker plot to visualise the data. [40]

p-values are available

The *p*-values from each study in the comparison should be summarised in a table as a minimum. When medians are available, the *p*-values can be provided alongside the median (and means, if only means were reported in some studies). Otherwise, they can be provided alongside a narrative description of the difference between intervention and comparator provided by the primary study author. The order in the table will be guided by the same principles as under the section 'Medians are available'.

Vote counting method and sign test

In addition to presenting the information on median and *p*-values in a table and box-whisker plot to summarise the medians, we recommend that the authors perform the vote counting method to determine the number of studies in which the treatment effect was in favour of the intervention and the number of studies in which the treatment effect was in favour of the comparator. *p*-values should not be considered in deciding whether the direction of effect was in favour of intervention or comparator. [40] Let us say that shorter lengths of intensive care unit stay indicate quicker recovery of patient. If the median length of intensive care unit stay was two days in the intervention group and one day in the comparator group but the difference was not statistically significant in a study, this study should still be counted as favouring the 'comparator'.

We recommend that the vote counting method be accompanied by a sign test. The sign test is used to test whether there was any evidence of treatment effect. [40] You can also provide the proportion of studies that favoured the intervention versus comparator, along with the CI.

The vote counting method may be accompanied by a harvest plot to indicate the direction of effect in each study. [40] We recommend that the harvest plot has information on the risk of bias (shown, for example, by different colours) to provide additional information on the relationship between the risk of bias and the direction of effect.

Vote counting, sign test and harvest plot are available in EQUAL-SR software.

SUMMARY

In this chapter, we have provided guidance on performing meta-analysis in a systematic review using appropriate effect measures and provided alternatives when meta-analysis is either inappropriate or not possible.

PRACTICE QUESTIONS

1. Two researchers want to compare the effect of two forms of surgeries (surgery_1 versus surgery_2) on short-term mortality, long-term mortality, length of hospital stay, and the number of complications during the immediate post-operative period and 30 days following surgery in patients with periampullary cancer. They have identified the data in Tables 6.6–6.9 from the studies.
   a. Short-term mortality
   b. Long-term mortality
   c. Length of hospital stay
   d. Number of complications

   Perform a meta-analysis of the four outcomes using appropriate effect measures, models and methods.
2. In the same study as Question 1, the study authors also recorded whether the origin of the periampullary cancer originated from the pancreas or from the ampulla. Table 6.10 shows which type of cancer different studies included.

   Perform a subgroup analysis and metaregression to find if the treatment effect (that is, the effect of surgery_1 versus surgery_2) was different in the different cancer types.
3. Perform a sensitivity analysis excluding the studies in which either the mean or the standard deviation or both were imputed from other measures or other studies.

4. What is reporting bias and why is it important?
5. Create funnel plots and perform tests for funnel plot asymmetry for each of the outcomes in Question 1. If it is inappropriate to create funnel plots or perform tests for funnel plot asymmetry for some outcomes, state the reason.

Answers to the practice questions can be found in the Appendix.

Table 6.6 Data on short-term mortality

Study	Number of patients with events in surgery_1 group	Total number of patients in surgery_1 group	Number of patients with events in surgery_2 group	Total number of patients in surgery_2 group
Study 1	1	261	2	258
Study 2	2	123	0	132
Study 3	2	248	1	254
Study 4	2	170	0	180
Study 5	3	181	1	178
Study 6	0	260	5	257
Study 7	2	129	2	139
Study 8	0	102	0	107
Study 9	0	86	5	89
Study 10	0	172	4	179
Study 11	0	298	1	292
Study 12	0	209	4	202

Table 6.7 Data on long-term mortality

Study	ln hazard ratio (HR)	Standard error of ln HR
Study 1	2.742	1.463
Study 3	2.861	1.364
Study 6	0.990	0.528
Study 7	2.185	0.151
Study 9	2.576	1.648

Table 6.8 Data on length of hospital stay (days)

Study	Mean hospital stay in surgery_1 group	Standard deviation of hospital stay in surgery_1 group	Total number of patients in surgery_1 group	Mean hospital stay in surgery_2 group	Standard deviation of hospital stay in surgery_2 group	Total number of patients in surgery_2 group
Study 1	12.3	11.8	261	15.8	13.5	258
Study 2	15.1	7	123	16	11.5	132
Study 3	12.2	15.4	248	12.1	13.9	254
Study 4	14.3	13.8	170	10.2	8.9	180
Study 5*	17.7	–	181	16.5	–	178
Study 6	16.5	10.3	260	11.6	9.7	257
Study 7†	15.4	15.6	129	9.3	12.6	139
Study 8*	17.2	–	102	12.8	–	107
Study 9	15.6	10.7	86	16.1	14.7	89
Study 10†	12.5	15.4	172	10.6	15.9	179
Study 11	11.2	8	298	13	13.4	292
Study 12	12.4	8.7	209	16.8	15.8	202

Notes:

* Standard deviation could not be imputed from any information available from the reports.

† Mean and standard deviation imputed from median and *p*-value.

Table 6.9 Data on number of complications

Study	ln hazard ratio (HR)	Standard error of ln HR
Study 1	–0.717	1.331
Study 3	2.929	0.939
Study 5	–1.820	1.705
Study 6	1.867	1.563
Study 8	2.756	0.308
Study 9	–1.851	0.072
Study 11	–1.546	0.208

Table 6.10 Data on type of cancer included in different studies

Study name	Cancer type
Study 1	Pancreatic cancer
Study 2	Pancreatic cancer
Study 3	Pancreatic cancer
Study 4	Pancreatic cancer
Study 5	Pancreatic cancer
Study 6	Ampullary cancer
Study 7	Pancreatic cancer
Study 8	Pancreatic cancer
Study 9	Ampullary cancer
Study 10	Pancreatic cancer
Study 11	Pancreatic cancer
Study 12	Pancreatic cancer

REFERENCES

1. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, 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(7): e1000100.
2. Fernández-Castilla B, Jamshidi L, Declercq L, Beretvas SN, Onghena P, Van den Noortgate W. The application of meta-analytic (multi-level) models with multiple random effects: a systematic review. Behavior Research Methods. 2020; 52(5):2031–52.
3. Campbell M. Statistics at Square One. 12th ed. Wiley-Blackwell. 2021.
4. Higgins JP, Li T, Deeks JJ. Chapter 6: Choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, Welch VA (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 6.4 (updated October 2023). Cochrane, 2023. Available from: www.training.cochrane.org/handbook. (accessed on 22 March 2024).
5. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Method. 2014; 14(1):135.

6. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med. 1998; 17(24):2815–34.
7. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials. 2007; 8(1):16.
8. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. J Clin Epidemiol. 1994; 47(8):881–9.
9. Heng M, Tang X, Schalet BD, Collins AK, Chen AF, Melnic CM, et al. Can the Knee Outcome and Osteoarthritis Score (KOOS) function subscale be linked to the PROMIS physical function to crosswalk equivalent scores? Clin Orthop Relat Res. 2021; 479(12):2653–64.
10. Putman S, Preda C, Girard J, Duhamel A, Migaud H. Mapping and crosswalk of the Oxford Hip Score and different versions of the Hip Disability and Osteoarthritis Outcome Score. Clin Orthop Relat Res. 2021; 479(7):1534–44.
11. Nauta MH, Scholing A, Rapee RM, Abbott M, Spence SH, Waters A. A parent-report measure of children's anxiety: psychometric properties and comparison with child-report in a clinic and normal sample. Behav Res Ther. 2004; 42(7):813–39.
12. Adams JB, Audhya T, Geis E, Gehn E, Fimbres V, Pollard EL, et al. Comprehensive nutritional and dietary intervention for Autism Spectrum Disorder: a randomized, controlled 12-month trial. Nutrients. 2018; 10(3):369.
13. Sarri G, Patorno E, Yuan H, Guo J, Bennett D, Wen X, et al. Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making. BMJ Evidence-Based Medicine. 2022; 27(2):109–19.
14. Bewick V, Cheek L, Ball J. Statistics review 12: survival analysis. Crit Care. 2004; 8(5):389–94.
15. Deeks JJ, Higgins JP, Altman DG. Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, Welch VA (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 6.4 (updated October 2023). Cochrane, 2023. Available from: www.training.cochrane.org/handbook. (accessed on 25 March 2024).
16. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med. 2002; 21(11):1575–600.
17. Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 2003; 3:21.

18. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med. 2000; 19(22):3127–31.

19. VanderWeele TJ. Optimal approximate conversions of odds ratios and hazard ratios to risk ratios. Biometrics. 2020; 76(3):746–52.

20. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000; 53(11):1119–29.

21. Liu D. Meta-Analysis of Rare Events. Wiley StatsRef: Statistics Reference Online. 2019. pp. 1–7.

22. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat Med. 2010; 29(29):3046–67.

23. Higgins JP, White IR, Wood AM. Imputation methods for missing outcome data in meta-analysis of clinical trials. Clin Trials. 2008; 5(3):225–39.

24. Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias: an updated review. PLoS One. 2013; 8(7): e66844.

25. Dwan K, Altman DG, Clarke M, Gamble C, Higgins JPT, Sterne JAC, et al. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. PLoS Med. 2014; 11(6): e1001666.

26. 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; 2009(1): Mr000006.

27. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One. 2008; 3(8): e3081.

28. 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. 2010; 340: c365.

29. Boutron I, Page M, Higgins J, Altman D, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, Welch VA (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 6.4 (updated October 2023). Cochrane, 2023. Available from: www.training.cochrane.org/handbook. (accessed on 26 March 2024).

30. Ioannidis JPA. Inverse publication reporting bias favouring null, negative results. BMJ Evidence-Based Medicine. 2024; 29(1):6–9.

31. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet. 1997; 350(9074):326–9.

32. Jüni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002; 31(1):115–23.
33. ICJME. Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. www.icmje.org/news-and-editorials/clin_trial_sep2004.pdf (accessed on 26 March 2024). 2004.
34. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997; 315(7109):629–34.
35. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Stat Med. 2006; 25(20):3443–57.
36. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol. 2001; 54(10):1046–55.
37. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011; 343: d4002.
38. Page MJ, Sterne JAC, Boutron I, Hróbjartsson A, Kirkham JJ, Li T, et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. BMJ. 2023; 383: e076754.
39. Page MJ, Sterne JAC, Boutron I, Hróbjartsson A, Kirkham JJ, Li T, et al. ROB-ME: A Tool for Assessing Risk of Bias Due to Missing Evidence in Systematic Reviews with Meta-Analysis. www.riskofbias.info/welcome/rob-me-tool (accessed on 28 March 2024). 2023.
40. McKenzie J, Brennan S. Chapter 12: Synthesizing and presenting findings using other methods. In: Higgins JPT, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, Welch VA (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 6.4 (updated October 2023). Cochrane, 2023. Available from: www.training.cochrane.org/handbook. (accessed on 20 March 2024).

7 Interpret the results

LEARNING OBJECTIVE

After studying this chapter, you should be able to:

- *Interpret the results of meta-analysis*

OVERVIEW

In systematic reviews of intervention, statistical interpretation of the results of the meta-analysis is largely based on risk of bias in the studies and the meta-analytical summary estimates and the 95% CI; although other levels of CI can be used, they rarely are. If the 95% CI do not include 'null' treatment effect, there is only a 5% probability that these intervals excluded the 'null' treatment effect by chance if the average treatment effect in the population (from which the trial participants were selected) was null (in the absence of bias). The interpretation of results is greatly enhanced using forest plots.

From an assessment of the risk of bias in the studies, an assessment should be made about whether any observed differences in the outcome between the intervention and comparator could be explained by other factors, such as systematic differences in the participants, how often they made planned clinic visits, received planned additional tests or interventions, how the outcome was measured in the two groups, and whether some participants who developed 'undesirable outcomes' were excluded from the analysis. Estimates are available about the deviations from the true average population treatment effect that each of the biases causes, although there is uncertainty around these deviations. Methods to adjust the treatment effects for these biases are available but are not routinely used.

Additional statistical measures such as I^2, which provides a measure of the statistical heterogeneity (differences in the results across the studies), can help with the interpretation. Meta-analytical treatment effect estimates in the presence of considerable statistical heterogeneity have no clinical application.

Subgroup analysis and metaregression might help but can lead to spurious conclusions. The Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) tool can help in deciding whether you should use the overall meta-analysis results or subgroup meta-analysis results for clinical interpretation of results. Statistical heterogeneity which cannot be explained by clinical differences in PICO, methodological differences in the study design, or risk of bias decreases the confidence in results.

Sensitivity analyses can help with testing whether any assumptions made while performing the meta-analysis are 'robust' or 'sensitive' to the assumptions.

Reporting biases should be considered while interpreting the results. The ROB-ME tool can help with the assessment of reporting biases.

Interpretation of conflicts of interest is likely to be guided by the Tool for Addressing Conflicts of Interest in Trials (TACIT) tool when it becomes available. Fraud in clinical trials continues to be an unresolved challenge with no clear methods for its detection, or guidance, available.

Summary estimates and CI alone do not provide information on whether this difference is clinically important, particularly in the context of continuous outcomes, where the observed differences may not be important. The interpretation of continuous outcomes should be made in the context of 'minimal important differences' in the outcomes. Other aspects, such as clinical applicability and number of participants included in the studies, should be considered while deciding whether the intervention improves outcomes compared to the comparator.

Confidence in the results can be estimated using GRADE (Grading of Recommendations, Assessment, Development and Evaluations) methodology which can help with arriving at conclusions.

INTRODUCTION

You may recollect from Chapter 2 that once the research question is determined, the next steps are to identify suitable methods for study selection, data extraction, analysis and reporting, and to complete protocol registration. In the previous chapter, we described the suitable methods for data analysis.

The next step is to describe the suitable methods for interpretation of the results of meta-analysis.

WHY SHOULD YOU USE SUITABLE METHODS FOR INTERPRETATION?

At the risk of stating the obvious, unless suitable methods for interpretation are used in a meta-analysis, you might arrive at wrong conclusions, which can have knock-on effects resulting in incorrect clinical practice guidelines, unnecessary or harmful treatments for patients and unnecessary costs to the healthcare funder. Therefore, you should use suitable methods for interpretation of the analysis results.

OVERVIEW

An overview of the process in interpreting the results is shown in Figure 7.1.

STATISTICAL INTERPRETATION OF THE MAIN RESULTS OF META-ANALYSIS

While the main reason for meta-analysis (as compared to a systematic review without meta-analysis) is to obtain more precise information about the treatment effect, meta-analysis and the resulting forest plots provide a lot more information to help decide whether an intervention improves an outcome compared to a comparator. Therefore, we provide an interpretation of what the results of the meta-analysis mean from a statistical perspective and what the results mean from a clinical perspective for all the information available from the meta-analysis. Since all the information is available from the forest plots (depending on which software you used and what options you chose to perform the meta-analysis), we start with the information found in forest plots.

What does the forest plot show?

The information found in the forest plot varies in different software, models and methods used for meta-analysis. Figure 7.2 shows the information found in a forest plot created by EQUAL-SR software (available from https://sites.google.com/view/equal-group/home).

Treatment effect across studies

A summary estimate represents the mean treatment effect across studies. The interpretation of 95% CI is provided in the section 'Line of no effect'. The treatment effect across studies is represented as a diamond in the forest plot.

Figure 7.1 Overview of interpreting the results

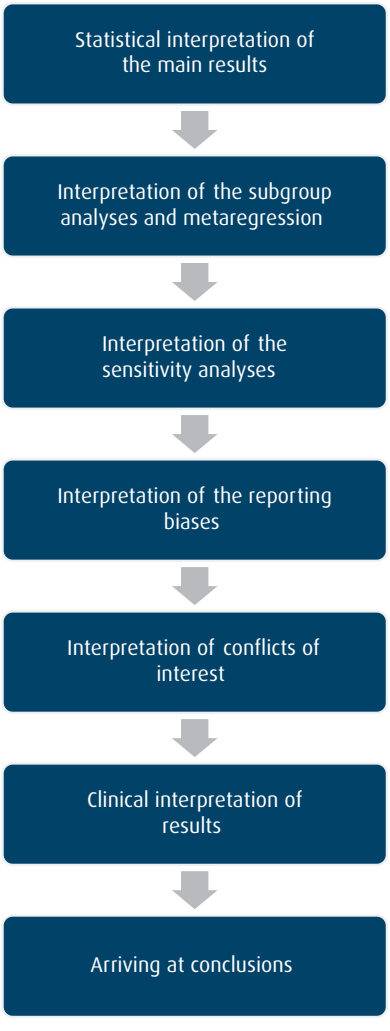


Figure 7.3 shows the summary estimates and 95% CI of treatment effects obtained in fixed-effect and random-effects models. The summary estimate is represented by the peak of the diamond and the 95% CI are represented by the left and right edges of the diamond. The width of the diamond represents the width of the 95% CI. Narrow CI means that there is less uncertainty about the mean treatment effect across studies ('precise estimates') and wide CI mean that there is greater uncertainty about the mean treatment effect across studies ('imprecise estimates'). Further interpretation of the treatment effect across studies is provided in the section 'Line of no effect'.

Figure 7.2 What does a forest plot show?

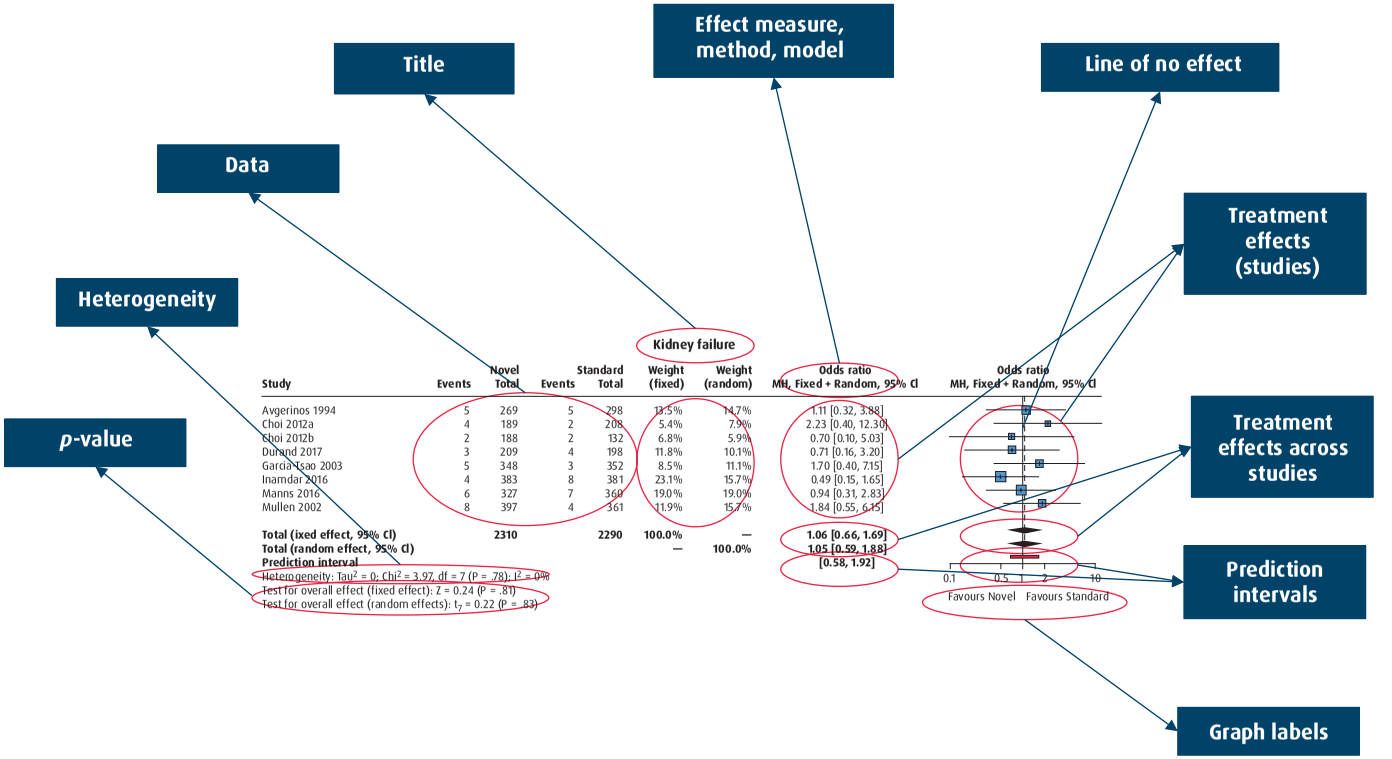
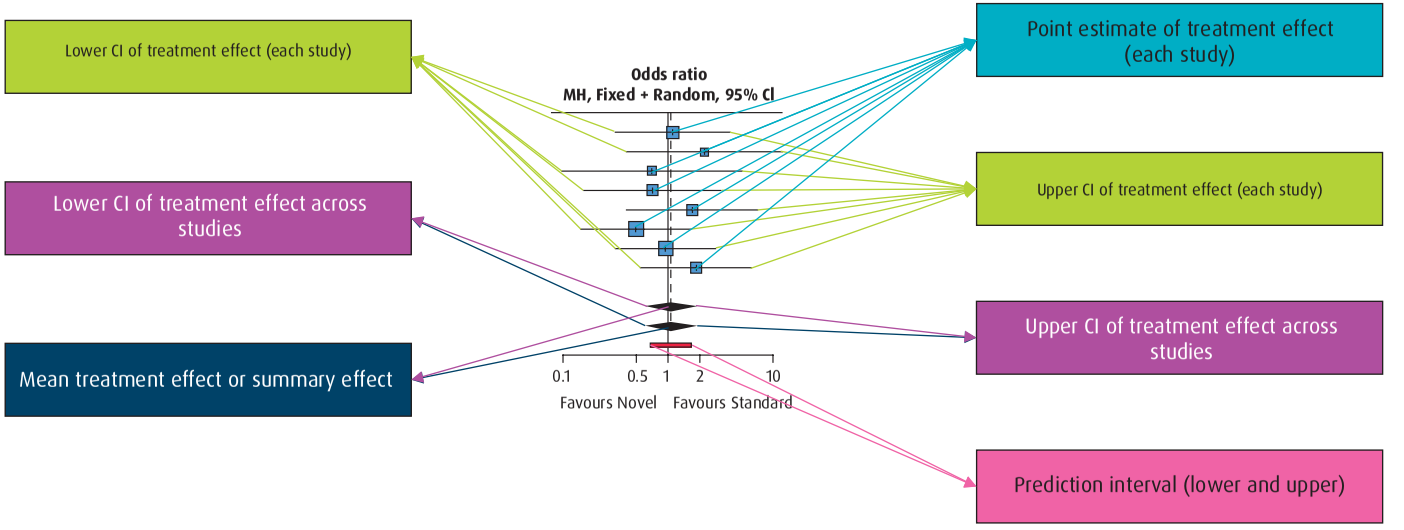


Figure 7.3 Treatment effects and 95% CI



Prediction intervals

When the number of studies in a meta-analysis is large, the CI of a random-effects model may be narrower than expected based on the heterogeneity. [1] Use of prediction intervals (which take into account the variability in the treatment effects across studies) rather than CI may overcome this problem, although they can be spuriously narrow or wide when there are few studies included in the meta-analysis. [2] Therefore, they are recommended if the meta-analysis includes 10 or more studies. [2]

In the forest plot, prediction intervals are represented as a red bar below the treatment effect across the studies, with the left and right edges of the bar representing the lower and upper prediction intervals (Figure 7.3).

Treatment effects in each study

Point estimate represents the treatment effect observed in the study. The interpretation of 95% CI is provided in the section 'Line of no effect'.

In the forest plot, the treatment effect in each study is represented by a horizontal line (Figures 7.2 and 7.3). The blue rectangle within the line indicates the point estimate while the edges of the line indicate the lower and upper 95% CI (Figure 7.3). Further interpretation of the treatment effects in each study is provided in the section 'Line of no effect'.

Line of no effect

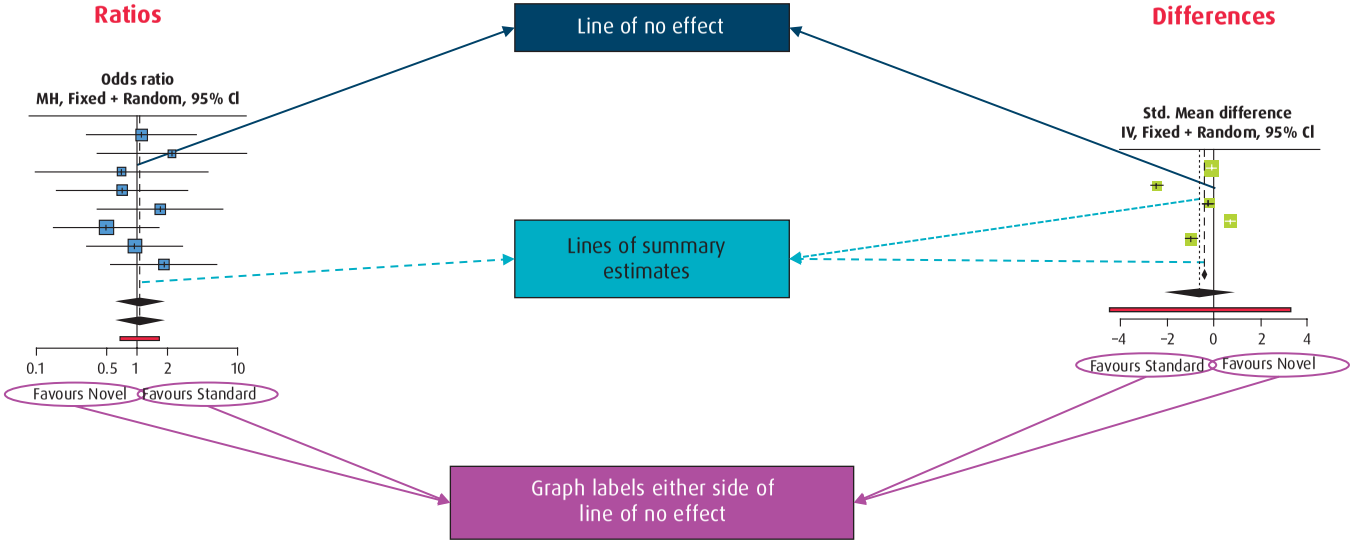
The line of no effect indicates the value when the effects of the intervention and the comparator on an outcome are identical. If the intervention and comparator have the same effect on an outcome, the treatment effect is null (the meaning of null is zero [3]).

The line of no effect is indicated by a solid vertical line in the middle of the graphical part of the forest plot (Figure 7.2 and Figure 7.4). You might notice in Figure 7.4 that the line of no effect corresponds to 1 for the outcome on the left and to 0 for the outcome on the right. This is because the outcome on the left used 'ratio' as effect measure and the outcome on the right used 'difference' as the effect measure. For all ratios, the line of no effect (that is, the line representing a treatment effect of null) is 1; for all differences, the line of no effect is 0. This is because of the way that these effect measures are calculated.

The line of no effect should not be confused with other dotted and dashed vertical lines in the plot which represent the summary estimates of the random-effects model and fixed-effect model, respectively.

The line of no effect is very helpful in interpreting whether there is a difference in the effect of the intervention versus comparator on the outcome. In this context, the interpretation of 95% CI of treatment differences between the intervention and comparator is no different from that of 95% CI of a single group described in Chapter 5: 95% CI of treatment difference mean that there is only a

Figure 7.4 Line of no effect



5% probability that the intervals exclude the average population treatment difference. If the 95% CI do not include the 'null' treatment effect, there is only a 5% probability that these intervals excluded the 'null' treatment effect by chance if the average treatment effect in the population (from which the trial participants were selected) was null (in the absence of bias). This is equivalent to rejecting the null hypothesis with a *p*-value of less than 0.05, commonly referred to as 'statistically significant difference' or in this context, 'statistically significant treatment effect'. If the diamond representing the random-effects or the fixed-effect model lies entirely to one side of the line of no effect—in other words, there is no crossing of paths between the diamond representing the summary estimate and 95% CI and the line of no effect—this means that the 95% CI do not include 'null' treatment effect; that is, there is a statistically significant treatment effect.

In Figure 7.4, for the outcome on the left, the diamonds of both the random-effects and fixed-effect models cross the line of no effect, indicating that there is no statistically significant treatment effect (across studies) for either model. For the outcome on the right, the diamond of the random-effects model crosses the line of no effect, indicating that there is no statistically significant treatment effect (across studies) for the random-effects model. However, the diamond of the fixed-effect model lies entirely to the left of the line of no effect (it does not cross the line of no effect), indicating that there is a statistically significant treatment effect (across studies) for the fixed-effect model. Whether this means that the intervention is better or the comparator is better for the outcome depends on the context and can be found in the graph labels on either side of the line of no effect.

In addition to allowing easy visualisation of whether the treatment effect (across studies) is statistically significant or not, the line of no effect can help with visualisation of whether the treatment effect in each study is statistically significant. In this context, the meaning of 95% CI of treatment effects in each study is the same as that of treatment effects across studies. If the horizontal line representing the point estimate and 95% CI of a study crosses the line of no effect, this indicates that the treatment effect was not statistically significant in that study; if the horizontal line lies entirely to the left or entirely to the right of the line of no effect, this indicates that the treatment effect was statistically significant in that study. As for the treatment effect across studies, if the 95% CI of the treatment effect include 1 for ratios or 0 for differences, there is no statistically significant treatment effect; if the 95% CI do not include 1 for ratios and 0 for differences, there is a statistically significant treatment effect. Whether this means that the intervention is better or the comparator is better for the outcome can be found in the graph labels on either side of the line of no effect (as for treatment effect across studies).

In Figure 7.4, for the outcome on the left, none of the studies cross the line of no effect, indicating that the treatment effect was not statistically significant in any study. For the outcome on the right, the studies represented by the first and

third lines cross the line of no effect (it is not clear whether the third line crosses the line of no effect in Figure 7.4, but the treatment effect of the study in text format, not shown in Figure 7.4, indicated that 0 was included in the 95% CI), the studies represented by the second line and last line are entirely to the left of the line of no effect, and the study represented by the fourth line is entirely to the right of the line of no effect. This means that there were no statistically significant treatment differences for the first and third studies; there were statistically significant treatment differences favouring 'standard' (short form for 'Standard treatment', the comparator) for the second and last studies, while there was a statistically significant treatment difference favouring 'novel' (short form for 'Novel treatment', the intervention) for the fourth study.

You might notice in Figure 7.4 that the graph labels for the outcome on the left and those for the outcome on the right are reversed: for the outcome on the left, 'Favours Novel' is to the left of the line of no effect and 'Favours Standard' is to the right of the line of no effect; for the outcome on the right, 'Favours Novel' is to the right and 'Favours Standard' is to the left of the line of no effect. This is because fewer participants developing the event indicated better health for the outcome on the left, while higher values indicated better health for the outcome on the right. In EQUAL-SR software, whether 'Favours intervention' or 'Favours comparator' is decided automatically from the information that is provided as part of the data for analysis: you have to indicate whether more events or higher values are better or worse for participants as part of the uploaded data. However, in most other software, this has to be decided manually. An algorithm to decide whether the 'Favours intervention' should be on the left or right of the line of no effect is provided in Figure 7.5 to allow reviewers using other software to perform the meta-analysis.

The forest plot also helps in the interpretation of the heterogeneity and clinical importance of the treatment effects. These are covered in the sections 'Heterogeneity' and 'Clinical interpretation of results'.

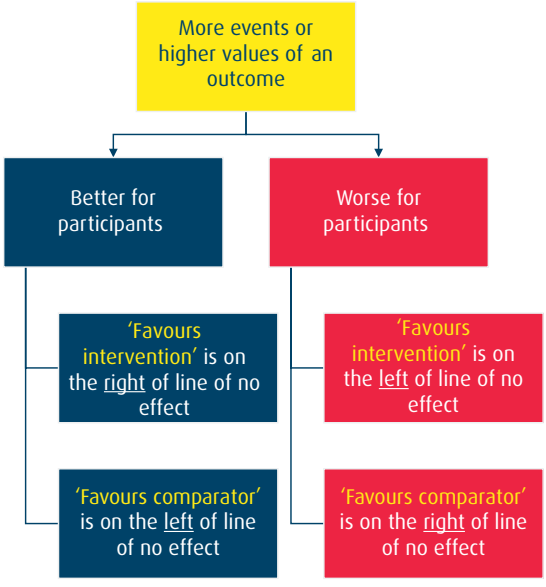
p-value

The *p*-value indicates whether the treatment effect across studies is statistically significant. A *p*-value is obtained for the random-effects model and another for the fixed-effect model (Figure 7.2). Generally, a *p*-value of less than 0.05 is considered statistically significant. However, we recommend against using *p*-value to arrive at conclusions about whether an intervention is effective against the comparator. Please see section 'Clinical interpretation of results' for further details.

Heterogeneity

You might recall that 'statistical heterogeneity' or simply 'heterogeneity' is a term used to describe differences in treatment effects observed between the studies. [2] In Chapter 6, in the coin-toss experiment analogy for the fixed-effect model,

Figure 7.5 Graph labels



we noted that some differences in results of multiple experiments may be expected by chance, even when the same coin is used. Therefore, it is reasonable to expect that there are some differences in the treatment effects observed between the studies. The question is when we can say that the observed differences in the treatment effects are due to chance. This is what the different measures of heterogeneity attempt to answer.

Different measures of heterogeneity are provided in the forest plot (Figure 7.2). These include I^2 and Chi^2 tests for heterogeneity, and between-study variance (Tau^2).

I^2

I^2 is the percentage of variability between treatment effects in the studies that cannot be explained by chance. [4] Using threshold values for interpretation of I^2 can be misleading and the following rough guide for interpretation of I^2 has been suggested. [2]

- 0% to 40%: may not be important.
- 30% to 60%: may represent moderate heterogeneity.
- 50% to 90%: may represent substantial heterogeneity.
- 75% to 100%: may represent considerable heterogeneity.

Other factors, such as the Chi^2 test of heterogeneity and overlap of CI (please see below) should be considered while interpreting I^2. You can also use CI of I^2 to

interpret heterogeneity. However, this might not be reliable when there are few studies in the meta-analysis. [2]

Chi^2 test of heterogeneity

The Chi^2 test of heterogeneity is a test to find out whether the differences in treatment effects between studies can be explained by chance. [2] A *p*-value is provided for the Chi^2 test of heterogeneity. A *p*-value of less than 0.10 is considered to indicate statistically significant heterogeneity. However, when there are few studies, the Chi^2 test of heterogeneity can miss heterogeneity. [2]

Between-study variance

This is a measure of the variability between the treatment effects across studies. If the Tau^2 95% CI do not include 0, this can indicate the presence of heterogeneity. The forest plot only contains the point estimate of the Tau^2. The Tau^2 is not usually used to interpret heterogeneity in meta-analysis involving two treatments, only because of the presence of other widely accepted measures for estimating heterogeneity.

Overlap of 95% CI

In addition to these measures, the point estimates and 95% CI of each study also help with the interpretation of heterogeneity. When there is overlap of 95% CI, this indicates that heterogeneity is low or absent; when there is no overlap of 95% CI (and this has to be the case for just one study's 95% CI not overlapping with the 95% CI of any other study in the meta-analysis), this indicates the presence of heterogeneity. For example, in Figure 7.4, for the outcome on the left, there is overlap between the horizontal lines representing the studies; for the outcome on the right, there is overlap between the horizontal lines representing the first and third studies, but no overlap in the lines representing the remaining studies. When there is no overlap in the 95% CI and the point estimates from the studies are on the same side of the line of no effect, it is called 'heterogeneity in the magnitude of effect'. This means that the treatment effects from the studies are favouring either the intervention or the comparator. When there is no overlap in the 95% CI and the point estimates from the studies are on the opposite side of the line of no effect, it is called 'heterogeneity in the direction of effect'. This means that the treatment effects from the studies are pointing to treatment effect in opposite directions, that is, some are favouring intervention and others the comparator. For the outcome on the right in Figure 7.4, there is no overlap of the 95% CI for the second and last studies, but the point estimates are on the same side of the line of no effect favouring standard treatment; therefore, the heterogeneity between these two studies is 'heterogeneity in magnitude of effect'. On the other hand, there is no

overlap of the 95% CI for the second (or last) study and fourth study and the point estimates are on opposite sides of the line of no effect. So, for the second (or last) study, the point estimates favour standard treatment, while for the fourth study, the point estimate favours novel treatment. Therefore, the heterogeneity between the second (or last) study and fourth study is 'heterogeneity in direction of effect'. For the outcome on the left in Figure 7.4, the point estimates of lines representing the third, fourth and sixth studies are on the left of the line of no effect while those representing the first, second, fifth and last studies are on the right of the line of no effect. However, there is overlap of CI between the studies. Therefore, you cannot interpret this as presence of heterogeneity even though the point estimates lie on either side of the line of no effect.

How do you deal with heterogeneity?

An overview of how to deal with heterogeneity is provided in Figure 7.6. The first thing to do is to check the data for errors. A list of common errors that I have encountered in the past is shown in Figure 7.6.

If heterogeneity persists after correcting any errors (this may include seeking clarifications from study authors), check whether the heterogeneity can be explained by subgroup analysis and metaregression. Meta-analytical treatment effect estimates in the presence of considerable statistical heterogeneity have no clinical application, as it is difficult to know which of PICO the treatment effect represents. Therefore, when subgroup analysis resolves the statistical heterogeneity, you can consider using the subgroup results for interpretation and present only subgroup results for categorical moderators. However, you must be cautious about performing too many subgroup analyses as this can lead to spurious conclusions. Please see section 'Interpretation of the subgroup analyses and metaregression' for further limitations of subgroup analyses. Furthermore, if you perform more than one subgroup analysis and present subgroup results for more than one subgroup analysis, the patient and healthcare professional involved in shared decision making will not know which subgroup analysis results should be used for deciding the treatment.

Statistical heterogeneity which cannot be explained by clinical differences in PICO or methodological differences in the study design or risk of bias decreases the confidence in results (please see section 'Certainty of evidence'). In this case, you should consider not performing the meta-analysis and take this unexplained heterogeneity into account while arriving at conclusions.

Other aspects available from forest plots

There are other aspects available from forest plots. These include the name of the outcome, data in a format used for meta-analysis, effect measure, method, and

Figure 7.6 Dealing with heterogeneity

models used for meta-analysis. These are self-explanatory. Some abbreviations are used for the method. MH indicates the Mantel-Haenszel method, GLMM indicates the Generalised Linear Mixed Model, and IV indicates the inverse variance method.

Statistical interpretation of treatment effects

In systematic reviews of intervention, statistical interpretation of the results of the meta-analysis is largely based on risk of bias in the studies and the meta-analytical summary estimates and the 95% CI. Although other levels of CI can be used, they rarely are.

In the section 'Line of no effect', we stated that if the 95% CI do not include 'null' treatment effect, there is only a 5% probability that these intervals excluded the 'null' treatment effect by chance if the average treatment effect in the population (from which the trial participants were selected) was null (in the absence of bias). What do you do in the presence of bias?

In Chapter 5, we described the use of RoB 2 [5,6] and ROBINS-I [7,8] for the assessment of the risk of bias in RCTs and non-randomised studies respectively. From this, an assessment should be made about whether any observed differences in the outcome between the intervention and comparator could be explained by other factors, such as systematic differences in the participants, how often they made planned clinic visits, received planned additional tests or interventions, how the outcome was measured in the two groups, and whether some participants who developed 'undesirable outcomes' were excluded from the analysis. Estimates are available for the deviations from the true average population treatment effect that each of the biases causes, [9,10] although there is uncertainty around these deviations as different studies provide different estimates. [9,10] Methods to adjust the treatment effects for these biases are available, [11–13] but are not routinely used.

INTERPRETATION OF THE SUBGROUP ANALYSES AND METAREGRESSION

The ICEMAN tool can help with the interpretation of subgroup analyses and metaregression. [14] In particular, this tool can help with deciding whether you should use the overall meta-analysis results or subgroup meta-analysis results for clinical interpretation of results.

This tool consists of eight questions for meta-analysis. The tool allows assessment of subgroup analysis performed by other systematic reviewers. Therefore, some of the options include 'unclear', which are not relevant if you used the tool to decide whether you should use the subgroup meta-analysis results or overall results. These are also not relevant if one is using the tool to decide whether subgroup

estimates or overall estimates should be used for clinical interpretation. Therefore, the question options and guidance here relate to the use of the tool for your own systematic review. For detailed guidance and a complete list of options available for each question, please refer to the manual available from Schandelmaier et al. [14]

1. Is the effect modification based on comparison within rather than between RCTs?

Except in factorial trials, where separate data are available for participants with and without the potential 'effect modifiers' or 'moderators', as mentioned in Chapter 6, the subgroup analyses are usually done at the 'study level' (all participants in a study are included in a subgroup) rather than 'within-study' level (that is, a subset of participants in a study included in different subgroups). However, the studies do not differ only by the characteristic for which the subgroup analysis is carried out. There are other confounders that can decrease the confidence in the results. [14]

- **a.** 'Completely between': subgroup analysis performed using only study-level data.
- **b.** 'Mostly between': subgroup analysis performed using mostly study-level data with occasional within-study-level data.
- **c.** 'Mostly within': subgroup analysis performed using mostly within-study-level data or individual participant data.
- **d.** 'Completely within': based on individual participant data (unlikely as the methods described in this book are for aggregate data).

2. If two or more within-trial comparisons are available, is the effect modification similar from trial to trial?

This involves calculating the ratio of the treatment effects in each study with 'within-study' data and assessing the consistency of this ratio. EQUAL-SR software can help with this assessment.

- **a.** 'Definitely not similar': ratio of the treatment effects is in different directions for each study with 'within-study' data (very much similar to the concept of 'heterogeneity in the direction of effect' discussed in the section 'Overlap of 95% CI').
- **b.** 'Mostly similar': ratio of the treatment effects is in the same direction for each study with 'within-study' data but differs by magnitude (very similar to the concept of 'heterogeneity in the magnitude of effect' discussed in the section 'Overlap of 95% CI').
- **c.** 'Definitely similar': ratio of the treatment effects is in the same direction for each study with similar magnitudes.

3. For between-RCT comparisons, is the number of studies large?

This involves a different response for subgroups and metaregression.

   a. 'Very small': 1 or 2 in the smallest subgroup; 5 or fewer studies in metaregression (we generally recommend at least 2 studies in each subgroup and 10 studies in metaregression to perform a subgroup analysis or metaregression).
   b. 'Rather small or unclear': 3 or 4 in the smallest subgroup; 6 to 10 studies in metaregression.
   c. 'Rather large': 5 to 9 in the smallest subgroup; 11 to 15 studies in metaregression.
   d. 'Large': 10 or more in the smallest subgroup; more than 15 in metaregression.

4. Was the direction of effect modification correctly hypothesised a priori?

Generally, you would be able to predict which subgroups are likely to benefit if there is biological plausibility. This is usually explained in the rationale for performing subgroup analysis.

   a. 'Definitely no': Clearly post hoc (that is, an unplanned subgroup analysis performed after reviewing the data), results inconsistent with expected direction, or biologically very implausible.
   b. 'Definitely yes': Results consistent with expected direction of effect modification.

5. Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification?

   a. 'Chance a very likely explanation': Test for subgroup differences or metaregression *p*-value > 0.05
   b. 'Chance a likely explanation': Test for subgroup differences or metaregression *p*-value > 0.01 up to 0.05
   c. 'Chance may not explain': Test for subgroup differences or metaregression *p*-value > 0.005 up to 0.01
   d. 'Chance an unlikely explanation': Test for subgroup differences or metaregression *p*-value 0.005 or less.

6. Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?

When you perform a lot of statistical tests, you can get statistically significant results by chance. You have to adjust the *p*-value for this multiple testing. EQUAL-SR software can help with adjustment of multiple testing.

   a. 'Definitely no': Explicitly exploratory analysis or more than 10 effect modifiers tested, and multiplicity not considered in analysis.
   b. 'Probably no': 4 to 10 effect modifiers tested, and multiplicity not considered in analysis.
   c. 'Definitely yes': 3 or fewer effect modifiers tested, or multiplicity considered in analysis.

7. Did the authors use a random-effects model?
   a. 'Definitely no': Only fixed-effect model used.
   b. 'Definitely yes': Random-effects model used.
8. If the effect modifier is a continuous variable, were arbitrary cut points avoided?

   We recommend metaregression rather than arbitrarily dividing continuous type moderators into categories; alternatively, you can divide continuous type data into categories at the protocol stage, with clear rationale for categorising the data.
   a. 'Definitely no': Analysis based on exploratory cut point(s), for example, choosing the cut-off point to obtain the best *p*-value.
   b. 'Definitely yes': Metaregression.

Rough guidance for interpretation is provided below. For detailed guidance, please refer to the manual available from Schandelmaier et al. [14] Please note that this is a simplified version for assessing the credibility of the subgroup analysis in your own systematic review.

- Very low credibility: all responses with reduced credibility.
- Low credibility: Two or more responses with reduced credibility.
- Moderate credibility: One response with reduced credibility.
- High credibility: No response with reduced credibility.

Separate estimates (of subgroups) are used for clinical interpretation if the credibility is moderate or high.

INTERPRETATION OF THE SENSITIVITY ANALYSES

Sensitivity analysis is performed at the meta-analysis level. For each meta-analysis, you assess whether the interpretation of treatment effects of the intervention versus comparator changes with different assumptions in the sensitivity analyses. If the interpretation of treatment effects does not change with different assumptions, you can conclude that the specific meta-analysis is 'robust' to different assumptions. On the other hand, if the interpretation of treatment effects changes with different assumptions, you can conclude that the specific meta-analysis is 'sensitive' to different assumptions. While the usual focus is on the statistical interpretation of treatment effects, it is reasonable to also assess whether the clinical interpretation of treatment effects is robust or sensitive to different assumptions.

INTERPRETATION OF THE REPORTING BIASES

You can use the ROB-ME tool to assess the risk of bias due to reporting bias. [15,16] However, to be able to assess the risk of bias due to missing evidence in meta-analysis, you need to be able to interpret funnel plots and to identify and distinguish the reasons for small-study effects. Therefore, we start with the interpretation of funnel plots.

Funnel plot asymmetry

What is funnel plot asymmetry?

Essentially, you look for whether the treatment effects from the studies are symmetrical on either side of the summary estimate in the funnel plot ('plane of symmetry').

Interpreting funnel plot asymmetry

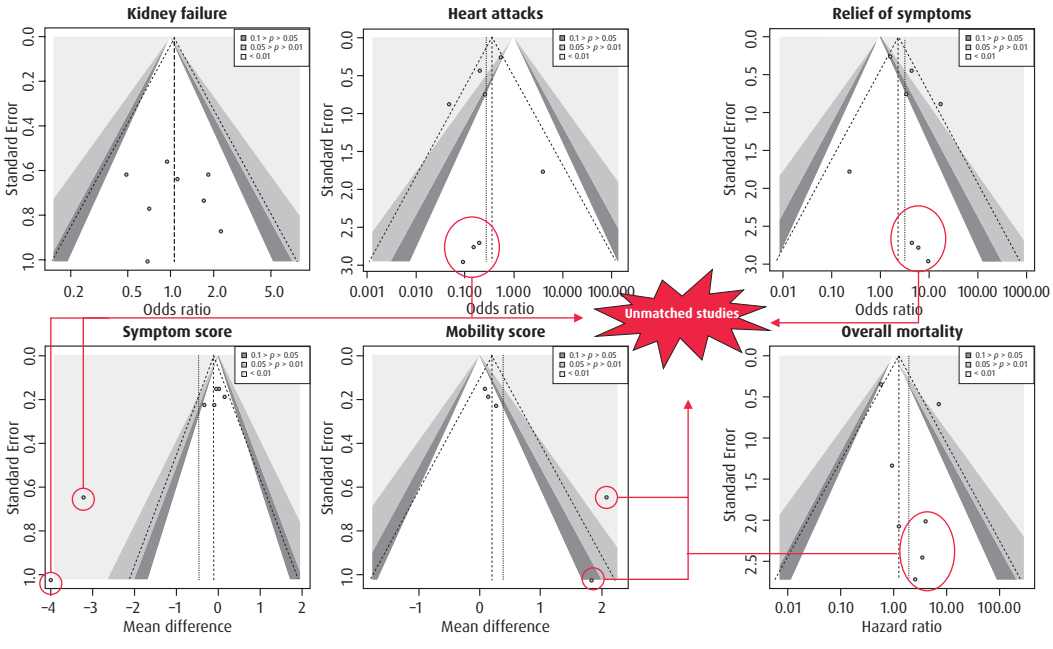
There are many causes of funnel plot asymmetry. These include reporting biases, small-study effects (which could be due to various reasons including reporting biases, poor methodological quality, differences in the type of participants included in the trials, the way the intervention was implemented and differences in comparators), statistical properties of the effect measure used, fraud and chance. [17,18] Detection of fraud in clinical trials is currently in its infancy and there is no consensus about how to do it. [19] Assessment of poor methodological quality is through the assessment of risk of bias. Differences in the type of participants, intervention or comparator can be dealt with by subgroup analysis. The choice of effect measure is based on other factors described in Chapter 6. That leaves you having to interpret whether there is funnel plot asymmetry and whether it is due to reporting biases (whether that is because of small-study effects or other reasons), small-study effect not due to reporting biases, or chance.

Visualisation

This is one of the main ways of interpreting funnel plot asymmetry.

Figure 7.7 shows the funnel plots of six different outcomes: kidney failure, heart attacks, relief of symptoms (top row, left to right), symptom score, mobility score and overall mortality (bottom row, left to right) based on hypothetical data. Of these, a higher number of participants with events, and higher values, mean better health of participants for relief of symptoms and mobility score. Conversely, a higher number of participants with events, or higher values, mean worse health of participants for the remaining outcomes. In Figure 7.7, the

Figure 7.7 Funnel plots



unmatched studies are marked by circles or ellipses. We have described the possibilities in terms of unmatched studies rather than missing studies (the usual way of describing funnel plot asymmetry) since, anecdotally, students appear to understand something that is present in the plot better than something that is missing from the plot.

In the funnel plot for kidney failure, there does not appear to be any asymmetry since the studies appear to be evenly spread on both sides of the summary estimate and broadly around the same levels in the plot.

In the funnel plot for heart attacks, some studies in the left lower quadrant appear to be unmatched with similar studies in the right lower quadrant (that is, studies are potentially missing from the right lower quadrant). The unmatched studies are in the white area of the funnel plot (zone where *p*-value is 0.10 or more). For an outcome where a higher number of participants with events or higher values mean worse health, if studies in the white area from the left half of the funnel plot are unmatched, this indicates that reporting biases are the likely causes of funnel plot asymmetry. The reverse is true for an outcome such as relief of symptoms where a higher number of participants with events, or higher values, mean better health. In this case, if studies in the right half are unmatched and the unmatched studies are in the white areas of the funnel plot, reporting biases should be suspected as the reason for funnel plot asymmetry.

In the funnel plot for symptom score, some studies in the left lower quadrant appear to be unmatched with similar studies in the right lower quadrant; the unmatched studies are in the grey areas of the funnel plot (indicating *p*-values less than 0.10). For an outcome where a higher number of participants with events, or higher values, mean worse health, if studies in the grey areas (particularly further away from the summary estimate) from the left lower quadrant (this is the area for small studies in favour of the intervention) are unmatched, this indicates that causes of small-study effects other than reporting biases are the likely causes of funnel plot asymmetry. The reverse is true for an outcome like mobility score where a higher number of participants with events, or higher values, mean better health. In this case, if studies in the right lower quadrant are unmatched and the unmatched studies are in the grey areas of the funnel plot, causes of small-study effects other than reporting biases should be suspected as the reason for funnel plot asymmetry.

While the funnel plot for overall mortality is similar to that of relief of symptoms (studies in the right half are unmatched and the unmatched studies are in the white area), this is not indicative of reporting bias, since the results favour the comparator and not the intervention.

Statistical tests for funnel plot asymmetry

A *p*-value of less than 0.10 can be considered as statistically significant. [17,20] You should not solely rely on statistical tests, as they may miss funnel plot asymmetry, particularly when the number of studies is low. [20] Furthermore, they do not help with distinguishing the reasons for funnel plot asymmetry. Therefore, they should be used in addition to visualisation.

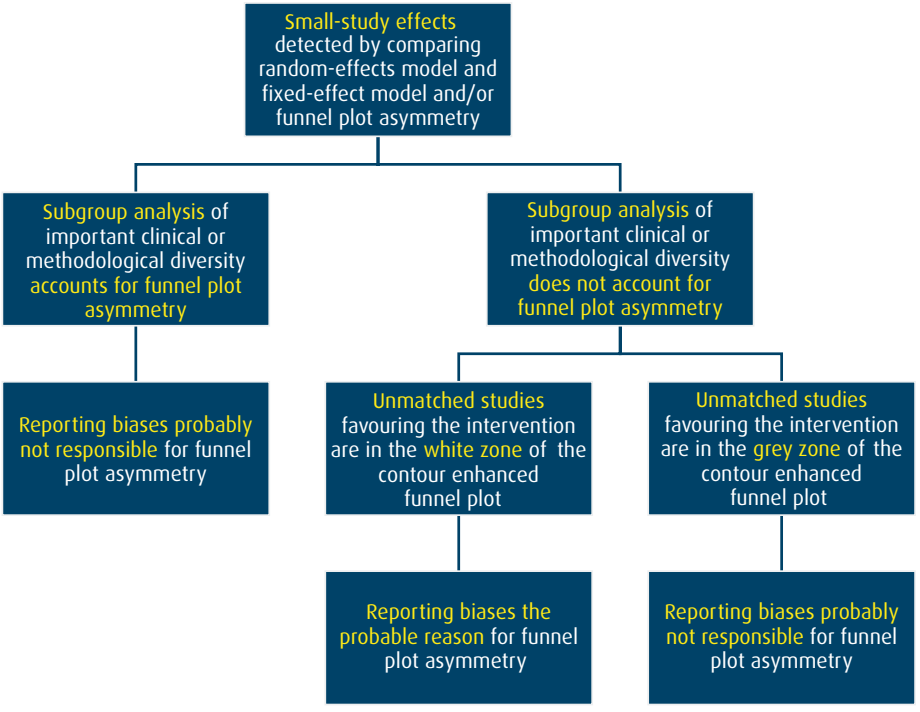
Small-study effects

If the summary estimate of the random-effects model shows greater treatment effect (further away from the line of no effect, that is, 1 for ratios and 0 for differences) in favour of the intervention, small-study effects can be suspected. As described in the section 'Interpreting funnel plot asymmetry', there are multiple reasons for small-study effects. Subgroup analysis to deal with differences in methodological quality or differences in participants, intervention and comparator along with visualisation of funnel plots may help with distinguishing the reason for funnel plot asymmetry. A suggested algorithm for checking whether reporting biases are probably responsible for funnel plot asymmetry is shown in Figure 7.8.

The ROB-ME tool

Having learnt how to interpret funnel plot asymmetry, you can use the ROB-ME tool to assess the risk of bias due to reporting bias. [15,16] The ROB-ME tool contains four steps.

Figure 7.8 Small-study effects



1. Step 1. Select and define meta-analyses that will be assessed for risk of bias due to missing evidence.
2. Step 2. Determine which studies meeting the inclusion criteria for the meta-analyses have missing results.
3. Step 3. Consider the potential for missing studies across the systematic review.
4. Step 4. Assess risk of bias due to missing evidence in a meta-analysis (complete for each meta-analysis).

Detailed guidance for the use of the tool is available from Page et al. [15,16] Please note that the authors of this tool have used the term 'non-reporting bias' to indicate reporting biases. The rationale for using the term 'non-reporting bias' rather than 'reporting bias' is sound. Perhaps the term 'non-reporting bias' will be the term that is preferred in future, but we have used the term 'reporting bias' as this is the term currently widely used to describe biases resulting from missing evidence in meta-analysis. We now provide a short description and some practical guidance.

Step 1. Select and define meta-analyses that will be assessed for risk of bias due to missing evidence

We recommend selecting only the outcomes that are critical to decision making. These are usually the primary outcomes in the systematic review. In terms of further description, you should note the study designs that were included in a meta-analysis, the outcome definitions included and the types of analysis included in the meta-analysis, for example, intention-to-treat effect versus per-protocol effect, final scores versus change scores, adjusted versus unadjusted estimates.

Step 2. Determine which studies meeting the inclusion criteria for the meta-analyses have missing results

This involves a classification of studies included in the systematic review into the following categories.

- A study result is available for inclusion in the meta-analysis.
- No study result is available for inclusion in the meta-analysis.
  - This is usually because a protocol prior to the start of the study is available in the public domain and the outcome was not measured. Outcomes may also not have been measured after the enrolment because of legitimate reasons, for example, equipment non-availability.
- Unclear whether an eligible study result was generated.
  - This is usually because a protocol prior to the start of the study is not available in the public domain.
- No study result is available for inclusion in the meta-analysis, likely because of the *p*-value, magnitude or direction of the result generated.
  - This is usually because it is clear from a protocol available publicly that an outcome stated in the protocol was not reported, or from the study reports from which it is clear the outcome was measured but not reported.

Step 3. Consider the potential for missing studies across the systematic review

This involves a systematic review-level assessment and depends on the type of systematic review undertaken and how thoroughly searches were performed. There are three signalling questions to arrive at the potential for missing studies across the systematic review.

3.1 Were prospectively registered studies or studies identified for a prospective meta-analysis the only type of study eligible for inclusion in the review? (Yes/No)

'Yes': The systematic review includes only studies whose results were not available prior to the start of the systematic review.

'No': The systematic review includes at least one study whose results were available prior to the start of the systematic review.

3.2 If 'No' to 3.1: Would you expect every eligible study to be identifiable regardless of its results?

(Not applicable/Yes/Probably Yes/Probably No/No)

'Yes' or 'Probably Yes': All the studies on the topic are likely to have been prospectively registered.

'No' or 'Probably No': It is unlikely that all non-randomised studies are prospectively registered. Therefore, you can answer this question as 'No' or 'Probably No' if the systematic review includes non-randomised studies. If the intervention and comparator were available prior to July 2005 when registration in a public trials registry became a prerequisite for publication in the ICJME, [21] trials may have been conducted which were never reported because of their findings. Therefore, you can answer this question as 'No' or 'Probably No' if the systematic review includes an intervention and comparator available prior to July 2005.

However, even if the intervention or comparator became available only after July 2005, you cannot be certain that all RCTs were prospectively registered, as an RCT may have been conducted with no intention of reporting in the ICJME, and with the intention to perform a subsequent prospectively registered trial only if the results of the unregistered trial show favourable results.

3.3 If 'Yes' or 'Probably Yes' to 3.2: Were you likely to have found all eligible studies regardless of their results? (Not applicable/Yes/Probably Yes/Probably No/No)

'Yes' or 'Probably Yes': Relevant trial registers have been searched and the search strategy did not include 'outcomes' as one of the domains. This is likely to be the case if the instructions in Chapter 4 were followed.

'No' or 'Probably No': Relevant trial registers have not been searched or the search strategy included 'outcomes' as one of the domains.

Overall, classify Step 3 as 'Yes' if 3.1 was classified as 'No' and either 3.2 or 3.3 was classified as 'No' or 'Probably No'. Overall, we expect most systematic reviews will have a classification of 'Yes'. In other words, the potential for missing studies is present in most systematic reviews.

Step 4. Assess risk of bias due to missing evidence in a meta-analysis (complete for each meta-analysis)

This assessment must be completed for each meta-analysis of interest. There are eight signalling questions which have different choices.

4.1 Of the studies identified, were there any for which no result was available for inclusion in the meta-analysis, likely because of the *p*-value, magnitude or direction of the result generated (refer to Step 2)? (Yes/No)

'Yes': At least one of the studies was classified as 'No study result is available for inclusion in the meta-analysis, likely because of the *p*-value, magnitude or direction of the result generated' in Step 2.

'No': None of the studies was classified as 'No study result is available for inclusion in the meta-analysis, likely because of the *p*-value, magnitude or direction of the result generated' in Step 2.

4.2 If 'Yes' to 4.1: Is it likely that there would be a notable change to the summary effect estimate if the omitted results had been included? (Not applicable/Yes/Probably Yes/Probably No/No/No information)

'Yes' or 'Probably Yes': In random-effects meta-analysis, in the presence of minimal heterogeneity, the addition of even one study could change the results. For fixed-effect meta-analysis, the impact on the results depends upon the number of participants not included in the meta-analysis relative to the number of participants included in the meta-analysis. There are no threshold levels for this ratio. Therefore, we suggest the following rough guidance.

- For random-effects meta-analysis, if there is minimal heterogeneity and intervention shows benefit or lack of harm, if there is even one study with missing data, classify this signalling question as 'Probably Yes'.
- For random-effects meta-analysis, if there are moderate or higher levels of heterogeneity, you can add hypothetical data opposite to the most favourable results for the intervention (but proportionate to the number of participants in the trial with missing data) and check the impact of adding such data on the results.
    - Let us say that the most favourable results in favour of an intervention for a binary outcome were as follows.
        - Number of participants with outcome in intervention group = 40.
        - Total number of participants in intervention group = 160.
        - Percentage of participants with outcome in intervention group = 40/160 = 25%.
        - Number of participants with outcome in comparator group = 80.
        - Total number of participants in comparator group = 160.
        - Percentage of participants with outcome in comparator group = 80/160 = 50%.

- Let us say that in the study with missing data, the number of participants in both the intervention group and the comparator group was 100. The hypothetical data in the study will be as follows.
  - Hypothetical percentage of participants in the intervention group = 50% (percentage of participants in the comparator group in the study with most favourable outcome in favour of the intervention).
  - Hypothetical number of participants with outcome in the intervention group = 50% of 100 = 50.
  - Hypothetical percentage of participants in the comparator group = 25% (percentage of participants in the intervention group in the study with most favourable outcome in favour of the intervention).
  - Hypothetical number of participants with outcome in the comparator group = 25% of 100 = 25.
- If the addition of these hypothetical data for the study with missing data changes the interpretation of the results, answer this signalling question, 'Probably Yes'.

- For fixed-effect meta-analysis, follow the same procedure as for random-effects meta-analysis with moderate or higher heterogeneity.
- EQUAL-SR software can help with this assessment with automated identification of the most favourable study in favour of an intervention and automated calculation of the hypothetical results of the study with missing data if the number of participants in each group is provided; if only the total number of participants in the study with missing data is provided it assumes an equal number of participants in the intervention and comparator groups.

'No' or 'Probably No': If the above guidance is followed and there is no impact on the results, classify the signalling question as 'Probably No'.

'No information': Sample size is missing for any of the studies with missing data.

4.3 Of the studies identified, were there any for which it was unclear whether an eligible result was generated (refer to Step 2)? (Yes/No)

'Yes': At least one of the studies was classified as 'Unclear whether an eligible study result was generated' in Step 2.

'No': None of the studies was classified as 'Unclear whether an eligible study result was generated' in Step 2.

4.4 If 'Yes' to 4.3: Is it likely that there would be a notable change to the summary effect estimate if the potentially omitted results had been included? (Not applicable/Yes/Probably Yes/Probably No/No/No information)

Same as for signalling question 4.2.

4.5 Do circumstances (identified in Step 3) indicate potential for some eligible studies not being identified because of the *p*-value, magnitude or direction of the results generated? (Yes/No)

'Yes': Classification in Step 3 is 'Yes'.

'No': Classification in Step 3 is 'No'.

4.6 If 'Yes' to 4.5: Is it likely that studies not identified had results that were eligible for inclusion in the meta-analysis? (Not applicable/Yes/Probably Yes/Probably No/No)

'Yes' or 'Probably Yes': 'Core outcome sets' [22] and regulatory requirements may provide an indication of whether the outcome is likely to have been measured. If core outcome sets have been in existence for a long period of time and have been used in trials that reported prior to enrolment of the study with missing data, it is likely that the outcome was measured in the trial with missing data. For some interventions, you would expect some outcomes to be measured routinely. For example, in trials of major surgical interventions, you can expect that the major complications and all-cause mortality is measured for at least 30 days; for trials in intensive care unit interventions, you can expect in-hospital mortality was measured; for trials involving two interventions related to oncological control in participants with cancer, you can expect all-cause mortality to be measured. In situations where there is a reasonable expectation that the outcome was measured in the trial but has not been reported, answer this signalling question as 'Probably Yes'.

'No' or 'Probably No': It is unlikely that the outcome was measured. For example, HRQoL assessment is not performed routinely for most interventions. In the absence of core outcome sets in existence for a considerable period of time prior to the start of the trial, HRQoL may not have been measured routinely.

4.7 If 'Yes' to 4.1, 4.3 or 4.5: Does the pattern of observed study results suggest that the meta-analysis is likely to be missing results that were systematically different (in terms of *p*-value, magnitude or direction) from those observed? (Not applicable/Yes/Probably Yes/Probably No/No)

'Yes' or 'Probably Yes': Reporting biases are identified as the probable cause for funnel plot asymmetry.

'No' or 'Probably No': There is no evidence of reporting biases.

4.8 If 'Yes', 'Probably Yes' or 'No Information' to 4.2, 4.4, 4.6 or 4.7: Did sensitivity analyses suggest that the summary effect estimate was biased due to missing results? (Not applicable/Yes/Probably Yes/Probably No/No)

'Yes' or 'Probably Yes': The role of complex sensitivity analyses such as selection models is unclear. However, if they are performed they clearly show the results change depending on the assumptions.

'No' or 'Probably No': No complex sensitivity analysis is performed or, if performed, it shows that the results do not change with assumptions.

Based on the answers to the signalling questions, the risk of bias due to missing data may be considered 'Low', 'Some concerns', or 'High' using algorithms available from Page et al. [15,16] If the answers to the signalling questions are provided, EQUAL-SR software provides the risk of bias using the algorithm.

INTERPRETATION OF CONFLICTS OF INTEREST

There is currently no consensus on how to incorporate conflicts of interest in studies in systematic reviews. Conflicts of interest probably play a significant role in reporting biases (which were covered previously). Other study design issues are likely to be addressed in the risk of bias (for example, appropriate definitions and measures of outcomes), and applicability issues concerning whether the participants studied are representative of the patients seen in the clinic can be handled within the indirectness domain of GRADE, thereby decreasing certainty of evidence (for further details, please section 'Certainty of evidence').

However, when numerical data are not reported and you have to accept the study author's interpretation of results, it can be problematic: Lundh et al. showed that there was more disagreement between the results and conclusions (which were more favourable to the sponsor) in industry-sponsored trials than non-industry-sponsored trials. [23] Therefore, we recommend a sensitivity analysis excluding industry-funded or industry-sponsored trials when using non-numerical data for interpretation of results.

TACIT is currently being developed [24,25] and may provide further information on incorporating conflicts of interest into the analysis and/or interpretation of the results of a systematic review.

FRAUD IN CLINICAL TRIALS

A book on systematic review cannot be considered complete if there is no mention of fraud in clinical trials. Systematic reviews rely on clinical trials usually performed by others. Some researchers estimate that approximately 25% of clinical trials in some subject areas could have serious flaws or be fraudulent. [19] However, exclusion of a genuine trial on the basis that it was flawed can lead to wrong conclusions. [19] There is currently no consensus on how to detect flawed trials. [19] Therefore, fraud in clinical trials continues to be an unresolved challenge with no clear methods for its detection or guidance available. Until such

methods and guidance are available, if fraud is suspected, you can perform a sensitivity analysis to consider the impact of a suspected fraudulent trial on the results. You might want to call this 'unusual pattern of results' to avoid a genuine trial being labelled as fraudulent.

CLINICAL INTERPRETATION OF RESULTS

In order to allow the clinical interpretation of the results, you need to learn three additional concepts. These are described next.

Clinically important differences

What are clinically important differences?

From the summary estimate and 95% CI, it is possible to know whether the observed apparent differences in the outcome between the intervention and comparator are due to chance. However, these estimates and CI do not provide information on whether this difference is clinically important.

Generally, a value of 10% or 20% relative decrease or increase are considered clinically important differences, although these vary across disease conditions and outcomes. A 10% relative decrease in risk means a risk ratio of 0.90. How did we arrive at a risk ratio of 0.90 from 10% relative decrease in risk? A 10% relative decrease in risk can be written in numerical terms as '–0.10'. Since null treatment effect is indicated by 1, we add 1 and '–0.10' to arrive at 0.90. Using the same method, 10% relative increase in risk is '+0.10' in numerical terms and when added to 1, we arrive at 1.10 for a 10% relative increase.

You might argue that any reduction in all-cause mortality is clinically important. However, even for all-cause mortality, there is a trade-off as the intervention might cause more complications than the comparator, particularly if the comparator was 'no treatment'. This can impact the patient's quality of life. Besides, a 10% relative decrease in risk translates to an absolute decrease in risk of 0.5% when the baseline risk of all-cause mortality is 5%. Therefore, we are considering small changes in all-cause mortality as clinically important in most instances.

Table 7.1 indicates relative changes and how that translates to clinically important treatment effects, expressed as ratios.

However, there is no standard way of finding clinically important differences for continuous outcomes. For example, let us say that a new intervention increases the HQRoL with a range of 0 to 100 by 2, which was statistically significant in the absence of bias. What does that mean to the patient? It is difficult to know. Therefore, for continuous outcomes, you should use the concept of 'minimal important

Table 7.1 Relative change and what it means for clinically meaningful treatment effects

Relative change	Clinically meaningful increase	Clinically meaningful decrease
10%	0.90	1.10
15%	0.85	1.15
20%	0.80	1.20
25%	0.75	1.25
30%	0.70	1.30

difference'. 'Minimal important difference' (MID) is the smallest difference in the score of the outcome of interest that patients (or their proxies) perceive as an important (beneficial or harmful) difference, and that would lead the patient or clinician to consider a change in the management. [26] Various methods are available to calculate MID: anchor-based methods appear to be the only ones that provide an answer to the MID as described above. [27] Other methods may provide similar [26] or different results. [28] We are not aware of any publicly available search databases that provide MID; PROMID appears to be a subscription-based service and claims to be an inventory of more than 7,000 MIDs for more than 600 patient-reported outcome measures. [29] Therefore, you have to search the medical literature databases for the MID of the measures that you use in the systematic review. For example, if you perform a systematic review of interventions for patients with degenerative knee diseases, you can use the systematic review by Devji et al. [30] to interpret what different scores mean; if you perform a systematic review of interventions for patients with shoulder conditions, you can use the systematic review by Hao et al. [31] For many conditions, you may have to search for primary studies from medical literature to obtain the MID.

Table 7.2 provides the MIDs reported by Devji et al. [30] and how that translates to clinically important treatment effects.

How do clinically important differences help with clinical interpretation of evidence?

If you accept that clinically important differences may not be 'null' differences, you need to understand whether the 95% CI include the clinically important decrease or clinically important increase. This is best understood with the help of an illustration. Let us consider the example of the EuroQol five dimensions questionnaire in patients with degenerative knee diseases. The MID is 0.15, [30] a clinically important decrease is '–0.15' and a clinically important increase is '+0.15' (Table 7.2).

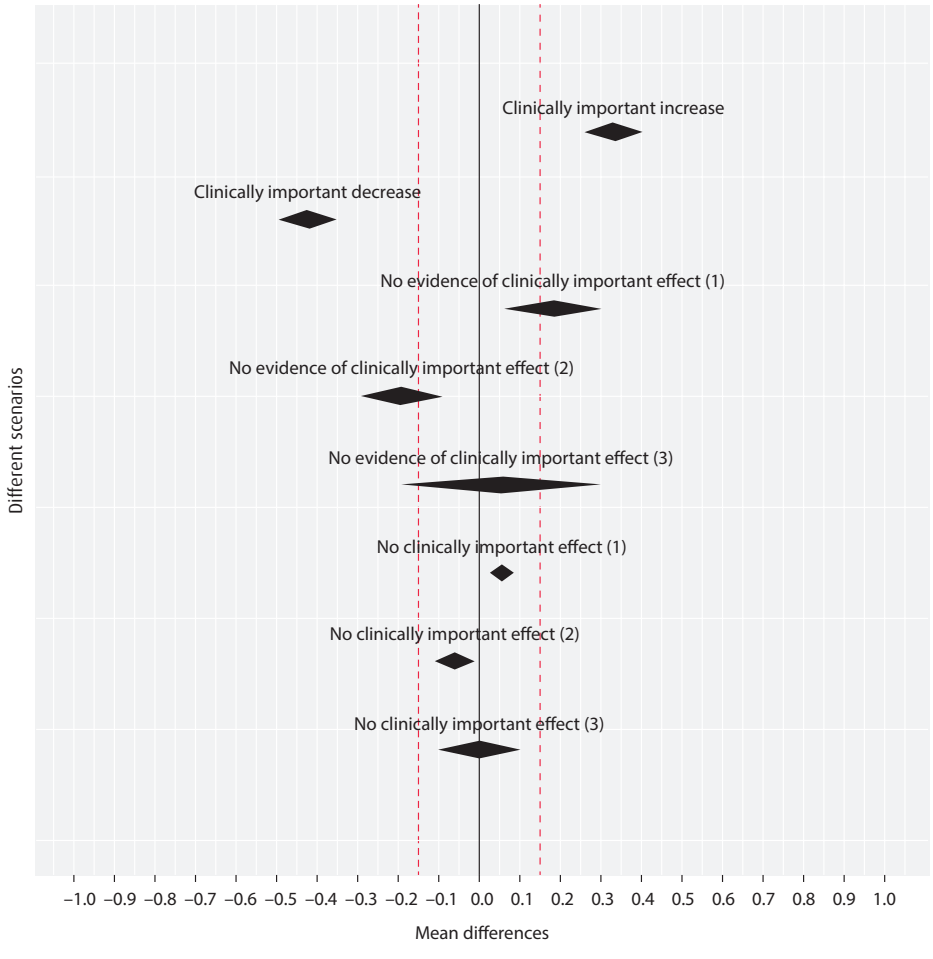
Table 7.2 Minimal important differences and what it means for clinically meaningful treatment effects

Measure	Patient group	Outcome	MID	Clinically meaningful decrease	Clinical meaningful increase
Western Ontario and McMaster University Osteoarthritis Index	Patients with degenerative knee diseases	Pain	12	–12	+12
Western Ontario and McMaster University Osteoarthritis Index	Patients with degenerative knee diseases	Function	13	–13	+13
Knee injury and Osteoarthritis Outcome Score	Patients with degenerative knee diseases	Pain	12	–12	+12
Knee injury and Osteoarthritis Outcome Score	Patients with degenerative knee diseases	Activities of daily living	8	–8	+8
EuroQol five dimensions Questionnaire	Patients with degenerative knee diseases	HRQoL	0.15	–0.15	+0.15

Figure 7.9 shows different eight different scenarios. As in the forest plots, the diamonds in Figure 7.9 refer to the treatment effect, with the left and right edges of the diamond representing the lower and upper CI of the treatment effect. The dotted red lines represent the clinically important decrease or increase.

1. Clinically important increase: The 95% CI do not include either clinically important decrease or clinically important increase and is entirely to the right of clinically important increase. In the absence of bias, you can conclude that this is clinically important. Whether this was in favour of intervention or comparator depends on context. Since a higher value on the EuroQol five dimensions questionnaire indicates better health, the intervention is better than the comparator for this outcome, on average. However, this has to be interpreted in the context of other aspects which we will describe in the section 'Certainty of evidence'.
2. Clinically important decrease: The 95% CI do not include either clinically important decrease or clinically important increase and is entirely to the left of clinically important decrease. As higher values indicate better health (in this example), the intervention is worse than the comparator for this outcome, on average. Again, this has to be interpreted in the context of other aspects described in the section 'Certainty of evidence'.

Figure 7.9 Interpretation of clinically important differences



3. No evidence of clinically important effect: In the next three scenarios in Figure 7.9, the 95% CI include either clinically important decrease or clinically important increase or both. In 'No evidence of clinically important effect (1)', the 95% CI include clinically important increase and clinically unimportant increase. Therefore, you can conclude that there was no evidence of a clinically important treatment effect, although you might notice that the diamond is entirely on the right side of the line of no effect indicating that there was a statistically significant increase in the score on the EuroQol five dimensions questionnaire. However, you cannot rule out a clinically important increase, as the 95% CI include the clinically important increase.

4. No clinically important effect: In the next three scenarios in Figure 7.9, the 95% CI are entirely between the clinically important decrease and

clinically important increase. This means that the intervention does not have any clinically important effect, even though in some of the scenarios, there were statistically significant treatment effects. Again, this has to be interpreted in the context of other aspects described in the section 'Certainty of evidence'.

EQUAL-SR software has a function that can help with interpreting whether the differences are clinically important, but this requires a relative change that is clinically important for ratios, and MID for continuous outcomes in addition to the 95% CI.

Applicability to patients

In the section 'Line of no effect', we stated that if the 95% CI do not include the 'null' treatment effect, there is only a 5% probability that these intervals excluded the 'null' treatment effect by chance if the average treatment effect in the population (from which the trial participants were selected) was null (in the absence of bias).

This also means that you cannot make any inference about the treatment effect on a population that differs from the types of participants included in the trial in important characteristics. As an example, consider a trial comparing major surgery versus supportive treatment only for cancer. If the trial included only participants likely to withstand major surgery based on the absence of comorbidities and limited extent of cancer and showed benefit for surgery, this evidence cannot be extrapolated to people with extensive comorbidities and extensive spread of cancer. Patients and healthcare professionals appear to intuitively understand this concept. However, in drug trials, many patients likely to use the treatment are excluded from clinical trials, [32,33] meaning that the same level of awareness about applicability issues may be lacking. This has the potential for the treatment effect to be different in the real-world population.

Sample size calculations

In Chapter 5, we described a coin-toss experiment where the number of heads observed in a 10 coin-toss experiment differs considerably from the expected number of heads. This is a phenomenon called random error, where the observed values differ from the expected value by chance. Random error can also occur when two or more treatments are compared with each other; the observed difference may be different from the expected difference ('true difference') just by chance. As a result, you might make false positive conclusions ('alpha error' or 'Type I error') or false negative conclusions ('beta error' or 'Type II error'). However, as you increase the sample size these errors decrease.

The general value of alpha error used in clinical trials is 0.05. This corresponds to the concept of 95% CI, that is, the probability of making a false positive

conclusion that there is a treatment effect when in fact there is none, is 5% (in the absence of bias). If you want to decrease the probability of making false positive conclusions, you can study a larger sample, but that would involve increased resources (including human resources and time) and delay in applying a potentially effective treatment in the population.

In addition to a possibility of making false conclusions, there is also a possibility that the average population treatment was not 'null', but this was not observed in the trial which included few participants. In a trial with few participants, the 95% CI are wide. This is because we use a formula that includes the variability of response to an intervention (or comparator) among the participants and the number of participants to calculate the 95% CI (rather than repeating the trials several times to find out the intervals within which the treatment differences lie 95% of the times, which of course is impossible in the real world). The CI become wider with increased variability of response to a treatment among participants, and fewer participants. As a result, the 95% CI include the 'null' treatment effect even though the average population treatment was not 'null' because of the inclusion of fewer participants.

Now, we are faced with a conundrum. If we include too few participants, there is a possibility of arriving at wrong conclusions, but if we included too many participants, this would consume resources and delay a treatment that could improve health outcomes. To find just the right number of participants that will help limit the wrong conclusions but also does not consume more resources than necessary, sample size calculations are carried out. Sample size calculations provide the minimum number of participants who should be included in a trial to limit the wrong conclusions to an acceptable level. If the sample size is inadequate (number of participants in the trial is fewer than the minimum number of participants required to limit the wrong conclusions to an acceptable level), you can reach wrong conclusions. This principle also applies to systematic reviews.

We have been discussing acceptable limits of wrong conclusions. What are these acceptable limits? Generally, an alpha error of 0.05 and a beta error of 0.10 or 0.20 (which corresponds to a power of 0.90 or 0.80, respectively) are used. Lower levels, say an alpha error of 0.01 and a beta error of 0.05, will increase the sample size compared to the usual levels used in clinical trials. In addition to the alpha and beta errors, the anticipated treatment effect (which should be at least the clinically important effect) and a measure of the variability of response to a treatment are required to perform the sample size calculations.

Sample size calculations can be performed using statistical software. If you provide the clinically important treatment differences, EQUAL-SR software can automatically perform sample size calculations for alpha error of 0.05 and beta error of 0.10 and 0.20 using the information extracted. For time-to-event

outcomes, the proportion of people who developed the outcome is necessary for sample size calculations.

Certainty of evidence

Having learnt all the concepts, we are now in a position to make a clinical interpretation of the effect of an intervention versus comparator for the outcome; that is, this is an outcome-level assessment rather than overall assessment of all the outcomes. We recommend that certainty of evidence is assessed for all outcomes and analysis that will help in shared decision making. This is most likely to be the primary outcomes in the systematic review but may contain important secondary outcomes. The choice of whether you should use the main analysis results or subgroup analysis results is described in the section 'Interpretation of the subgroup analyses and metaregression'. If per-protocol effect is of importance to patients in shared decision making, you should assess the certainty of evidence for the per-protocol effect and present those results in addition to the intention-to-treat effect.

The GRADE Working Group developed a series of publications describing how to assess the certainty of evidence. [34–42] Please see these for detailed information. We provide a broad overview of how to assess the certainty of evidence. For the most up-to-date guidance, please see www.gradepro.org/.

Overall, there are four categories of certainty of evidence with different levels of confidence in the results.

- High certainty: You are very confident that the average population treatment effect is close to the treatment effect found in the systematic review for the outcome.
- Moderate certainty: You are moderately confident that the average population treatment effect is close to the treatment effect found in the systematic review for the outcome, but there is a possibility that the average population treatment effect is substantially different from the treatment effect found in the systematic review for the outcome.
- Low certainty: You have limited confidence in the treatment effect found in the systematic review for the outcome. The average population treatment effect may be substantially different from the treatment effect found in the systematic review for the outcome.
- Very low certainty: You have very little confidence in the treatment effect found in the systematic review for the outcome. The average population treatment effect is likely to be substantially different from the treatment effect found in the systematic review for the outcome.

Broadly speaking, you start with the 'high certainty evidence' for RCTs. For non-randomised studies, if ROBINS-I is used for the assessment of risk of bias, then

you start with 'high certainty evidence'; otherwise, you start with 'low certainty evidence'. [43] The certainty of evidence is downgraded based on the observations on five aspects of the evidence. There are also three aspects based on which you can upgrade the certainty of evidence.

Figure 7.10 provides the aspects of the evidence which are considered for determining the uncertainty of evidence.

Table 7.3 provides some rough guidance on downgrading and upgrading the certainty of evidence. It also provides information on the assessments and analyses that can provide the information.

SUMMARY OF FINDINGS TABLE

GRADE suggests summarising evidence in succinct, transparent and informative summary of findings tables that show the quality of evidence and the magnitude of relative and absolute effects for each important outcome and/or as evidence profiles that provide, in addition, detailed information about the reason for the quality of evidence rating.

Summary of findings is a table that provides the relative and absolute effects (that is, ratios if appropriate and converting them to differences) along with the certainty of evidence for each important outcome. [34] For details of how to

Figure 7.10 Aspects taken into account for determining the uncertainty of evidence
Aspects in red downgrade the certainty of evidence; the aspects in green upgrade the certainty of evidence.

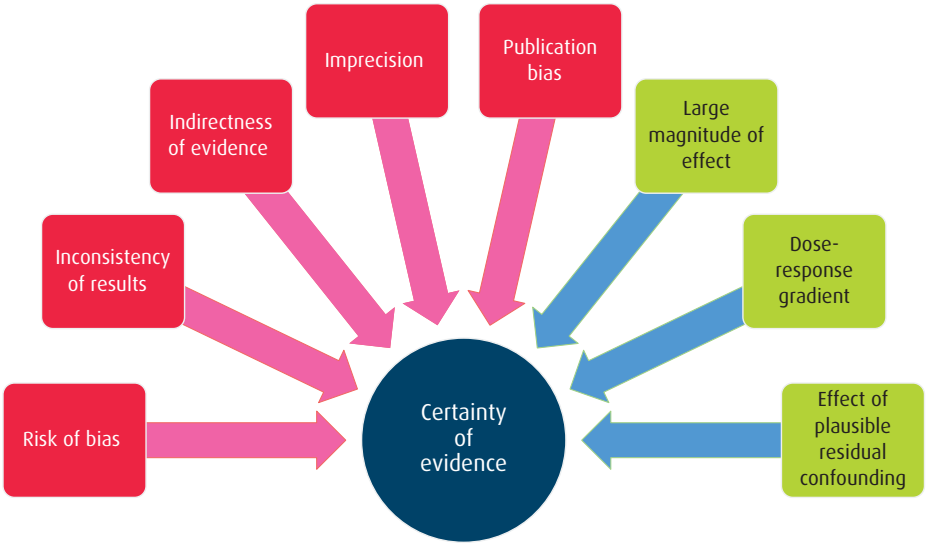


Table 7.3 Summary of domains in GRADE

Aspect	Maximum levels of downgrade or upgrade	Criteria for downgrading or upgrading	Sections that guide in assessment	Additional comments
Risk of bias	2	No downgrading: Most information is from low risk of bias studies or at moderate risk of bias due to unclear information. Downgrade one level: Most information is from moderate risk of bias studies with limitations, or most information is from studies with high or serious risk of bias in one domain. Downgrade two levels: Most information is from high or serious risk of bias with high or serious risk of bias in multiple domains or critical risk of bias in non-randomised studies.	Outcome-level risk of bias assessments	EQUAL-SR software can provide the contribution of studies to the meta-analysis in the different categories of risk of bias.
Inconsistency of results	2	No downgrading: No evidence of heterogeneity, heterogeneity in magnitude of effect but unlikely to change treatment decision. Downgrade one level: Heterogeneity in magnitude of effect with potential to change treatment decision. Downgrade two levels: Heterogeneity in direction of effect.	Heterogeneity assessment	
Indirectness of evidence	2	No downgrading: No concerns about applicability to patients and the outcome is not a surrogate outcome. Downgrade one level: Some concerns that the trial participants may not be similar to the patients in clinic or implementation of the intervention in the setting, surrogate outcomes where there is some evidence that they lead to clinical outcomes. Downgrade two levels: Major concerns that the trial participants may not be similar to the patients in clinic or implementation of the intervention in the setting, surrogate outcomes where there is some evidence that they lead to clinical outcomes.	Applicability to patients, Surrogate outcomes (Chapter 2)	

(*Continued*)

Table 7.3 Summary of domains in GRADE (*Continued*)

Aspect	Maximum levels of downgrade or upgrade	Criteria for downgrading or upgrading	Sections that guide in assessment	Additional comments
Imprecision	2	No downgrading: Adequate sample size and treatment effect suggests clinically important differences (clinically important increase or decrease scenarios in section 'Clinical interpretation of results') or rules out clinically important differences (no clinically important effect scenarios in section 'Clinical interpretation of results'). If sample size is inadequate, at least 2,000 to 4,000 participants for binary outcomes is acceptable. If standardised mean difference was used, approximately 400 participants may be acceptable. Downgrade one level: Inadequate sample size (or alternate sample sizes mentioned above) or clinically important differences cannot be ruled out (no evidence of clinically important effect scenarios in section 'Clinical interpretation of results'). Downgrade two levels: Inadequate sample size (or alternate sample sizes mentioned above) and clinically important differences cannot be ruled out (no evidence of clinically important effect scenarios in section 'Clinical interpretation of results').	Clinically important differences, sample size calculations	EQUAL-SR software can provide information on sample size, whether the meta-analysis included the sufficient sample size, whether the CI excluded no effect and clinically important effects, although for time-to-event outcomes you need to provide the number of events in the comparator group at the specified time point of interest from the studies in which this information is available directly or can be estimated from other information such as proportion without the event. EQUAL-SR software can help with these estimations.

Publication bias	1	No downgrading: Low risk of reporting bias. Downgrade one level: Some concerns or high risk of reporting bias.	Assessment of reporting bias	EQUAL-SR software can provide the reporting bias if you upload the signalling questions of ROB-ME tool.
Large magnitude of effect	2	Upgrade one level: lower 95% CI of risk ratio (or hazard ratio) are more than 2 or upper 95% CI are less than 0.5 and there is no confounding. Upgrade two levels: lower 95% CI of risk ratio (or hazard ratio) are more than 5 or upper 95% CI are less than 0.2 and there are no serious problems with risk of bias.	Meta-analytical results	EQUAL-SR software can provide information on whether the magnitude of effect was large or very large.
Dose-response gradient	1	Upgrade one level: If increased doses of intervention are associated with greater treatment effect.	Subgroup analysis and metaregression	
Effect of plausible residual confounding	1	Upgrade one level: If confounding is against the intervention.	Outcome-level risk of bias assessments	
Overall certainty	–	Each level of downgrading decreases the category from high certainty by 1 and each level of upgrading improves the category from the existing category by 1. The lowest category is very low certainty of evidence.		

create a summary of findings table using GRADEpro software (software developed by the GRADE Working Group), please refer to www.gradepro.org/.

EQUAL-SR software automatically converts any ratios to differences using the information used for meta-analysis, although for time-to-event outcomes you need to provide the number of events in the comparator group at the specified time point of interest from the studies in which this information is available directly or can be estimated from other information such as proportion without the event. EQUAL-SR software can help with these estimations.

ARRIVING AT CONCLUSIONS

Arriving at conclusions about the effectiveness of an intervention versus comparator depends upon following major aspects.

1. Whether the treatment effect of an intervention versus comparator is clinically important for each outcome.
2. Certainty of evidence for each outcome contributing to shared decision making.
3. Relative importance of the outcome.
4. Biological plausibility.

While the first two aspects are covered in the sections 'Clinically important differences' and 'Certainty of evidence', the next two require explanation.

Relative importance of the outcome

You have to review the treatment effect of an intervention versus comparator along with certainty of evidence for each outcome. The different possible scenarios in relation to the relative importance of outcomes are as follows.

- Clinically important benefits without clinically important harms.
- Clinically important harms without clinically important benefits.
- Clinically important benefits for some outcomes and clinically important harms for others.
- No clinically important benefits or harms.

Biological plausibility

In RCTs, the likely reason for conducting the trial is because of potential mechanisms by which an intervention improves outcomes. However, in non-randomised studies, biological plausibility of the association should be considered before arriving at conclusions. This is likely to have been considered as a part of

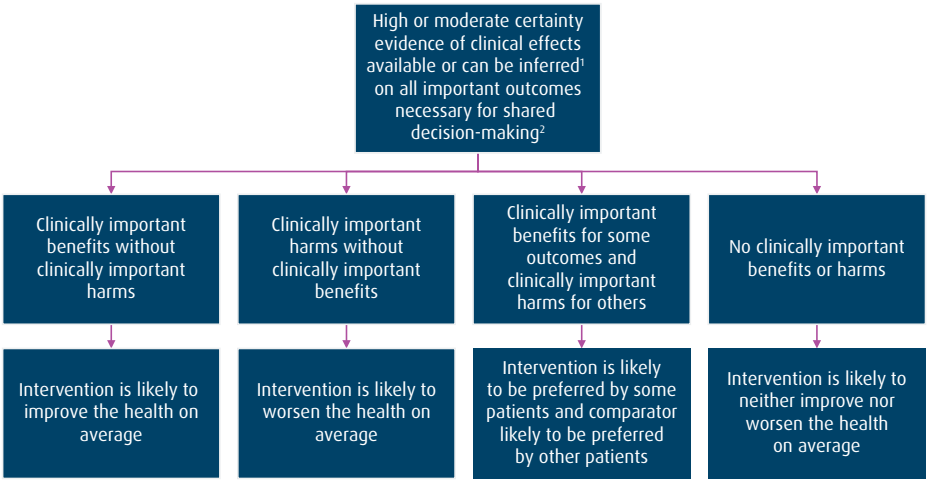
assessing the risk of bias due to confounding, but you might want to reconsider this while arriving at conclusions, particularly in light of the information gathered from the other studies and the results of the analysis.

Implications for clinical practice

Figures 7.11, 7.12, 7.13 and 7.14 show how to arrive at implications on clinical practice based on the information available on the clinical importance of treatment effect and certainty of evidence of each outcome, and the relative importance of the outcomes.

When there are clinically important benefits for some outcomes and clinically important harms for other outcomes due to the intervention, sometimes it is possible to find systematic reviews of patients' preferences and values, for example, León-García et al., [44] which can help with determining whether the intervention improves or decreases the average health of the patients in addition to the statement on some patients preferring the intervention and some preferring the comparator.

Figure 7.11 Relative importance of outcomes: high or moderate certainty evidence of clinically important effects available on all important outcomes

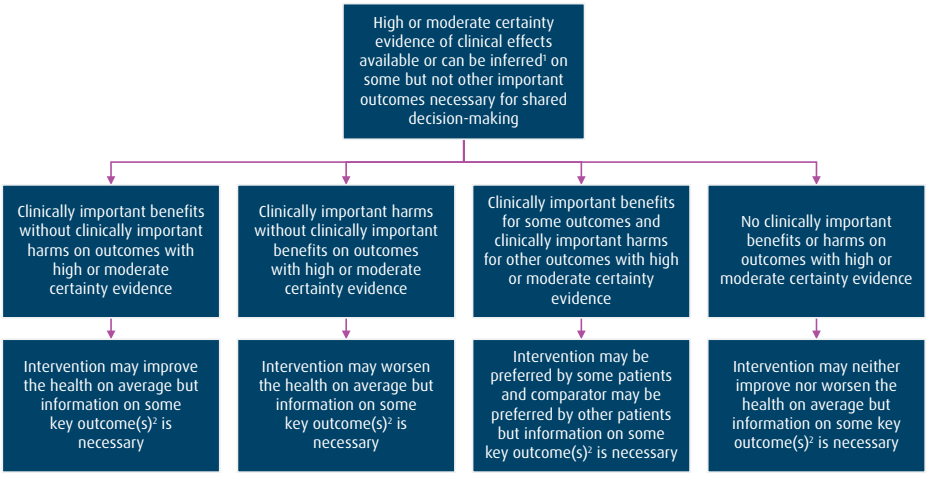


Notes:

[1] Some important outcomes can be inferred from the nature of the intervention and other important outcomes. For example, if the intervention was considerably more invasive than the comparator and results in increased complications, it is unlikely that the short-term HRQoL improves with the intervention.

[2] It is unlikely but not impossible that you find a combination of moderate certainty evidence with 'No evidence of clinically important difference' scenarios in Figure 7.9. However, if you do find such scenarios, you should use conclusions suggested in Figure 7.12.

Figure 7.12 Relative importance of outcomes: high or moderate certainty evidence available on some but not all important outcomes



Notes:

[1] Some important outcomes can be inferred from the nature of the intervention and other important outcomes. For example, if the intervention was considerably more invasive than the comparator and results in increased complications, it is unlikely that the short-term HRQoL improves with the intervention.

[2] You should state the key outcome(s) for which information is necessary.

Figure 7.13 Relative importance of outcomes: only low certainty evidence available on some or all important outcomes

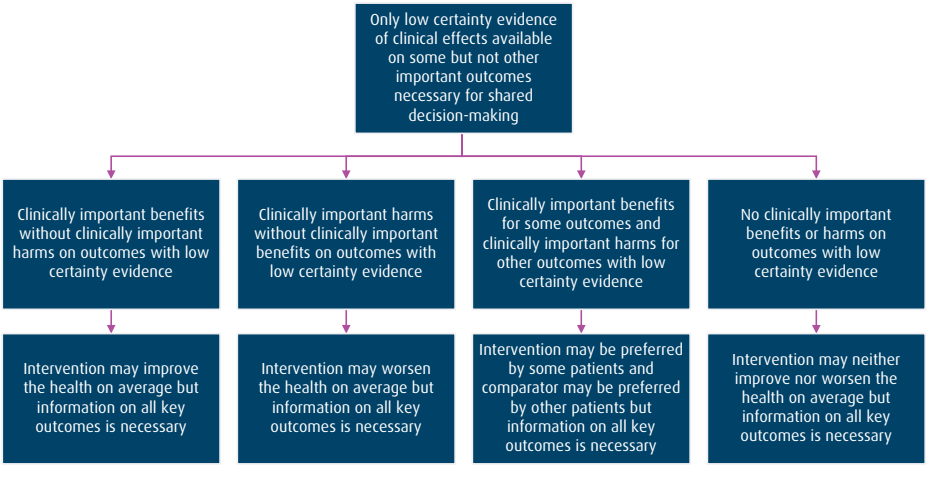
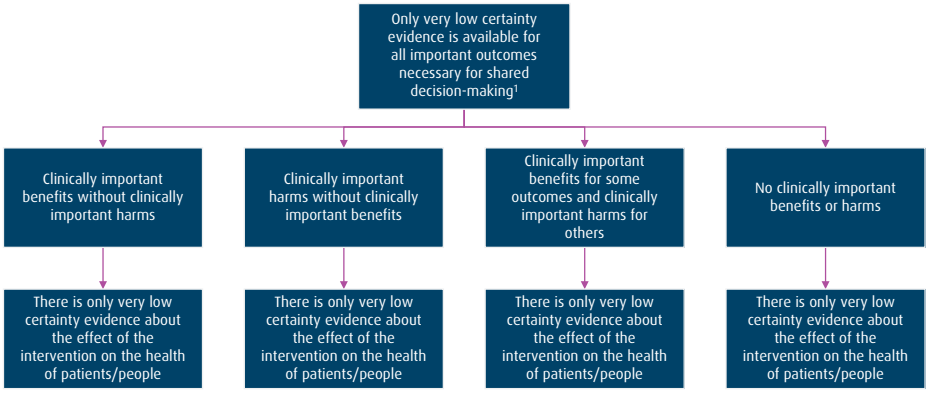


Figure 7.14 Relative importance of outcomes: only very low certainty evidence available on all important outcomes



Note:

[1] When there is only very low certainty evidence available for all the important outcomes necessary for shared decision making, the conclusion is 'There is only very low certainty evidence about the effect of the intervention on the health of patients/people'.

When formulating the implications for clinical practice, it is important to consider and highlight whether additional expertise or financial resources are required to implement the intervention. The implications for clinical practice should not be changed solely on the basis of lack of expertise or financial resources required to implement the intervention.

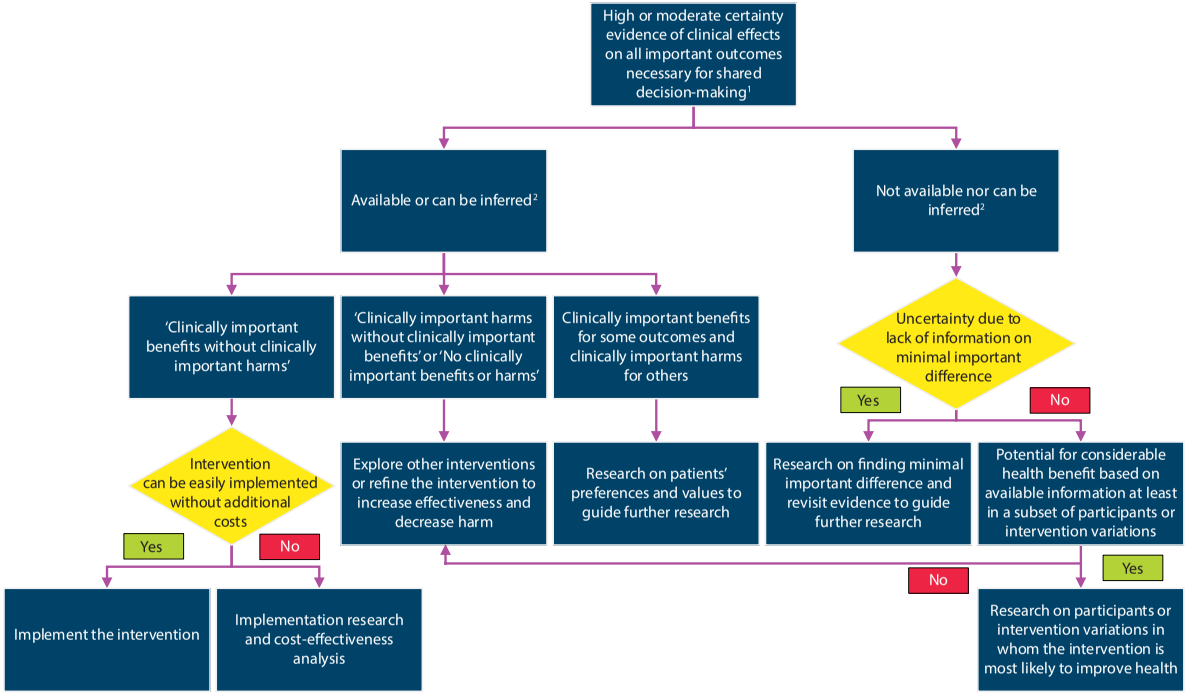
Implications for research

Figure 7.15 shows how to arrive at implications for research based on the information available on the clinical importance of treatment effect and certainty of evidence of each outcome, and the relative importance of the outcomes.

CLINICAL GUIDELINE RECOMMENDATIONS

One of the primary motives for completing a systematic review might be to develop clinical recommendations. The GRADE methodology can be used to determine the strength of the recommendation. [45] Certainty of evidence is one of the four domains that determine the strength of the recommendation, the other three domains being balance between desirable and undesirable outcomes (which is a similar concept to relative importance of outcomes), confidence in values and preferences and variability, and resource use. The confidence in

Figure 7.15 Implications for research



Notes:

[1] It is unlikely but not impossible that you find a combination of moderate certainty evidence with 'No evidence of clinically important difference' scenarios in Figure 7.9. However, if you do find such scenarios, you should use the 'Not available nor can be inferred' pathway.

[2] Some important outcomes can be inferred from the nature of the intervention and other important outcomes. For example, if the intervention is considerably more invasive than the comparator and results in increased complications, it is unlikely that the short-term HRQoL improves with the intervention.

values and preferences and variability involves additional systematic reviews similar to that of León-García et al. [44] Resource use involves availability of expertise (if applicable) and cost-effectiveness analysis which is beyond the scope of this book.

SUMMARY

In this chapter, we have provided guidance on how to interpret the results of meta-analysis.

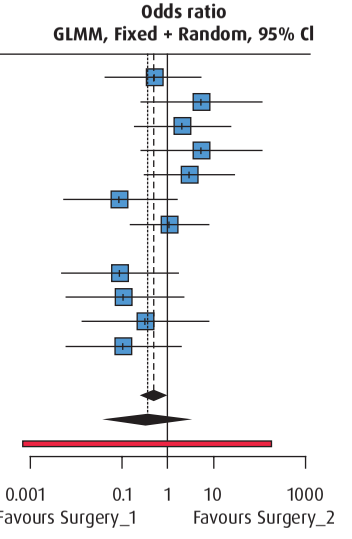
PRACTICE QUESTIONS

1. Interpret the forest plot shown in Figure 7.16.

Figure 7.16 Short-term mortality

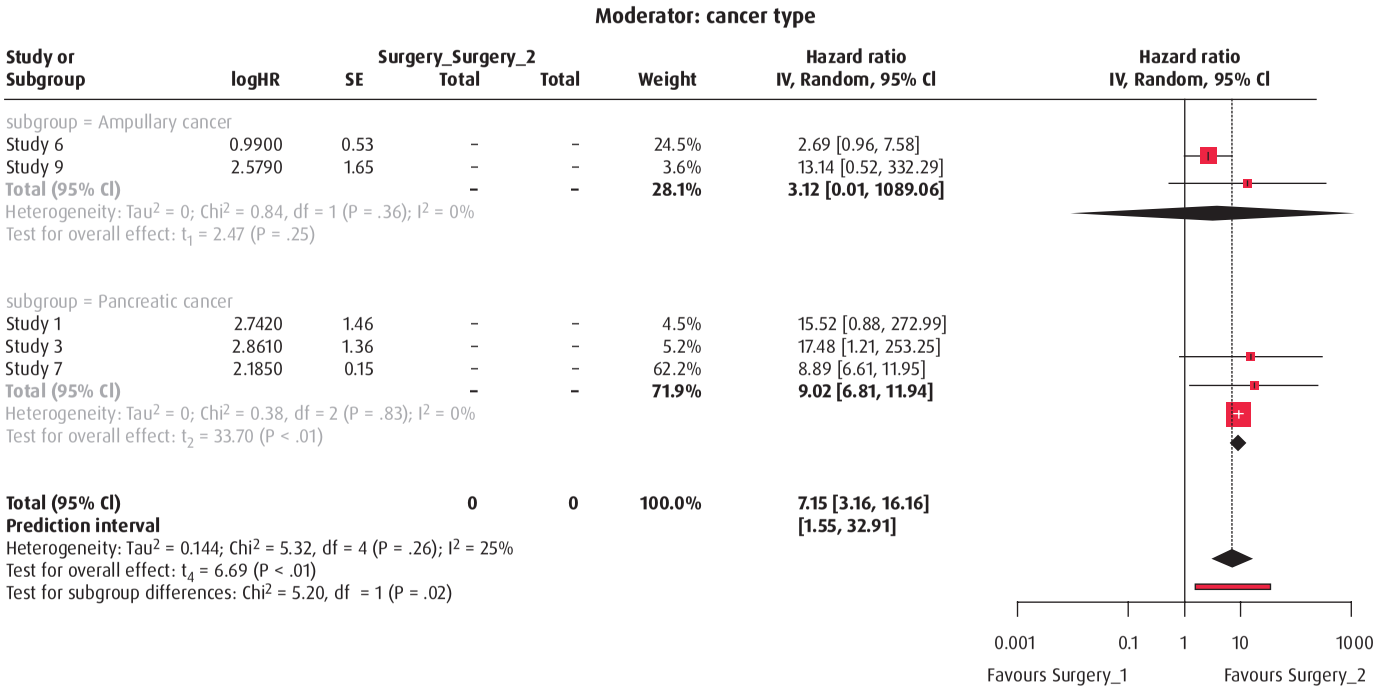
Study	Surgery_1 Events	Surgery_1 Total	Surgery_2 Events	Surgery_2 Total	Odds ratio GLMM, Fixed + Random, 95% CI
Study 1	1	261	2	258	0.49 [0.04, 5.46]
Study 2	2	123	0	132	5.45 [0.26, 114.71]
Study 3	2	248	1	254	2.06 [0.19, 22.83]
Study 4	2	170	0	180	5.36 [0.26, 112.38]
Study 5	3	181	1	178	2.98 [0.31, 28.95]
Study 6	0	260	5	257	0.09 [0.00, 1.60]
Study 7	2	129	2	139	1.08 [0.15, 7.77]
Study 8	0	102	0	107	
Study 9	0	86	5	89	0.09 [0.00, 1.63]
Study 10	0	172	4	179	0.11 [0.01, 2.12]
Study 11	0	298	1	292	0.33 [0.01, 8.02]
Study 12	0	209	4	202	0.11 [0.01, 1.97]
Total (fixed effect, 95% CI)		**2239**		**2267**	**0.49 [0.25, 0.97]**
Total (random effect, 95% CI)					**0.35 [0.04, 3.32]**
Prediction interval					**[0.00, 173.94]**

Heterogeneity: $Tau^2 = 6.519$; $Chi^2 = 1.30$, df = 10 (P = 1.00); $I^2 = 0\%$
Test for overall effect (fixed effect): Z = −2.04 (P = .04)
Test for overall effect (random effects): $t_{10} = -1.04$ (P = .32)



2. Interpret the results of subgroup analysis shown in the forest plot in Figure 7.17. Assume that it was hypothesised that Surgery_1 would be more beneficial (compared to Surgery_2) in participants with pancreatic cancer than ampullary cancer. This was one of the three planned subgroup analyses.

Figure 7.17 Subgroup analysis: long-term mortality



3. Interpret the results of sensitivity analysis in the forest plots shown in Figures 7.18 and 7.19. The first plot shows the analysis that included studies in which the mean or standard deviation were imputed. The second plot shows the analysis that excluded such studies. The MID was 1 day for length of hospital stay.

Figure 7.18 Sensitivity analysis: hospital stay (all studies included)

(days)

Study	Surgery_1 Mean	Surgery_1 SD	Surgery_1 Total	Surgery_2 Mean	Surgery_2 SD	Surgery_2 Total	Weight (fixed)	Weight (random)	Mean difference IV, Fixed + Random, 95% CI
Study 1	12.30	11.80	261	15.80	13.50	258	10.9%	8.9%	−3.50 [−5.68, −1.32]
Study 2	15.10	7.00	123	16.00	11.50	132	9.7%	8.8%	−0.90 [−3.22, 1.42]
Study 3	12.20	15.40	248	12.10	13.90	254	7.9%	8.6%	0.10 [−2.47, 2.67]
Study 4	14.30	13.80	170	10.20	8.90	180	8.7%	8.7%	4.10 [1.65, 6.55]
Study 5	17.70	15.60	181	16.50	15.90	178	4.9%	7.9%	1.20 [−2.06, 4.46]
Study 6	16.50	10.30	260	11.60	9.70	257	17.5%	9.3%	4.90 [3.18, 6.62]
Study 7	15.40	15.60	129	9.30	12.60	139	4.5%	7.8%	6.10[2.69, 9.51]
Study 8	17.20	15.60	102	12.80	15.90	107	2.8%	6.9%	4.40 [0.13, 8.67]
Study 9	15.60	10.70	86	16.10	14.70	89	3.6%	7.4%	−0.50 [−4.30, 3.30]
Study 10	12.50	15.40	172	10.60	15.90	179	4.8%	7.9%	1.90 [−1.37, 5.17]
Study 11	11.20	8.00	298	13.00	13.40	292	16.3%	9.2%	−1.80 [−3.59, −0.01]
Study 12	12.40	8.70	209	16.80	15.80	202	8.5%	8.7%	−4.40 [−6.88, −1.92]
Total (fixed effect, 95% CI)			**2239**			**2267**	**100.0%**	**-**	**0.62 [−0.11, 1.34]**
Total (random effect, 95% CI)							**-**	**100.0%**	**0.87 [−1.43, 3.17]**
Prediction interval									**[−6.89, 8.62]**

Heterogeneity: Tau^2 = 11.027; Chi^2 = 83.72, df = 11 (P < .01); I^2 = 87%
Test for overall effect (fixed effect): Z = 1.67 (P = .09)
Test for overall effect (random effects): t_{11} = 0.83 (P = .42)

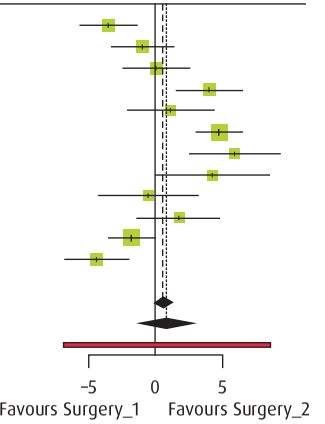
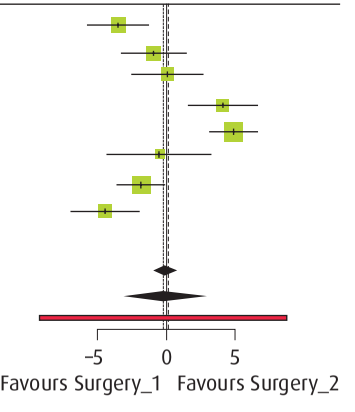


Figure 7.19 Sensitivity analysis: hospital stay (studies with imputation excluded)

(days)

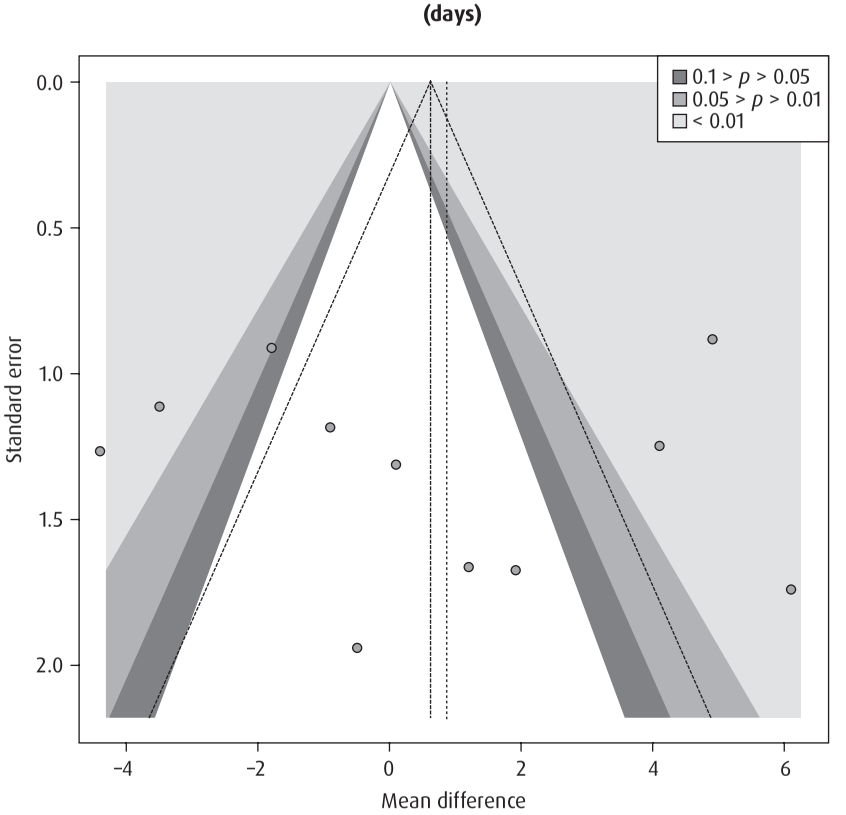
Study	Surgery_1 Mean	Surgery_1 SD	Surgery_1 Total	Surgery_2 Mean	Surgery_2 SD	Surgery_2 Total	Weight (fixed)	Weight (random)	Mean difference IV, Fixed + Random, 95%, CI
Study 1	12.30	11.80	261	15.80	13.50	258	13.2%	12.8%	−3.50 [−5.68, −1.32]
Study 2	15.10	7.00	123	16.00	11.50	132	11.6%	12.7%	−0.90 [−3.22, 1.42]
Study 3	12.20	15.40	248	12.10	13.90	254	9.5%	12.4%	0.10 [−2.47, 2.67]
Study 4	14.30	13.80	170	10.20	8.90	180	10.5%	12.5%	4.10 [1.65, 6.55]
Study 6	16.50	10.30	260	11.60	9.70	257	21.1%	13.3%	4.90 [3.18, 6.62]
Study 9	15.60	10.70	86	16.10	14.70	89	4.3%	10.7%	−0.50 [−4.30, 3.30]
Study 11	11.20	80.0	298	13.00	13.40	292	19.7%	13.2%	−1.80 [−3.59, −0.01]
Study 12	12.40	8.70	209	16.80	15.80	202	10.2%	12.5%	−4.40 [−6.88, −1.92]
Total (fixed effect, 95% CI)			**1655**			**1664**	**100.0%**	**—**	**0.08 [−0.71, 0.87]**
Total (random effect, 95% CI)							**—**	**100.0%**	**−0.23 [−3.27, 2.82]**
Prediction interval									**[−9.17, 8.72]**

Heterogeneity: $Tau^2 = 11.705$; $Chi^2 = 68.30$, df = 7 ($P < .01$); $I^2 = 90\%$
Test for overall effect (fixed effect): Z = 0.20 (P = .84)
Test for overall effect (random effect): $t_7 = -0.18$ (P = .87)



4. Assess whether there is funnel plot asymmetry and if present whether it is due to reporting bias or other reasons in the funnel plot shown in Figure 7.20.

Figure 7.20 Length of hospital stay (funnel plot)

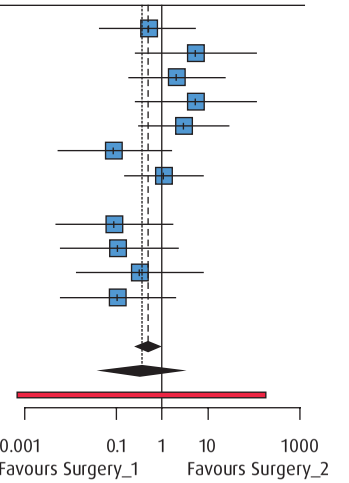


5. Look at Figure 7.21. Assess whether the CI of the random-effects model suggest clinically important increase, clinically important decrease, no evidence of clinically important difference, or no clinically important difference. A 15% relative change in odds of mortality was considered clinically important.

Figure 7.21 Short-term mortality (clinical interpretation)

Study	Surgery_1 Events	Surgery_1 Total	Surgery_2 Events	Surgery_2 Total	Odds ratio GLMM, Fixed + Random, 95% CI
Study 1	1	261	2	258	0.49 [0.04, 5.46]
Study 2	2	123	0	132	5.45 [0.26, 114.71]
Study 3	2	248	1	254	2.06 [0.19, 22.83]
Study 4	2	170	0	180	5.36 [0.26, 112.38]
Study 5	3	181	1	178	2.98 [0.31, 28.95]
Study 6	0	260	5	257	0.09 [0.00, 1.60]
Study 7	2	129	2	139	1.08 [0.15, 7.77]
Study 8	0	102	0	107	
Study 9	0	86	5	89	0.09 [0.00, 1.63]
Study 10	0	172	4	179	0.11 [0.01, 2.12]
Study 11	0	298	1	292	0.33 [0.01, 8.02]
Study 12	0	209	4	202	0.11 [0.01, 1.97]
Total (fixed effect, 95% CI)		**2239**		**2267**	**0.49 [0.25, 0.97]**
Total (random effect, 95% CI)					**0.35 [0.04, 3.32]**
Prediction interval					**[0.00, 173.94]**

Heterogeneity: $Tau^2 = 6.519$; $Chi^2 = 1.30$, df = 10 (P = 1.00); $I^2 = 0\%$
Test for overall effect (fixed effect): Z = −2.04 (P = .04)
Test for overall effect (random effects): $t_{10} = -1.04$ (P = .32)



6. For the answer to Question 5, assess the certainty of evidence, stating any reasons for downgrading or upgrading. For this question consider that the studies were all RCTs at moderate risk of bias which could affect the estimates and that there were no concerns about applicability to patients. There was no evidence of reporting bias. Sample size calculations were not available.

Answers to the practice questions can be found in the Appendix.

REFERENCES

1. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009; 172(1):137–59.
2. Deeks JJ, Higgins JP, Altman DG. Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, Welch VA (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 6.4 (updated October 2023). Cochrane, 2023. Available from: www.training.cochrane.org/handbook. (accessed on 25 March 2024).
3. Merriam-Webster Dictionary. Null. www.merriam-webster.com /dictionary/null (accessed on 29 March 2024). 2024.
4. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003; 327(7414):557–60.
5. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019; 366: l4898.
6. 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. https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool /current-version-of-rob-2 (accessed on 14 March 2024). 2019.
7. Sterne J, Higgins J, Elbers R, Reeves B, Development group for ROBINS-I. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance. www.riskofbias.info (accessed on 10 September 2023). 2016.
8. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016; 355: i4919.
9. Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials:

systematic review of meta-epidemiological studies. PLoS One. 2016; 11(7): e0159267.
10. Savović J, Turner RM, Mawdsley D, Jones HE, Beynon R, Higgins JPT, et al. Association between risk-of-bias assessments and results of randomized trials in Cochrane reviews: the ROBES meta-epidemiologic study. Am J Epidemiol. 2018; 187(5):1113–22.
11. Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG. Bias modelling in evidence synthesis. J R Stat Soc Ser A Stat Soc. 2009; 172(1):21–47.
12. Rhodes KM, Savović J, Elbers R, Jones HE, Higgins JPT, Sterne JAC, et al. Adjusting trial results for biases in meta-analysis: combining data-based evidence on bias with detailed trial assessment. J R Stat Soc Ser A Stat Soc. 2020; 183(1):193–209.
13. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. J R Stat Soc Ser A Stat Soc. 2008; 172(1):119–36.
14. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ. 2020; 192(32): e901–e906.
15. Page MJ, Sterne JAC, Boutron I, Hróbjartsson A, Kirkham JJ, Li T, et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. BMJ. 2023; 383: e076754.
16. Page MJ, Sterne JAC, Boutron I, Hróbjartsson A, Kirkham JJ, Li T, et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. www.riskofbias.info/welcome/rob-me-tool (accessed on 28 March 2024). 2023.
17. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997; 315(7109):629–34.
18. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011; 343: d4002.
19. Van Noorden R. Medicine is plagued by untrustworthy clinical trials. How many studies are faked or flawed? Nature. 2023; 619(7970):454–8.
20. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000; 53(11):1119–29.
21. ICJME. Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. www.icmje.org/news-and-editorials/clin_trial_sep2004.pdf (accessed on 26 March 2024). 2004.
22. COMET Initiative. Core Outcome Measures in Effectiveness Trials. www.comet-initiative.org/ (accessed on 10 September 2023). 2023.

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

24. Lundh A, Boutron I, Stewart L, Hróbjartsson A. What to do with a clinical trial with conflicts of interest. BMJ Evidence-Based Medicine. 2020; 25(5):157–8.

25. TACIT. Tool for Addressing Conflicts of Interest in Trials. www.tacit.one/ (accessed on 28 March 2024). 2024.

26. Schünemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the Chronic Respiratory disease Questionnaire (CRQ). Copd. 2005; 2(1):81–9.

27. Devji T, Carrasco-Labra A, Guyatt G. Mind the methods of determining minimal important differences: three critical issues to consider. Evid Based Ment Health. 2021; 24(2):77–81.

28. Franceschini M, Boffa A, Pignotti E, Andriolo L, Zaffagnini S, Filardo G. The minimal clinically important difference changes greatly based on the different calculation methods. Am J Sports Med. 2023; 51(4):1067–73.

29. PROMID. PROMID. www.promid.org/ (accessed on 29 March 2024). 2024.

30. Devji T, Guyatt GH, Lytvyn L, Brignardello-Petersen R, Foroutan F, Sadeghirad B, et al. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid Recommendations. BMJ Open. 2017; 7(5): e015587.

31. Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC, et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. BMJ Open. 2019; 9(2): e028777.

32. Tan YY, Papez V, Chang WH, Mueller SH, Denaxas S, Lai AG. Comparing clinical trial population representativeness to real-world populations: an external validity analysis encompassing 43 895 trials and 5 685 738 individuals across 989 unique drugs and 286 conditions in England. Lancet Healthy Longev. 2022; 3(10): e674–e689.

33. Ingrasciotta Y, Spini A, L'Abbate L, Fiore ES, Carollo M, Ientile V, et al. Comparing clinical trial population representativeness to real-world users of 17 biologics approved for immune-mediated inflammatory diseases: an external validity analysis of 66,639 biologic users from the Italian VALORE project. Pharmacol Res. 2024; 200:107074.

34. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011; 64(4):383–94.

35. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011; 64(4):395–400.
36. 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):401–6.
37. 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):407–15.
38. 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):1277–82.
39. 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):1283–93.
40. 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):1294–302.
41. 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):1303–10.
42. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011; 64(12):1311–16.
43. Schünemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol. 2019; 111:105–14.
44. León-García M, Humphries B, Maraboto A, Rabassa M, Boehmer KR, Perestelo-Perez L, et al. Women's values and preferences on low-molecular-weight heparin and pregnancy: a mixed-methods systematic review. BMC Pregnancy Childbirth. 2022; 22(1):747.
45. Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation—determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013; 66(7):726–35.

8 Report the findings

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- *Report the findings of the systematic review*
- *Complete the protocol of a systematic review*

OVERVIEW

The reporting of the systematic review should follow the PRISMA guidance.

There are multiple extensions for PRISMA covering different aspects. These extensions should be used in conjunction with the PRISMA statement and are not standalone statements. PRISMA-S (PRISMA literature search extension) should be used routinely in all systematic reviews. When no meta-analysis is performed in intervention reviews, the Synthesis Without Meta-analysis (SWiM) guideline should be used in conjunction with PRISMA.

The remaining extensions depend on the focus or type of the systematic review. For systematic reviews focusing on interventions in disadvantaged populations or aimed at reducing social gradients across populations, PRISMA-Equity extension (PRISMA-E 2012) should be used. When systematic reviews focusing on harms of an intervention are performed, PRISMA-harms should be used. For systematic reviews focusing on acupuncture, use the PRISMA extension for acupuncture (PRISMA-A).

When performing scoping reviews, use the PRISMA extension for scoping reviews (PRISMA-ScR). In terms of intervention reviews, when a network meta-analysis is performed, PRISMA-NMA should be used. For individual participant data meta-analyses, use PRISMA-IPD.

The general structure of reporting includes a title, abstract, rationale for the review, objectives of the review, methods of the review (including study eligibility, study selection, data extraction, data analysis, interpretation of data), results of the review including results of the search, study characteristics, effect estimates, subgroup analysis, sensitivity analysis, reporting bias, discussion

around the clinical context of the results, reliability of the results, and implications of the results for practice and research.

The protocol registration should be completed prior to commencement of a systematic review.

INTRODUCTION

You may recollect from Chapter 2 that once the research question is determined, the next steps are to identify suitable methods for study selection, data extraction, analysis and reporting, and to complete protocol registration. In the previous chapter, we described the suitable methods for interpretation of the review.

The next step is to describe the suitable methods for reporting a systematic review.

WHY SHOULD YOU USE SUITABLE METHODS FOR REPORTING A SYSTEMATIC REVIEW?

One of the key characteristics of a systematic review is that it has an explicit, reproducible methodology. [1] This means that some other researcher should be able to arrive at the same findings if they performed the same steps used in performing the systematic review. This is only possible if the reporting is adequate.

REPORTING GUIDELINES FOR SYSTEMATIC REVIEW

The reporting of the systematic review should follow the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidance. [2,3]

There are multiple extensions for PRISMA covering different aspects. These extensions should be used in conjunction with the PRISMA statement. PRISMA-S (PRISMA literature search extension) [4] should be used routinely in all systematic reviews. When no meta-analysis is performed in intervention reviews, the Synthesis Without Meta-analysis (SWiM) guideline [5] should be used in conjunction with PRISMA.

The remaining extensions depend on the focus or type of the systematic review. For systematic reviews focusing on interventions in disadvantaged populations or aimed at reducing social gradients across populations, the PRISMA-Equity extension (PRISMA-E 2012) should be used. [6,7] When systematic reviews focusing on harms of an intervention are performed, the PRISMA-harms

checklist should be used. [8] For systematic reviews focusing on acupuncture, use the PRISMA extension for acupuncture (PRISMA-A). [9]

When performing scoping reviews, use the PRISMA extension for scoping reviews (PRISMA-ScR). [10] In terms of intervention reviews, when a network meta-analysis (NMA) is performed, the PRISMA-NMA extension should be used. [11] For individual participant data meta-analyses (IPD), use the PRISMA-IPD extension. [12]

GENERAL STRUCTURE FOR REPORTING SYSTEMATIC REVIEW

The general structure of reporting includes a title, abstract, rationale for the review, objectives of the review, methods of the review (including study eligibility, study selection, data extraction, data analysis, interpretation of data), results of the review including results of the search, study characteristics, effect estimates, subgroup analysis, sensitivity analysis, reporting bias, discussion around the clinical context of the results, reliability of the results, and implications of the results for practice and research.

This general structure of reporting a systematic review can be linked to the general structure of reporting in a journal: title, abstract, main text which contains the background (or introduction), methods, results, discussion and conclusions, and references. In this chapter, we provide guidance for reporting in which tables and figures are interspersed with text. You might notice that we indicate tables and figures coming before the related text. This is because the information for the text comes from these tables and figures. However, while reporting, you will start with the text and provide links to the table and figures from the text. Some journal submissions will require the researcher to provide the tables and figures separately from the text. If that is the case, then you need to move the tables and figures.

Most of these items have been covered in previous chapters. Where this is the case, details of where they can be found in previous chapters have been provided. When these have not been covered in previous chapters, the details of how to report these items are provided.

We have focused on items related to a systematic review of interventions. For detailed guidance and explanations on why some information is required, please refer to the PRISMA statement, its extensions, or additional guidance used in special circumstances. [2–12]

Title

The title should provide key information about the objectives and identify the research as a systematic review. [2,3] It can include other information related to methods, for example, if meta-analysis has also been performed. [2,3]

For example, for the research objective *'To assess the benefits and harms of laparoscopic distal pancreatectomy versus open distal pancreatectomy for people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma of the body or tail of the pancreas, or both'*, the suggested title will be *'Laparoscopic distal pancreatectomy versus open distal pancreatectomy for people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma: a systematic review'*.

If the systematic review included only RCTs and included a meta-analysis, the title could be *'Laparoscopic distal pancreatectomy versus open distal pancreatectomy for people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma: a systematic review and meta-analysis of randomised controlled trials'*.

If the systematic review is an update of a previous systematic review, you should consider including this information in the title. [2,3,13] For example, if the above systematic review was an update of a previous systematic review, the title could be *'Laparoscopic distal pancreatectomy versus open distal pancreatectomy for people undergoing distal pancreatectomy for pancreatic ductal adenocarcinoma: an updated systematic review and meta-analysis of randomised controlled trials'*.

Abstract

The abstract of a systematic review should aim to contain the information provided in Table 8.1. [2,3] However, because of word count restrictions, it may not always be possible to include some aspects. As a minimum, we recommend reporting the results of major outcomes regardless of statistical significance or clinically important differences.

Main text

Background

The background section includes the rationale for the review and objectives of the review sections of the general structure of reporting a review. Table 8.2 provides the information that should be included in the background section of the main text.

Table 8.3 provides some common sources of information that can help with finding the information on the number of new cases per year or existing cases of a disease, health condition or the number of procedures performed in a year.

Of course, PubMed is a good source of such information. You can combine the 'population' domain of the formal search strategy with filters for incidence and prevalence available from the 'Epidemiological studies: Filters' subsection of the InterTASC Information Specialists' Sub-group Search Filter Resource. [14]

Methods

The methods section includes the study eligibility, study selection, data extraction, data analysis and interpretation of data sections of the general structure of

Table 8.1 Information to be included in the abstract

Identification as systematic review
Main objective
Major inclusion and exclusion criteria
Searches • Databases and trial registers searched • Date of last search
Methods used to assess risk of bias
Method used for analysis
Number of included studies and participants
Key information on studies • Risk of bias • Other important relevant characteristics, for example, study design, type of participants, variations in interventions, comparators, outcomes
Results for main outcomes *This should be done regardless of statistical significance or clinically important differences* • Number of studies and participants included in the analysis • Summary estimate and CI • Interpretation of the summary estimate and CI • Certainty of evidence
Overall interpretation and important implications
Primary source of funding for the review
Register name and registration number

Table 8.2 Information to be included in the main text (background)

- **Population (disease, health condition, procedure)**
  - Definition or description of the disease, health condition, or procedure (please see Chapter 3)
    - This is particularly important if there are many ways of defining a disease, health condition or a procedure
  - Magnitude of the disease or health condition
    - Example: number of new cases per year, number of existing cases (for chronic conditions), number of procedures (if the population is participants undergoing procedures)
  - Impact on people suffering from the health condition
    - Example: decreased longevity of life, decreased quality of life due to symptoms related to the health condition or its treatment, loss of productivity due to symptoms related to the health condition or its treatment, economic loss due to symptoms related to the health condition or its treatment
  - Mechanism of the impact on people
    - This can help the readers of the review understand how the disease, health condition or procedure causes ill-health in people and help them understand how the intervention and comparator (if active treatment) work
  - Current standard treatment

- **Intervention**
  - What is the proposed mechanism of action (why do people think that the intervention will work)?
  - Any evidence that it works in a different disease, health condition, or procedure
    - Example: in a review of routine antibiotics for a surgical procedure, you can include information about its effectiveness in a different surgical procedure
  - What are the known side effects due to the intervention?
- **Comparator**
  - Same details as intervention if the comparator is an active treatment
  - If the comparator is no treatment, why is no treatment being chosen as the comparator (for example, this might be the standard treatment)?
- **Why is this research important?**
  - Please see Chapter 3, section 'Choosing a research question of importance to stakeholders'
- **Research objectives**
  - Please see Chapter 3, section 'From "research question" to "objectives"'

Table 8.3 Common sources of epidemiological information

- Cancers
  - https://gco.iarc.fr/en
- Benign diseases and operations
  - www.cdc.gov/nchs/
  - https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics
  - www.odt.nhs.uk/statistics-and-reports/annual-activity-report/
  - www.srtr.org/

reporting a review. Table 8.4 provides the information that should be included in the methods section of the main text. These have been described in detail in the previous chapters of this book.

Results

The results section includes the results of the search, study characteristics, effect estimates, subgroup analysis, sensitivity analysis and reporting bias sections of the general structure of reporting a review. Table 8.5 provides the information that should be included in the results section of the main text. The table is self-explanatory. The chapter numbers in brackets indicate how or where the information can be found.

Discussion and conclusions

The discussion and conclusions section includes the discussion around the clinical context of the results, reliability of the results, and implications of the results for practice and research sections of the general structure of reporting a review. Table 8.6 provides the information that should be included in the discussion and conclusion section of the main text. The table is self-explanatory. The chapter numbers in brackets indicate where the information can be found.

Table 8.4 Information to be included in the main text (methods)

- **Selection criteria**
  - Please see Chapter 3, section 'From "objectives" to "eligibility criteria"'
- **Outcomes**
  - Please see Chapter 3, section 'From "objectives" to "outcomes"'
- **Identification of studies**
  - Please see Chapter 4, section 'Study selection process'
- **Risk of bias**
  - Please see Chapter 5, section 'Risk of bias'
- **Data extraction**
  - Please see Chapter 5, section 'What data should be extracted?'
- **Data analysis**
  - Please see Chapter 6
- **Reporting bias**
  - Please see Chapter 6, section 'Explore reporting bias'
- **Certainty of evidence**
  - Please see Chapter 7, section 'Certainty of evidence'

Table 8.5 Information to be included in the main text (results)

- Search results
  - Text (Chapter 4)
    - Number of records retrieved from each database
    - De-duplication of searches (including any software used for de-duplication)
    - Number of records excluded by screening titles and abstracts
    - Number of records excluded by review of full text and summary of reasons
    - Number of records included in the systematic review
    - Number of records of ongoing studies
  - PRISMA flow diagram
    - Information in text format is provided as a flow diagram
    - Template for flow diagram is available from www.prisma-statement.org/
- Characteristics of included studies (Chapter 5)
  - Table containing the following information[1]
    - Study design (if different study designs are included)
    - Inclusion and exclusion criteria of trial participants
    - Major participant characteristics
      - PROGRESS-PLUS framework for identifying disadvantaged people (Chapter 3)
      - Other potential effect modifiers
    - Intervention
    - Comparator
    - Number of participants in each group
    - Outcomes reported
  - Text summarising the information in the 'Characteristics of included studies' table
    - Total number of studies and participants
    - Number of arms (if the study is multi-armed)
    - Number of studies and participants in each study design (if different study designs are included)
    - Number of studies and participants in each comparison (if the systematic review has a broad focus)
    - Summary details about participants

- Risk of bias[2] (Chapter 5)
  - Risk of bias table
    - Classification of risk of bias for each domain and overall classification of risk of bias in each study for each outcome
  - Risk of bias summary (graph)
    - Proportion of studies at different risks of bias for each domain for each outcome
  - Risk of bias summary (text)
    - Proportion of studies at different risks of bias for each domain for each outcome, that is, the same information as risk of bias summary (graph) but in text format with details of which studies are included in each category
- Treatment effect for each analysis (table) (Chapter 6)
  - Number of studies and number of participants included in the meta-analysis
  - Method
  - Model
  - Treatment effect (summary estimates with 95% CI)
- Treatment effect for each analysis (forest plots) (Chapter 6)
- Treatment effect for each outcome (for each comparison if the review has broad focus) (text) (Chapters 6 and 7)
  - Number of studies and number of participants included in the outcome
  - Number of studies and number of participants included in the meta-analysis (if meta-analysis was performed)
  - Treatment effect
  - Heterogeneity
  - Reason for exclusion of any studies (if excluded from the meta-analysis)
  - Reason for not performing the meta-analysis (if not performed)
- Subgroup analysis and metaregression[3] (Chapters 6 and 7)
  - Forest plots
  - Text
    - Test for subgroup differences
    - Metaregression results
- Sensitivity analysis[3] (Chapters 6 and 7)
  - Any alteration in results because of assumptions (are the results robust or sensitive to assumptions?)
  - Any forest plots
- Reporting bias[3] (Chapters 6 and 7)
  - Funnel plots
  - Tables helping with ROB-ME assessments
  - Text
    - Test for funnel plot asymmetry
    - Risk of reporting bias
- Certainty of evidence (Chapter 7)
  - Summary of findings table
  - Text[4]

Notes:

[1] Depending upon the amount of information, you might consider separate tables for description of trial participants, intervention versus comparator, and outcomes. For example, if there was variation of participants, you might consider a table that contains information on whether different types of participants were included in the trial; if there was variation in intervention and comparator, you might provide a table that provides detailed description of the intervention and comparator; if there was variation in the outcome definition or scales, you might provide a separate table for the outcomes with varied definitions or scales.

[2] Outcomes with the same studies included and similar risk of bias in the different domains can be combined to avoid repeating the same information multiple times.

[3] This can be included under each outcome or can be provided as a different section depending on whether the results are similar or different for the outcomes.

[4] Certainty of evidence is included under each outcome along with the text description of treatment effect for each outcome.

Table 8.6 Information to be included in the main text (discussion and conclusion)

- Summary of main results (Chapter 6)
  - This is a synopsis of the results highlighting the major findings
- Interpretation of the results
  - This is the section where you discuss what the results mean. You might want to consider the following:
    - Interpretation of results (Chapter 7)
    - Alternative interpretations of results: this is guided by the assumptions made and what if those assumptions were not true
    - How does the evidence from the results of the systematic review and external evidence support your assumptions rather than the assumptions under alternative interpretations of results?
    - If the alternative interpretation of results has a high probability of being true, what is needed to support or refute the alternate interpretation? This can also help with being more precise about the implications of research.
- Reliability of the results
  - Strengths and weaknesses of the review process
    - Strengths
      - Efforts made to identify most studies on the topic by searching multiple databases, trial registers
      - Two researchers identifying studies and extracting data to minimise errors
      - Appropriateness of the methods used for analysis
    - Weaknesses
      - Any deviation from the planned protocol including the reasons for the deviation and the implications on the conclusions
  - Summarise the reasons for downgrading or upgrading the certainty of evidence (Chapter 7)
  - Comment on bias within the studies and reporting bias (although these are part of certainty of evidence) (Chapter 7)
  - If the results are unreliable, what additional information or changes to study design are necessary to make the results more reliable? Again, this can help with being more precise about the implications of research.
- Applicability of results
  - Applicability to patients (Chapter 7)
  - Is the intervention available globally?
  - Is there any special expertise or setting required to implement the intervention?
  - Could costs of the intervention or lack of resources be a factor in implementing the intervention?
- How the findings of this study compare to those of other studies, recommendations or policies on the topic
  - Are the findings, recommendations or policies similar or different?
  - If the findings, recommendations or policies on the topic are different, what could be the potential reasons for the differences, for example, newer studies, different methodology, different interpretation of results?
- Conclusions (Chapter 7)
  - Implications for clinical practice
  - Implications for research
    - If applicable, provide the guidance for further research using the PICO framework and study design

References

The citations used for any section of the report should be provided.

Additional information to be provided for reporting systematic reviews

In addition to the information above, you also need to provide information relating to the registration. This includes the registration number, link to a publicly available protocol, link to any information such as the data and code used for meta-analysis which are publicly available. [2,3] Some repositories that are publicly available and free for use by the researchers and public include Open Science Framework (https://osf.io/) and zenodo (https://zenodo.org/). Researchers can upload their protocol, data and codes in these repositories free and link them to the review, to allow replicability of findings by others.

REGISTRATION OF PROTOCOL

In the steps in performing a systematic review outlined in Chapter 2, registration of protocol is done after identifying methods for selecting studies, data extraction, analysis and reporting. Now that we have learnt the best methods for these, we can register the protocol.

What is a protocol?

In the context of systematic reviews and meta-analyses, the protocol is a document that presents an explicit plan for the review. [13]

What details should a protocol contain?

The protocol details the rationale and methodological and analytical approach of the review prior to performing the systematic review and meta-analysis. [13] PRISMA for systematic reviews protocol (PRISMA-P) should be followed for the protocol. [13,15] Broadly speaking, the sections covered in the background and methods are included in a protocol. Of course, in the protocol you detail how you will do things rather than how you did things.

What does registration of protocol mean?

Protocol registration means recording the protocol in a database such as PROSPERO (www.crd.york.ac.uk/prospero/). The PROSPERO registration involves answering questions related to the background and methods sections of the systematic review. If it is not possible to register the systematic review in PROSPERO, for example because the systematic review has no direct implications on the healthcare of the individual, you can consider making the protocol publicly

available via open access journals, Open Science Framework (https://osf.io/), zenodo (https://zenodo.org/), or similar repositories that can be accessed publicly.

When should the protocol registration be completed?

We have provided a description of the details about selecting studies, extracting data, analysing data, interpreting data and reporting the findings to make you aware of the methods to be used in the protocol. This does not mean that the registration of protocol is done after completing all the steps. The protocol should be registered prior to starting the systematic review; once the search strategy has been designed, the protocol should be registered.

SUBSEQUENT STEPS IN CONDUCTING THE SYSTEMATIC REVIEW

These have been covered in detail in Chapters 4, 5, 6 and 7.

SUMMARY

In this chapter, we have provided guidance on how to report a standard systematic review of intervention and complete a protocol.

Is this the end or the beginning?

This brings us to the end of this book. We hope that this is the beginning of your quest for knowledge.

REFERENCES

1. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, 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(7): e1000100.
2. 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: n71.
3. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021; 372: n160.

4. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA Statement for reporting literature searches in systematic reviews. Systematic Reviews. 2021; 10(1): 39.
5. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020; 368: l6890.
6. Welch V, Petticrew M, Petkovic J, Moher D, Waters E, White H, et al. Extending the PRISMA statement to equity-focused systematic reviews (PRISMA-E 2012): explanation and elaboration. J Clin Epidemiol. 2016; 70:68–89.
7. Welch V, Petticrew M, Tugwell P, Moher D, O'Neill J, Waters E, et al. PRISMA-Equity 2012 extension: reporting guidelines for systematic reviews with a focus on health equity. PLoS Med. 2012; 9(10): e1001333.
8. Zorzela L, Loke YK, Ioannidis JP, Golder S, Santaguida P, Altman DG, et al. PRISMA harms checklist: improving harms reporting in systematic reviews. BMJ. 2016; 352: i157.
9. Wang X, Chen Y, Liu Y, Yao L, Estill J, Bian Z, et al. Reporting items for systematic reviews and meta-analyses of acupuncture: the PRISMA for acupuncture checklist. BMC Complement Altern Med. 2019; 19(1):208.
10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018; 169(7):467–73.
11. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, 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(11):777–84.
12. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, et al. Preferred reporting items for a systematic review and meta-analysis of individual participant data: the PRISMA-IPD statement. JAMA. 2015; 313(16):1657–65.
13. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews. 2015; 4(1):1.
14. The InterTASC Information Specialists' Sub-Group. The ISSG Search Filter Resource. https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home (accessed on 31 March 2024). 2024.
15. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015; 349: g7647.

Appendix
Answers to practice questions

Chapter 1: Introduction to study designs and randomised controlled trials

1. A case-control study is an observational study. Therefore, the correct answer is 'False'.
2. A cohort study is an observational study in which the researcher takes a passive role in making measurements on the study subjects. Therefore, the correct answer is 'True'.
3. An RCT is an experimental study. Therefore, the correct answer is 'True'.
4. Cohort studies can be prospective or retrospective or could be a combination of both. Therefore, the correct answer is 'True'.
5. In an RCT, the researcher applies an intervention and examines its effect. Therefore, RCTs are always prospective and the correct answer is 'True'.

Chapter 2: Introduction to systematic reviews

1. Qualitative systematic reviews are systematic reviews of qualitative data and are mostly of narrative type. Therefore, the correct answer is 'False'.
2. Meta-analysis without performing a systematic review usually results in combining a biased selection of studies to provide a biased result. Therefore, the correct answer is 'False'.

Chapter 3: Determine the research question

1.
   a. Population: people with type II diabetes mellitus
   b. Intervention: exercise in addition to dietary advice
   c. Comparator: dietary advice
   d. Outcome(s): stroke
2.
   a. Population: people with heart attack
   b. Intervention: heart bypass surgery
   c. Comparator: coronary stent
   d. Outcome(s): longevity of life

Chapter 4: Select the studies

1. Please check your answers with the search available at www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD006338.pub3/full

Chapter 5: Extract the data

1.

a. Number of patients who died within one year: binary outcome (occasionally can be treated as time-to-event outcome)

b. Number of tonsillitis episodes in six months: count outcome

2.

a. www.ncbi.nlm.nih.gov/pubmed/29079379 (deaths, serious adverse events, any adverse events all at maximal follow-up available).

i. Intervention = omarigliptin

ii. Comparator = placebo

The extracted data are shown in Table A.1.

b. www.ncbi.nlm.nih.gov/pubmed/28898281 (deaths, serious adverse events, any adverse events all at maximal follow-up available).

i. Intervention = tenofovir disoproxil fumarate (TDF) prophylaxis

ii. Comparator = observation

The extracted data are shown in Table A.2.

3.

a. www.ncbi.nlm.nih.gov/pubmed/29419692 (length of hospital stay, VAS scores at 12 months, HRQoL using a validated scale at 12 months).

i. Intervention = non-operative group

ii. Comparator = operative group

The extracted data are shown in Table A.3.

Table A.1 Extracted data on binary outcomes (publication 1)

Study name	Outcome	Number of participants with outcome (intervention)	Total number of participants (intervention)	Number of participants with outcome (comparator)	Total number of participants (comparator)
Home 2018	Death	0	165	1	164
Home 2018	Serious adverse events	5	165	8	164
Home 2018	Any adverse events	90	165	99	164

The above information is from Table 3 (Phase A + B). Please note that footnote c of Table 3 indicates that there is one death even though it states 0 vs 0 in the table.

Table A.2 Extracted data on binary outcomes (publication 2)

Study name	Outcome	Number of participants with outcome (intervention)	Total number of participants (intervention)	Number of participants with outcome (comparator)	Total number of participants (comparator)
Buti 2017	Death	4	33	5	28
Buti 2017	Serious adverse events	8	33	7	28
Buti 2017	Any adverse events	9	33	7	28

The above information is from the 'Adverse events' section of the text.

b. www.ncbi.nlm.nih.gov/pubmed/29079379 (HRQoL, change in HBA1C%, change in plasma fasting glucose, all at 24 weeks).

i. Intervention = omarigliptin

ii. Comparator = placebo

The extracted data are shown in Table A.4.

4.

a. www.ncbi.nlm.nih.gov/pubmed/29183844 (total number of adverse events, total number of serious adverse events, and number of hypoglycaemic episodes).

i. Intervention = Insulin Degludec/Insulin Aspart

ii. Comparator = Biphasic Insulin Aspart 30

The extracted data are shown in Table A.6.

5.

a. www.ncbi.nlm.nih.gov/pubmed/29367198 (time-to-death, time-to-any postoperative pulmonary complication).

i. Intervention = physiotherapy education and breathing exercise training session

ii. Comparator = information booklet

The extracted data are shown in Table A.7.

6.

a. https://pubmed.ncbi.nlm.nih.gov/30988825/ (HRQoL)

The classification of risk of bias is shown in Table A.8.

b. https://pubmed.ncbi.nlm.nih.gov/30912414/ (overall survival)

The classification of risk of bias is shown in Table A.9.

Table A.3 Extracted data on continuous outcomes (publication 1)

Study name	Outcome	Mean (intervention)	Standard deviation (intervention)	Total number of participants (intervention)	Mean (comparator)	Standard deviation (comparator)	Total number of participants (comparator)
Mohamadi 2018	Length of hospital stay	4.2	2.2	25	4.1	1.3	25
Mohamadi 2018	VAS scores at 12 months	2.6	1.6	25	1.02	1.1	25

HRQoL using a validated scale at 12 months was not reported in the trial. Length of hospital stay was from Table 4 and VAS scores were from Table 2.

Note: It is fine to consider operative group as the intervention and non-operative group as the comparator, but the mean, standard deviation and number of participants have to change accordingly.

Table A.4 Extracted data on continuous outcomes (publication 2)

Study name	Outcome	Mean (intervention)	Standard deviation (intervention)	Total number of participants (intervention)	Mean (comparator)	Standard deviation (comparator)	Total number of participants (comparator)
Home 2018	Change in HBA1C%	−0.49	Not reported	165	−0.10	Not reported	164
Home 2018	Change in plasma fasting glucose	−0.7	Not reported	165	−0.10	Not reported	164

The above information is from Table 2.

HRQoL was not reported in the study.

Standard deviations for change in HBA1C% and change in plasma fasting glucose were not reported. Therefore, 95% CI were extracted as shown in Table A.5.

Table A.5 Calculation of standard deviation (publication 2)

Study name	Outcome	Lower CI (intervention)	Upper CI (intervention)	Total number of participants (intervention)	Lower CI (comparator)	Upper CI (comparator)	Total number of participants (comparator)
Home 2018	Change in HBA1C%	–0.73	–0.24	165	–0.34	0.14	164
Home 2018	Change in plasma fasting glucose	–1.4	0	165	–0.8	0.6	164

The above information is from Table 2.

Table A.6 Extracted data on count outcomes (publication 1)

Study name	Outcome	Number of events (intervention)	Total number of participants (intervention)	Number of events (comparator)	Total number of participants (comparator)
Hassanein 2018	Adverse events	271	131	200	132
Hassanein 2018	Serious adverse events	9	131	9	132
Hassanein 2018	Number of hypoglycaemic episodes	93	131	389	132

The number of adverse events and serious adverse events was extracted from the 'Adverse events' section. The number of hypoglycaemic episodes was obtained from Table S2.

Table A.7 Extracted data on count outcomes (publication 2)

Study name	Outcome	Hazard ratio	Lower CI of hazard ratio	Upper CI of hazard ratio	Total number of participants (intervention)	Total number of participants (comparator)
Boden 2018	Death	0.78	0.41	1.48	222	219
Boden 2018	Post-operative pulmonary complication	0.48	0.30	0.75	222	219

Time-to-death was from 'Secondary outcomes' section (only the adjusted analysis was reported); time-to-any postoperative pulmonary complication was from the adjusted analysis from Table 3. Since this was an RCT, you could have obtained the unadjusted analysis from Table 3, that is, 0.43, 0.27 and 0.67 from Table 3 instead of 0.48, 0.30 and 0.75.

Table A.8 Data extraction and classification of risk of bias (publication 1)

RoB 2	Quotes	Comments	Classification
Bias arising from the randomisation process		**As per algorithm**	**Low risk**
1.1 Was the allocation sequence random?	Randomisation was implemented using a central computerised randomisation system managed by an independent statistician		Y
1.2 Was the allocation sequence concealed until participants were enrolled and assigned to interventions?	Randomisation was implemented using a central computerised randomisation system managed by an independent statistician		Y
1.3 Did baseline differences between intervention groups suggest a problem with the randomisation process?		Table 1 shows no evidence of baseline differences	PN
Risk of bias due to deviations from the intended interventions (intention-to-treat effect)		**As per algorithm**	**High risk**
2.1. Were participants aware of their assigned intervention during the trial?	Open label		Y
2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	Open label		Y
2.3 If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention that arose because of the trial context?			PN
2.4 If Y/PY to 2.3: Were these deviations likely to have affected the outcome?			NA
2.5 If Y/PY/NI to 2.4: Were these deviations from intended intervention balanced between groups?			NA

RoB 2	Quotes	Comments	Classification
2.6 Was an appropriate analysis used to estimate the effect of assignment to intervention?		Authors performed an intention-to-treat analysis and a per-protocol analysis; however, it is not clear which analyses were used for reporting the HRQoL	NI
2.7 If N/PN/NI to 2.6: Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomised?			NI
Risk of bias due to deviations from the intended interventions (per-protocol effect)		**As per algorithm**	**High risk**
2.1. Were participants aware of their assigned intervention during the trial?	Open label		Y
2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	Open label		Y
2.3. [If applicable] If Y/PY/NI to 2.1 or 2.2: Were important non-protocol interventions balanced across intervention groups?			NI
2.4 [If applicable] Were there failures in implementing the intervention that could have affected the outcome?			NA
2.5 [If applicable] Was there non-adherence to the assigned intervention regimen that could have affected participants' outcomes?			PY
2.6 If N/PN/NI to 2.3, or Y/PY/NI to 2.4 or 2.5: Was an appropriate analysis used to estimate the effect of adhering to the intervention?		Authors performed an intention-to-treat analysis and a per-protocol analysis; however, it is not clear which analyses were used for reporting the HRQoL	NI

(*Continued*)

Table A.8 Data extraction and classification of risk of bias (publication 1) (*Continued*)

RoB 2	Quotes	Comments	Classification
Risk of bias due to missing outcome data		**As per algorithm**	**High risk**
3.1 Were data for this outcome available for all, or nearly all, participants randomised?		10 participants were excluded	PN
3.2 If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data?		No sensitivity analysis was performed by the study authors. As this was a continuous outcome, there is insufficient information to perform a sensitivity analysis across different scenarios	PN
3.3 If N/PN to 3.2: Could missingness in the outcome depend on its true value?		Missing data could be related to outcomes; for example, withdrawals could be related to poor HRQoL	PY
3.4 If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?			NI
Risk of bias in measurement of the outcome		**As per algorithm**	**High risk**
4.1 Was the method of measuring the outcome inappropriate?		HRQoL appears to be measured using appropriate measures	PN
4.2 Could measurement or ascertainment of the outcome have differed between intervention groups?		The outcomes seemed to be measured similarly in the two groups	PN
4.3 If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants?	Open label		Y
4.4 If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received?		HRQoL is a subjective outcome	PY

RoB 2	Quotes	Comments	Classification
4.5 If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?			NI
Risk of bias in selection of the reported result		**As per algorithm**	**Some concerns**
5.1 Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalised before unblinded outcome data were available for analysis?		No pre-specified analysis plan was available	NI
5.2 Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible outcome measurements (for example, scales, definitions, time points) within the outcome domain?		No pre-specified analysis plan was available. From the methods section, it appears that various HRQoL measures were used. Detailed reports and figures are available only for some (and this has been selected based on statistical significance), but the study authors also mention that there were no statistically significant differences in other measures. Therefore, it is probable that the authors mentioned all the measures in the report. However, you cannot be sure as no pre-specified analysis plan was available. Therefore, a classification of NI is also acceptable.	PY
5.3 Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible analyses of the data?		No pre-specified analysis plan was available	NI
Overall risk of bias			**High risk**

Abbreviations: Y = Yes, PY = Probably Yes, N = No, PN = Probably No, NI = No Information, NA = Not Applicable

(*Continued*)

Table A.9 Data extraction and classification of risk of bias (publication 2)

RoB 2	Quotes	Comments	Classification
Bias arising from the randomisation process		**As per algorithm**	**Low risk**
1.1 Was the allocation sequence random?	Randomisation was performed by assigning random numbers according to the random number tables to the intervention or control group		Y
1.2 Was the allocation sequence concealed until participants were enrolled and assigned to interventions?	Supplementation of Omega-3 fatty acids (1 g/day) in an oil-fish capsule or placebo to the intervention or control group was administered once a day along with administration of three cycles of neoadjuvant chemotherapy	While the precise method of allocation concealment was not reported, this was a double-blinded trial and allocation was also probably achieved by the use of placebo	PY
1.3 Did baseline differences between intervention groups suggest a problem with the randomisation process?		Table 1 shows no evidence of baseline differences	PN
Risk of bias due to deviations from the intended interventions (intention-to-treat effect)		**As per algorithm**	**Low risk**
2.1. Were participants aware of their assigned intervention during the trial?	Double-blind RCT... placebo	Blinding was achieved by the use of placebo	PN
2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	Double-blind RCT... placebo	Blinding was achieved by the use of placebo	PN
2.3 If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention that arose because of the trial context?			NA
2.4 If Y/PY to 2.3: Were these deviations likely to have affected the outcome?			NA

RoB 2	Quotes	Comments	Classification
2.5 If Y/PY/NI to 2.4: Were these deviations from intended intervention balanced between groups?			NA
2.6 Was an appropriate analysis used to estimate the effect of assignment to intervention?	Drop-out was found in this study in neither intervention nor control groups	All participants were included in the analysis according to the group to which they were randomised	Y
2.7 If N/PN/NI to 2.6: Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomised?			NA
Risk of bias due to deviations from the intended interventions (per-protocol effect)		**As per algorithm**	**Low risk**
2.1. Were participants aware of their assigned intervention during the trial?	Double-blind RCT… placebo	Blinding was achieved by the use of placebo	PN
2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	Double-blind RCT… placebo	Blinding was achieved by the use of placebo	PN
2.3. [If applicable] If Y/PY/NI to 2.1 or 2.2: Were important non-protocol interventions balanced across intervention groups?			NA
2.4 [If applicable] Were there failures in implementing the intervention that could have affected the outcome?			NA
2.5 [If applicable] Was there non-adherence to the assigned intervention regimen that could have affected participants' outcomes?	During the treatment, 3 patients suffered from diarrhoea and the supplementation of Omega-3 was paused for 5–7 days	It is unlikely that the temporary pause in the treatment will affect the overall mortality	PN
2.6 If N/PN/NI to 2.3, or Y/PY/NI to 2.4 or 2.5: Was an appropriate analysis used to estimate the effect of adhering to the intervention?			NA

(*Continued*)

Table A.9 Data extraction and classification of risk of bias (publication 2) (*Continued*)

RoB 2	Quotes	Comments	Classification
Risk of bias due to missing outcome data		**As per algorithm**	**Low risk**
3.1 Were data for this outcome available for all, or nearly all, participants randomised?	Drop-out was found in this study in neither intervention nor control groups	All participants were included in the analysis according to the group to which they were randomised	Y
3.2 If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data?			NA
3.3 If N/PN to 3.2: Could missingness in the outcome depend on its true value?			NA
3.4 If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?			NA
Risk of bias in measurement of the outcome		**As per algorithm**	**High risk**
4.1 Was the method of measuring the outcome inappropriate?		There is only one way to measure all-cause mortality	PN
4.2 Could measurement or ascertainment of the outcome have differed between intervention groups?		The outcome seems to have been measured similarly in the two groups	PN
4.3 If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants?	Double-blind RCT... placebo	Blinding was achieved by the use of placebo	N
4.4 If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received?			NA
4.5 If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?			NA

RoB 2	Quotes	Comments	Classification
Risk of bias in selection of the reported result		**As per algorithm**	**Some concerns**
5.1 Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalised before unblinded outcome data were available for analysis?		No pre-specified analysis plan was available	NI
5.2 Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible outcome measurements (for example, scales, definitions, time points) within the outcome domain?		There is only one way to measure all-cause mortality	Y
5.3 Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible analyses of the data?		No pre-specified analysis plan was available	NI
Overall risk of bias			**Some concerns**

Abbreviations: Y = Yes, PY = Probably Yes, N = No, PN = Probably No, NI = No Information, NA = Not Applicable

Chapter 6: Analyse the data

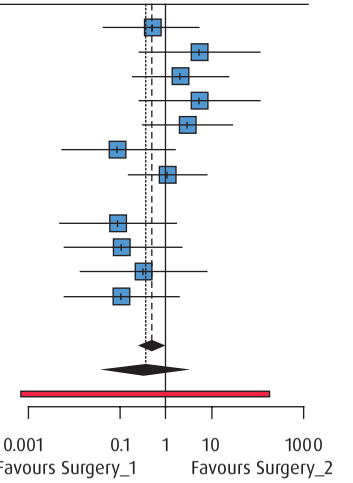
1.

a. Short-term mortality

Figure A.1 Short-term mortality

Study	Surgery_1 Events	Surgery_1 Total	Surgery_2 Events	Surgery_2 Total	Odds ratio GLMM, Fixed + Random, 95% CI
Study 1	1	261	2	258	0.49 [0.04, 5.46]
Study 2	2	123	0	132	5.45 [0.26, 114.71]
Study 3	2	248	1	254	2.06 [0.19, 22.83]
Study 4	2	170	0	180	5.36 [0.26, 112.38]
Study 5	3	181	1	178	2.98 [0.31, 28.95]
Study 6	0	260	5	257	0.09 [0.00, 1.60]
Study 7	2	129	2	139	1.08 [0.15, 7.77]
Study 8	0	102	0	107	
Study 9	0	86	5	89	0.09 [0.00, 1.63]
Study 10	0	172	4	179	0.11 [0.01, 2.12]
Study 11	0	298	1	292	0.33 [0.01, 8.02]
Study 12	0	209	4	202	0.11 [0.01, 1.97]
Total (fixed effect, 95% CI)		**2239**		**2267**	**0.49 [0.25, 0.97]**
Total (random effect, 95% CI)					**0.35 [0.04, 3.32]**
Prediction interval					**[0.00, 173.94]**

Heterogeneity: $Tau^2 = 6.519$; $Chi^2 = 1.30$, df = 10 (P = 1.00); $I^2 = 0\%$
Test for overall effect (fixed effect): Z = −2.04 (P = .04)
Test for overall effect (random effects): $t_{10} = -1.04$ (P = .32)



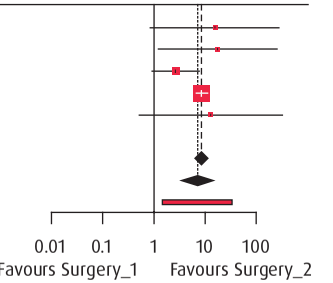
Note: Odds ratio was used as data were sparse.

b. Long-term mortality

Figure A.2 Long-term mortality

Study	logHR	SE	Surgery_Surgery_2 Total	Total	Weight (fixed)	Weight (random)	Hazard ratio IV, Fixed + Random, 95% CI
Study 1	2.7420	1.46	–	–	1.0%	4.5%	15.52 [0.88, 272.99]
Study 3	2.8610	1.36	–	–	1.1%	5.2%	17.48 [1.21, 253.25]
Study 6	0.9900	0.53	–	–	7.3%	24.5%	2.69 [0.96, 7.58]
Study 7	2.1850	0.15	–	–	89.8%	62.2%	8.89 [6.61, 11.95]
Study 9	2.5760	1.65	–	–	0.8%	3.6%	13.14 [0.52, 332.29]
Total (fixed effect, 95% CI)			**0**	**0**	**100.0%**	**–**	**8.27 [6.25, 10.95]**
Total (random effect, 95% CI)					**–**	**100.0%**	**7.15 [3.16, 16.16]**
Prediction interval							**[1.55, 32.91]**

Heterogeneity: $Tau^2 = 0.144$; $Chi^2 = 5.32$, df = 4 (P = .26); $I^2 = 25\%$
Test for overall effect (fixed effect): Z = 14.76 (P < .01)
Test for overall effect (random effects): $t_4 = 6.69$ (P < .01)

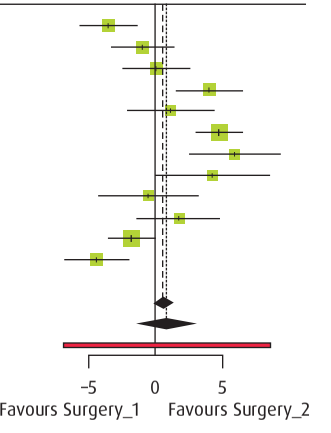


c. Length of hospital stay (days)

Figure A.3 Length of hospital stay (days)

Study	Surgery_1 Mean	SD	Total	Surgery_2 Mean	SD	Total	Weight (fixed)	Weight (random)	Mean difference IV, Fixed + Random, 95% CI
Study 1	12.30	11.80	261	15.80	13.50	258	10.9%	8.9%	−3.50 [−5.68, −1.32]
Study 2	15.10	7.00	123	16.00	11.50	132	9.7%	8.8%	−0.90 [−3.22, 1.42]
Study 3	12.20	15.40	248	12.10	13.90	254	7.9%	8.6%	0.10 [−2.47, 2.67]
Study 4	14.30	13.80	170	10.20	8.90	180	8.7%	8.7%	4.10 [1.65, 6.55]
Study 5	17.70	15.60	181	16.50	15.90	178	4.9%	7.9%	1.20 [−2.06, 4.46]
Study 6	16.50	10.30	260	11.60	9.70	257	17.5%	9.3%	4.90 [3.18, 6.62]
Study 7	15.40	15.60	129	9.30	12.60	139	4.5%	7.8%	6.10[2.69, 9.51]
Study 8	17.20	15.60	102	12.80	15.90	107	2.8%	6.9%	4.40 [0.13, 8.67]
Study 9	15.60	10.70	86	16.10	14.70	89	3.6%	7.4%	−0.50 [−4.30, 3.30]
Study 10	12.50	15.40	172	10.60	15.90	179	4.8%	7.9%	1.90 [−1.37, 5.17]
Study 11	11.20	8.00	298	13.00	13.40	292	16.3%	9.2%	−1.80 [−3.59, −0.01]
Study 12	12.40	8.70	209	16.80	15.80	202	8.5%	8.7%	−4.40 [−6.88, −1.92]
Total (fixed effect, 95% CI)			**2239**			**2267**	**100.0%**	**–**	**0.62 [−0.11, 1.34]**
Total (random effect, 95% CI)							**–**	**100.0%**	**0.87 [−1.43, 3.17]**
Prediction interval									**[−6.89, 8.62]**

Heterogeneity: $Tau^2 = 11.027$; $Chi^2 = 83.72$, df = 11 (P < .01); $I^2 = 87\%$
Test for overall effect (fixed effect): Z = 1.67 (P = .09)
Test for overall effect (random effects): $t_{11} = 0.83$ (P = .42)



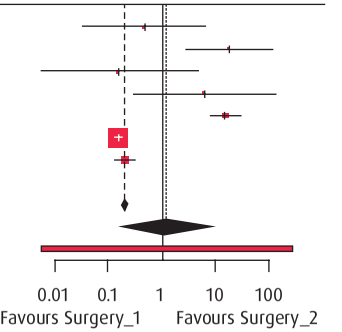
Note: Missing standard deviations were imputed from the highest standard deviation in the group of the remaining trials.

d. Number of complications

Figure A.4 Number of complications

Study	logIRR	SE	Surgery_Surgery_2 Total	Total	Weight (fixed)	Weight (random)	Incidence rate ratio IV, Fixed + Random, 95% CI
Study 1	−0.7170	1.33	–	–	0.2%	12.1%	0.49 [0.04, 6.63]
Study 3	2.9290	0.94	–	–	0.5%	14.4%	18.71 [2.97, 117.85]
Study 5	−1.8200	1.71	–	–	0.2%	10.0%	0.16 [0.01, 4.58]
Study 6	1.8670	1.56	–	–	0.2%	10.7%	6.47 [0.30, 138.44]
Study 8	2.7560	0.31	–	–	4.6%	17.4%	15.74 [8.60, 28.78]
Study 9	−1.8510	0.07	–	–	84.2%	17.8%	0.16 [0.14, 0.18]
Study 11	−1.5460	0.21	–	–	10.1%	17.6%	0.21 [0.14, 0.32]
Total (fixed effect, 95% CI)			**0**	**0**	**100.0%**	**–**	**0.21 [0.18, 0.24]**
Total (random effect, 95% CI)					**–**	**100.0%**	**1.26 [0.16, 10.00]**
Prediction interval							**[0.01, 277.14]**

Heterogeneity: $Tau^2 = 3.685$; $Chi^2 = 240.74$, df = 6 ($P < 0.01$); $I^2 = 98\%$
Test for overall effect (fixed effect): $Z = -23.83$ ($P < 0.01$)
Test for overall effect (random effects): $t_6 = 0.27$ ($P = 0.79$)



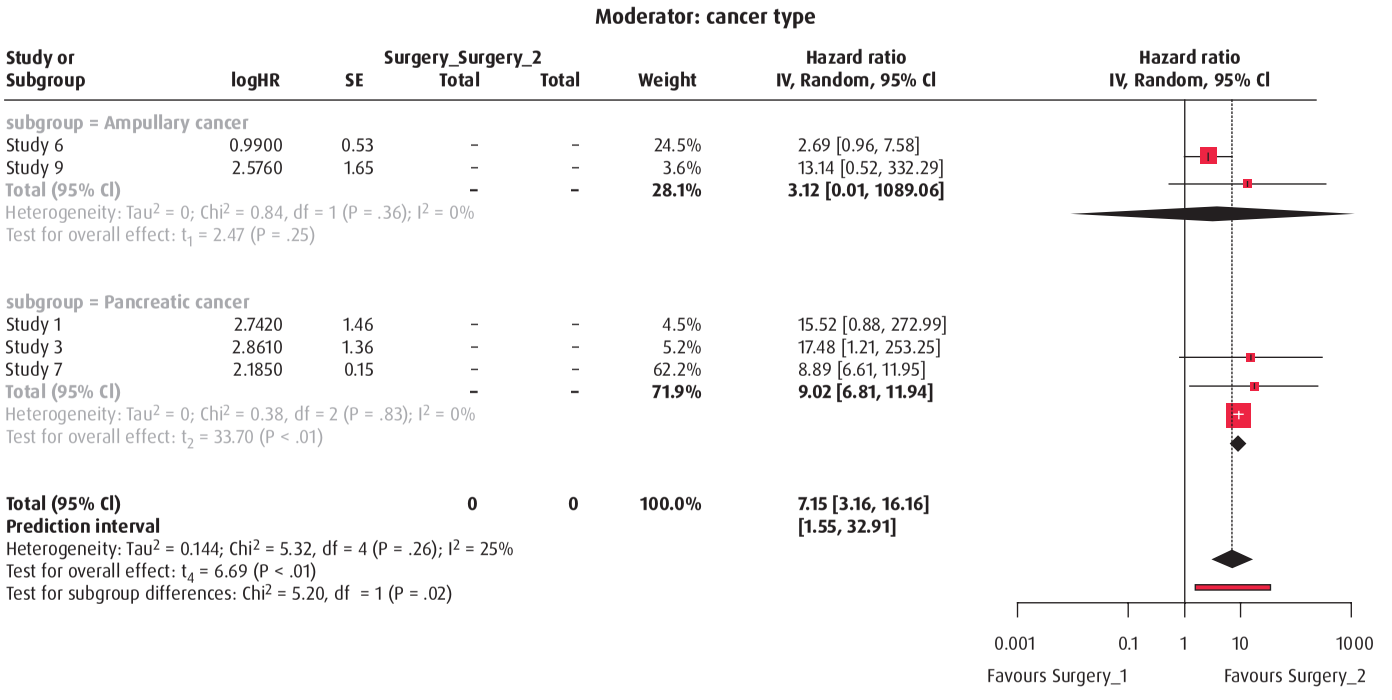
2.

a. Short-term mortality

A subgroup analysis was not performed because of sparse data. The metaregression revealed that the Omnibus Test of Moderators was not statistically significant ($p = 0.9614$).

b. Long-term mortality

Figure A.5 Subgroup analysis: long-term mortality



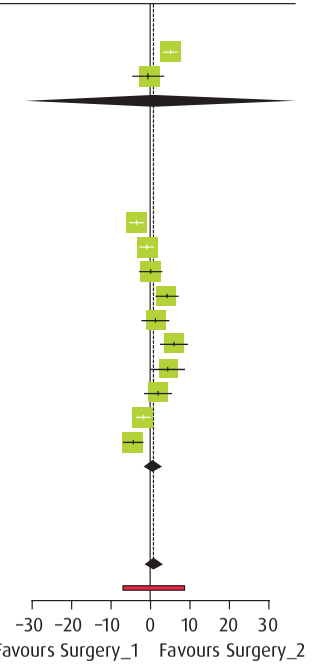
The forest plot showing the test for subgroup differences is shown above. The test for subgroup differences is statistically significant (P = 0.02). The Omnibus Test of Moderators (by metaregression) *p*-value was 4e-04, that is, $4 \times 10^{-4} = 0.0004$, which is statistically significant.

c. Length of hospital stay

Figure A.6 Subgroup analysis: length of hospital stay

Moderator: cancer type

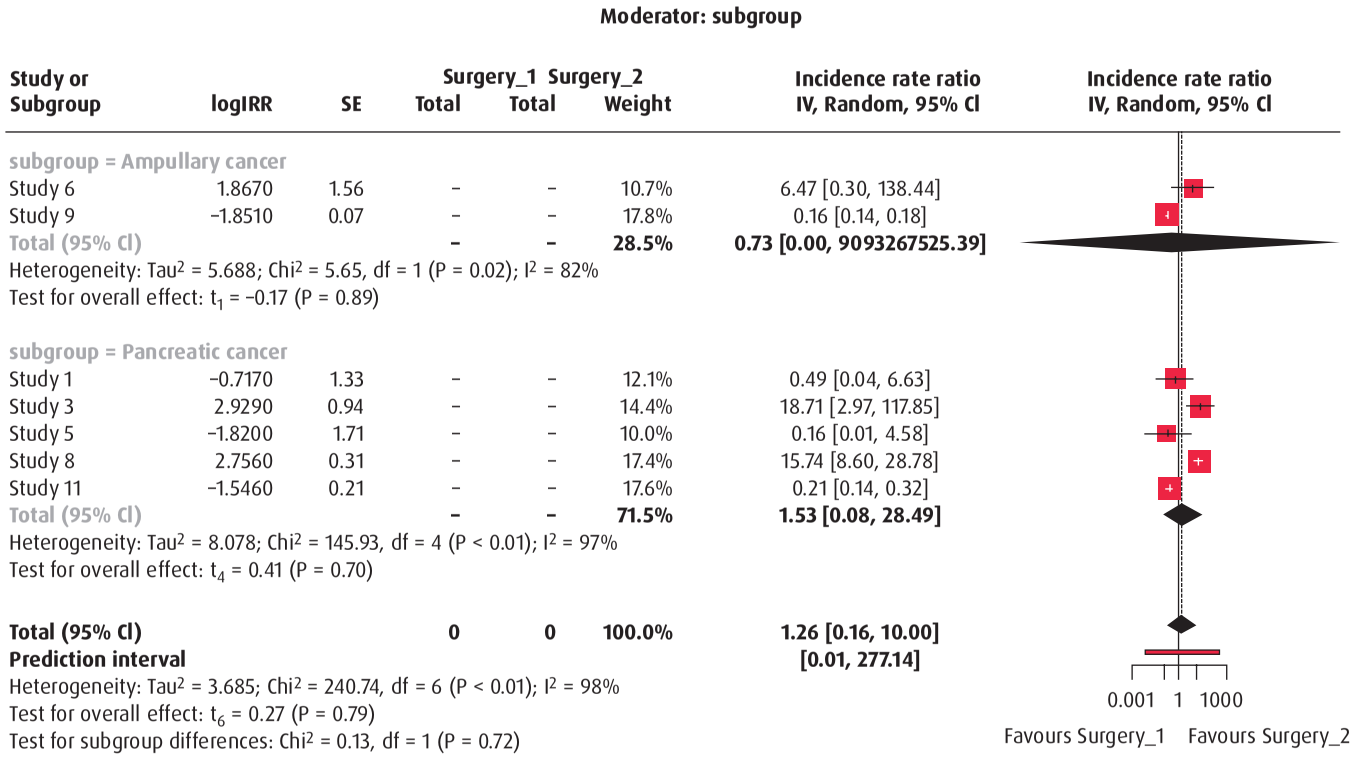
Study or Subgroup	Surgery_1 Mean	SD	Total	Surgery_2 Mean	SD	Total	Weight	Mean difference IV, Random, 95% CI
subgroup = Ampullary cancer								
Study 6	16.50	10.30	260	11.60	9.70	257	9.3%	4.90 [3.18, 6.62]
Study 9	15.60	10.70	86	16.10	14.70	89	7.4	−0.50 [−4.30, 3.30]
Total (95% CI)			**346**			**346**	**16.6%**	**2.48 [−31.65, 36.60]**
Heterogeneity: $Tau^2 = 12.313$; $Chi^2 = 6.43$, df = 1 (P = .01); $I^2 = 84\%$								
Test for overall effect: $t_1 = 0.92$ (P = .53)								
subgroup = Pancreatic cancer								
Study 1	12.30	11.80	261	15.80	13.50	258	8.9%	−3.50 [−5.68, −1.32]
Study 2	15.10	7.00	123	16.00	11.50	132	8.8%	−0.90 [−3.22, 1.42]
Study 3	12.20	15.40	248	12.10	13.90	254	8.6%	0.10 [−2.47, 2.67]
Study 4	14.30	13.80	170	10.20	8.90	180	8.7%	4.10 [1.65, 6.55]
Study 5	17.70	15.60	181	16.50	15.90	178	7.9%	1.20 [−2.06, 4.46]
Study 7	15.40	15.60	129	9.30	12.60	139	7.8%	6.10 [2.69, 9.51]
Study 8	17.20	15.60	102	12.80	15.90	107	6.9%	4.40 [0.13, 8.67]
Study 10	12.50	15.40	172	10.60	15.90	179	7.9%	1.90 [−1.37, 5.17]
Study 11	11.20	8.00	298	13.00	13.40	292	9.2%	−1.80 [−3.59, −0.01]
Study 12	12.40	8.70	209	16.80	15.80	202	8.7%	−4.40 [−6.88, −1.92]
Total (95% CI)			**1893**			**1921**	**83.4%**	**0.51 [−1.96, 2.98]**
Heterogeneity: $Tau^2 = 8.976$; $Chi^2 = 54.97$, df = 9 (P < .01); $I^2 = 84\%$								
Test for overall effect: $t_9 = 0.47$ (P = .65)								
Total (95% CI)			**2239**			**2267**	**100.0%**	**0.87 [−1.43, 3.17]**
Prediction interval								**[−6.85, 8.59]**
Heterogeneity: $Tau^2 = 11.027$; $Chi^2 = 83.72$, df = 11 (P < .01); $I^2 = 87\%$								
Test for overall effect: $t_{11} = 0.83$ (P = .42)								
Test for subgroup differences: $Chi^2 = 0.46$, df = 1 (P = .50)								



Neither the test for subgroup differences *p*-value ($p = 0.50$) nor the Omnibus Test of Moderators by metaregression ($p = 0.5467$) was statistically significant.

d. Number of complications

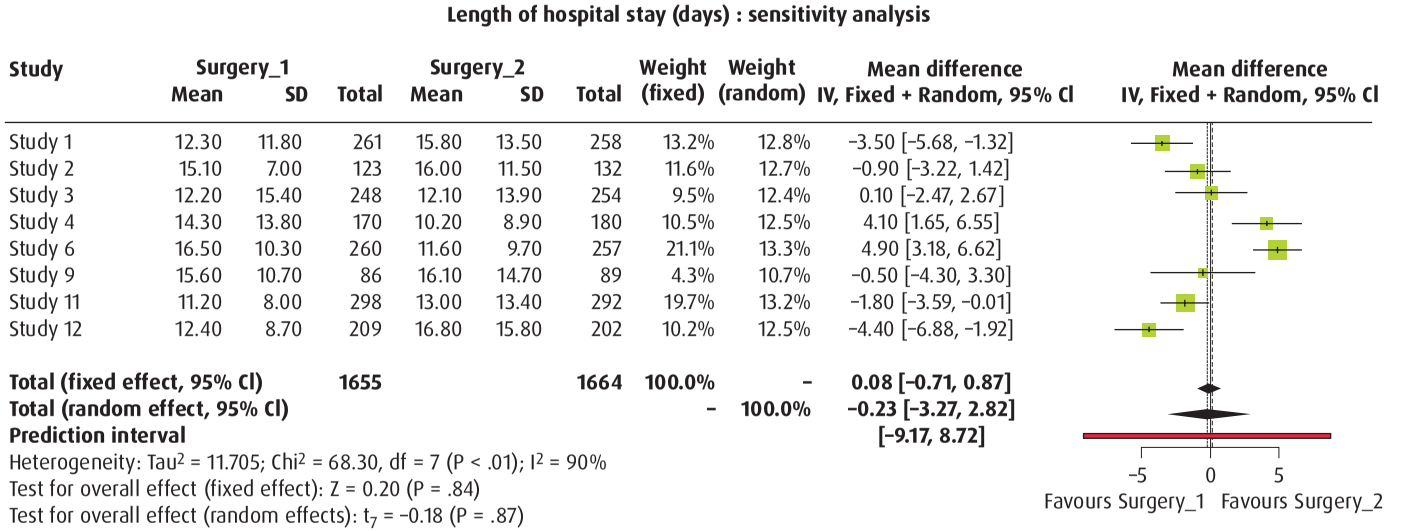
Figure A.7 Subgroup analysis: number of complications



Neither the test for subgroup differences *p*-value ($p = 0.72$) nor the Omnibus Test of Moderators ($p = 0.9159$) was statistically significant.

3.

Figure A.8 Sensitivity analysis: length of hospital stay



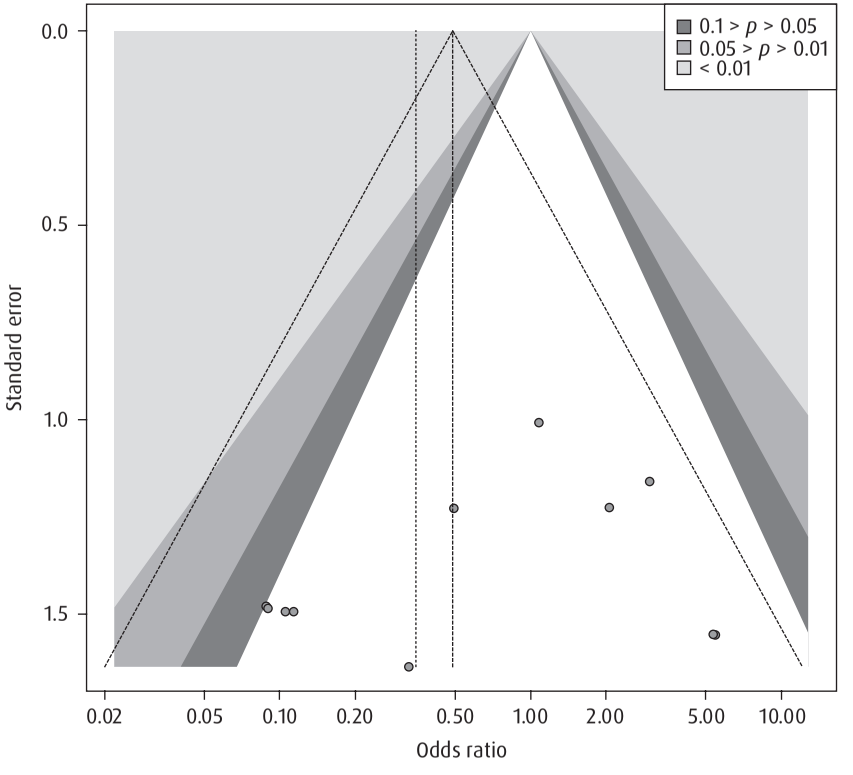
The forest plot in Figure A.8 shows the results of the meta-analysis after excluding 'Study 5' and 'Study 8' in which the standard deviations were missing and 'Study 7' and 'Study 10' in which mean and standard deviation were imputed.

4. Reporting bias is reporting of studies or outcomes within studies based on the results. Reporting bias can lead to meta-analysis resulting in incorrect summary estimates and 95% CI. In other words, you can no longer trust the results.

5.

a. Short-term mortality

Figure A.9 Short-term mortality (funnel plot)



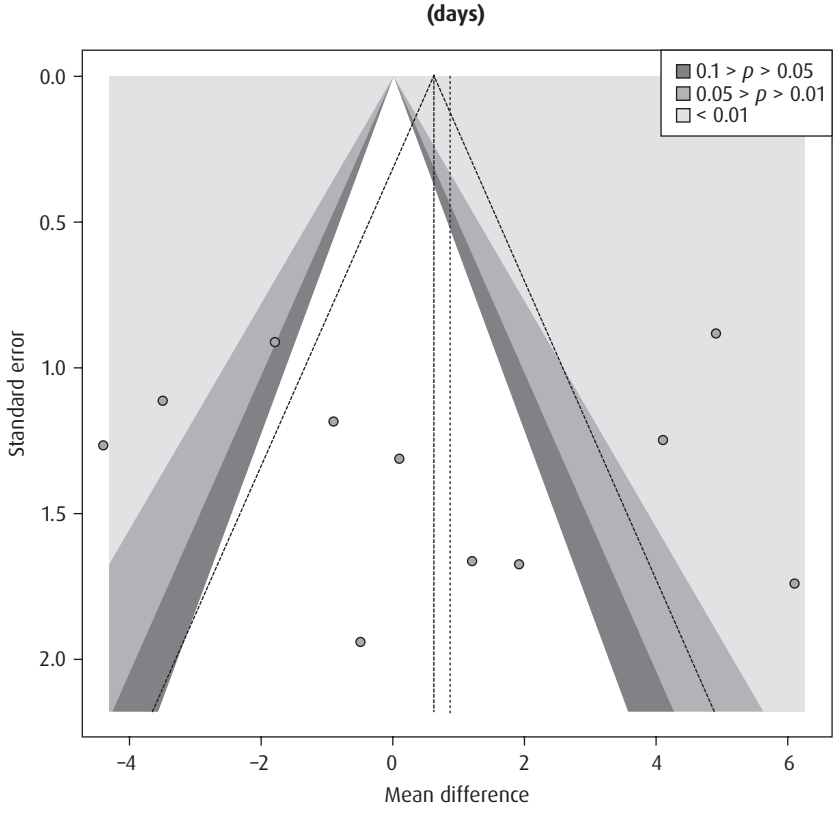
The test for funnel plot asymmetry was not performed since there were fewer than 10 trials that provided data for analysis.

b. Long-term mortality

Funnel plot was not created and tests for funnel plot asymmetry were not performed since there were fewer than 10 trials for this outcome.

c. Length of hospital stay

Figure A.10 Length of hospital stay (funnel plot)



The *p*-value for test for funnel plot symmetry was 0.6842.

d. Number of complications

Funnel plot was not created and tests for funnel plot asymmetry were not performed since there were fewer than 10 trials for this outcome.

Chapter 7: Interpret the results

1. There is no evidence of difference in the short-term mortality between Surgery_1 and Surgery_2 since the odds ratio is 0.35 with a 95% CI of 0.04 and 3.32, which overlaps 1 which is the line of no effect for ratios. This can

be interpreted as the odds of dying if the patient has undergone Surgery_1 is 35% lower (96% lower to 232% higher) than that of patients undergoing Surgery_2. The test for overall effect also shows that there is no significant difference between the two groups according to the random model. There was no evidence of heterogeneity based on I^2, Chi^2 test for heterogeneity (P = 1.0), and overlap of CI.

2. The ICEMAN tool was used to assess the credibility of this subgroup analysis. The assessments are as follows.

a. Is the effect modification based on comparison within rather than between RCTs?

'Completely between': subgroup analysis performed using only study-level data (this is evidenced from the study names).

b. If two or more within-trial comparisons are available, is the effect modification similar from trial to trial?

Not applicable.

c. For between-RCT comparisons, is the number of studies large?

'Very small': 2 in the smallest subgroup.

d. Was the direction of effect modification correctly hypothesised a priori?

'Definitely no': Surgery_1 was hypothesised to be more beneficial for pancreatic cancer than ampullary cancer. In the results, Surgery_1, the long-term mortality was 9.02 and 3.12 with pancreatic cancer and ampullary cancers, that is, all participants receiving Surgery_1 fared worse than those receiving Surgery_2, but this was worse in participants with pancreatic cancer than in participants with ampullary cancer.

e. Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification?

'Chance a likely explanation': Test for subgroup differences was $p = 0.02$

f. Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?

'Definitely yes': This was one of the three planned subgroup analyses.

g. Did the authors use a random-effects model?

'Definitely yes': Random-effects model was used.

h. If the effect modifier is a continuous variable, were arbitrary cut points avoided?

Not applicable.

Overall credibility: Low credibility: reasons a, c, d and possibly e had possibly reduced credibility.

3. Sensitivity analysis did not change the statistical or clinical interpretation of the treatment effects. Therefore, the meta-analysis results were robust to the assumption about means and standard deviations.

4. There was no asymmetry around the plane of symmetry (the summary estimate).
5. The 95% CI of random-effects model were 0.04 and 3.32. This included 0.85 and 1.15, indicating a 0.15 decrease and 0.15 increase in relative odds. This suggests that there is 'no evidence of clinically important difference'.
6. There was downgrading or upgrading for the following aspects.
    a. Risk of bias: Downgraded one level as there was moderate risk of bias which could affect the estimates.
    b. Heterogeneity: No downgrading as there was no heterogeneity suggested by I^2, Chi^2 test for heterogeneity, and good overlap of CI.
    c. Indirectness: No downgrading as there were no concerns about applicability to patients.
    d. Imprecision: Downgraded one level as the treatment effect indicated 'no evidence of clinically important difference' and although the sample size calculations were not available, the total number of participants included in the analysis was in excess of 4,000.
    e. Publication bias: No downgrading as there was no evidence of reporting bias.
    f. Large magnitude of effect: No upgrading as the lower CI were above 2 and upper CI were not below 0.5.
    g. Dose-response gradient: No upgrading as this was a comparison involving two surgeries.
    h. Effect of plausible residual confounding: No upgrading as the studies were all RCTs.
    i. Overall certainty: Low certainty evidence (two levels of downgrading in total without any levels of upgrading resulted in decreasing two categories from high certainty to low certainty).

Index

'An understanding of how treatment options are assessed and compared has never been more important for those involved in the healthcare industry… With years of teaching experience Professor Gurusamy takes us through the basics of study design, an introduction to systematic reviews and how to identify an important research question. This is an essential read and step-by-step guide for early enthusiasts in systematic reviews.'
Brian R. Davidson, *Royal Free Hospital*

'This textbook on systematic reviews is an indispensable reference for novice researchers…The clear explanations and structured approach make complex concepts accessible, ensuring that even beginners can confidently navigate and execute systematic reviews. A must-have resource for anyone looking to deepen their understanding of research methodologies.'
Stefan Bouwense, *Maastricht UMC+*

This practical guide to performing systematic reviews in a healthcare context provides a step-by-step approach for students and health professionals. Using free, open source software to extract data and perform the necessary meta-analyses, this guide navigates the process of reviews, from study design and randomised controlled trials to interpreting results and reporting your findings.

Each chapter includes an overview, learning outcomes and practice questions. Aimed particularly at those who do not have an extensive statistical background, it will help enable you to confidently perform high quality systematic reviews.

KURINCHI GURUSAMY is Professor of Evidence-based Medicine and Surgery, Head of Research at Division of Surgery and Interventional Science, and Lead of Surgical and Interventional Group at Comprehensive Clinical Trials Unit at UCL.

Free open access version available from **www.uclpress.co.uk**

UCLPRESS

Cover Design:
EMC Design Ltd