Synthesizing Quantitative Evidence

Craig Lockwood

Raluca Sfetcu

Eui Geum Oh

Synthesizing Quantitative Evidence

Craig Lockwood

Raluca Sfetcu

Eui Geum Oh

DISCLAIMER

Care has been taken to confirm the accuracy of the information present and to describe generally accepted practices. However, the authors, editors, and publisher are not responsible for errors or omissions or for any consequences from application of the information in this book and make no warranty, expressed or implied, with respect to the currency, completeness, or accuracy of the contents of the publication. Application of this information in a particular situation remains the professional responsibility of the practitioner.

To purchase additional copies of this book, please visit Lippincott's NursingCenter.com or call our customer service department at (800) 638-3030 or fax orders to (301) 223-2320. International customers should call (301) 223-2300.

Visit Lippincott Williams & Wilkins on the Internet: http://www.lww.com. Lippincott Williams & Wilkins customer service representatives are available from 8:30 am to 6:00 pm, EST.

Series Editor: Professor Alan Pearson AM

This series of concise texts is designed to provide a "toolkit" on synthesising evidence for healthcare decision-making and for translating evidence in action in both policy and practice. The series seeks to expand understandings of the basis of evidence-based healthcare and brings together an international range of contributors to describe, discuss and debate critical issues in the field.

Incredible developments have occurred in the synthesis and use of evidence in healthcare over the last several years, but the science and emerging practices that underpin evidence based healthcare are often poorly understood by policy makers and health professionals. Several emerging and exciting developments have much to offer health professionals. Firstly, new, deeper understandings of the nature of evidence and of ways to appraise and synthesise it have led to the development of more sophisticated methodologies for synthesis science. Secondly, the realization that the rapid increase in the availability of high quality evidence has not been matched by increases in the translation of this evidence into policy and/or clinical action has spurred on developments in the science of knowledge implementation and practice improvement.

The burgeoning publications in this area – particularly books on evidence based healthcare - go only so far in informing responsible and conscientious policy makers and healthcare practitioners. This new series Lippincott/Joanna Briggs Institute, "Synthesis Science in Healthcare", is devoted to communicating these exciting new interventions to researchers, clinicians on the frontline of practice and policy makers.

The books in this series contain step-by-step detailed discussions and practical processes for assessing, pooling, disseminating and using the best available international evidence. In all healthcare systems, the growing consensus is that evidence-based practice offers the most responsible course of action for improving health outcomes. All clinicians and health scientists want to provide the best possible care for patients, families and communities. In this series, our aim is to close the evidence to action gap and make that possible.

About the Authors

Craig Lockwood RN, BNurs, MNSc is the Associate Director for Translation Science in the Joanna Briggs Institute, Faculty of Health Sciences, The University of Adelaide, Australia. Craig's interests are in the pragmatic development of resources and systems for the implementation and evaluation of evidence based practice from a multi disciplinary perspective (Translation Science). Craig has particular interests in systematic review methodology, in qualitative synthesis methodology, and is an active reviewer for the Cochrane Renal Group, a convenor of the Cochrane Qualitative Research Methods Group, and actively teaches, and supervises as well as conducts synthesis science projects of qualitative, quantitative and mixed methods research findings.

Raluca Sfetcu MS Diploma in Human Resources Management is a lecturer at Spiru Haret University in Bucharest and PhD candidate at Babes-Bolyai University, Cluj-Napoca, Faculty of Psychology. Raluca is also an affiliated researcher for the Institute of Economic Prognosis, Bucharest, an affiliated staff member of the national school for public health, Training and Health Management, a member of the Romanian Centre for Evidence Based Nursing and Midwifery: an Affiliated centre of the Joanna Briggs Institute, a Comprehensive Systematic Review Trainer and a member of the Young Gastein Task Force, within the European Health Forum, Gastein.

Dr Eui Geum Oh RN, BSN, MSN, PhD is a Professor and an Associate Dean of Academic Affairs for the College of Nursing at Yonsei University, Korea. She is also the Director of Yonsei Evidence Based Nursing Center which is the Korean Affiliate Center of the Joanna Briggs Institute. She is an expert in the synthesis and translation of scientific evidence in nursing practice. She has conducted systematic reviews on the effects of therapeutic lifestyle modification programs in osteoporosis. Dr Oh has developed nursing curricula integrating the evidence-based practice concept into undergraduate and graduate educational programs. She is currently a national review board member representing the nursing field at the National Research Foundation, Korea.

Contents

Introduction

T he objective of a systematic review is to summarize the evidence on a specific clinical question using a transparent, a-priori protocol driven approach. Characteristics of the systematic review methodology are that in achieving this summary of the evidence, there is a critical evaluation of primary studies, data extraction is undertaken in a reliable and repeatable way, and that the results (either meta analysis, or narrative summary if meta analysis is not possible) are scrutinized for validity. Even reviews that are unable to identify a statistically significant difference are very helpful in providing an overview of the available evidence, highlighting gaps in the evidence base that can be used to inform research agenda's and therefore new research questions.

This book provides an overview of the fundamental knowledge, principals and processes for the synthesis of quantitative data in reviews of the effectiveness of health care interventions. As such, it is designed for new reviewers, for students and as an introductory text for academics looking for a book on the fundamentals rather than advanced statistical processes.

It is very much our hope that this book assists you to understand, and to then undertake systematic reviews that provide pragmatic guidance for policy or practice. Systematic reviews are considered hierarchically as the highest form of evidence as they systematically search, identify, and summarize the available evidence that answers a focused clinical question with particular attention to the methodological quality of studies (all papers are critically appraised) or the credibility of opinion and text. As such, the methods described in this book are intended for practical reading and, we trust, will increase the conduct of systematic reviews across the health sciences.

1

The Synthesis of Quantitative Evidence

Chapter 1:
Positivism and its Role in Knowledge Generation

Positivism

The notions that underpin positivism are rarely discussed as its relevance to generating knowledge in the health domain is largely accepted, based on widely held, well-grounded assumptions and understandings. Why then include a section on positivism and its under-lying assumptions in a text on conducting reviews of effects? The intent of this section is to illustrate the origins of thought and logic that underpin the processes described in more prag-matic detail in the following chapters. These are the roots; the bases for our understandings of the natural world and how we investigate it, and therefore warrant some consideration. Our conventions regarding inference and probability are based on these notions and schools of thought. Positivism, (i.e. logical positivism, or the empirico-analytical paradigm) attempts to view the world objectively in order to manipulate and control it. Fay (1975) says that the positivist stance attempts to:

> *'...understand a state of affairs scientifically only to the extent that we have knowledge of what to do in order to control it, and it is thus that the ability to control phenomena provides the framework in terms of which scientific explanation proceeds.' (Fay 1975)*

Consequently, knowledge is generated through the use of controlled observation and ex-perimentation in an attempt to confirm explanations. Thus the researcher starts off with a theoretical idea that is then transformed into a hypothesis to be tested using objective meth-ods. If the hypothesis is confirmed, it is then assumed that this will occur in the same way in the future and thus the event and its results become predictable. There is no doubt that areas of interest to health care are amenable to such an approach, especially those that relate to physiological processes.

The positivist paradigm is concerned with quantitative theory that aims at controlling the physical world. It is seen as quite distinct from qualitative, subjective, personal experience. In order to maintain objectivity, distance between the research process/researcher and the research subject(s) is meticulously maintained and the subtle, non-quantifiable subjective components of human existence are devalued.

Its central approach is deductive in that theory is developed and then tested. If it is supported through testing then it becomes law that is generalizable when the same events occur in the future. If one adopts a positivist stance, then theory and practice become separated.

Positivism in perspective

The social and political world in general, and western health care systems in particular, are currently driven by a view which values objectivity and control and they are thus grounded in a view which sees the positivist paradigm as the legitimate framework for research in to the efficacy, or effectiveness of health care interventions.

It is obvious to most of us that research based on the positivist paradigm has served humanity well, having uncovered numerous explanations for phenomena in the physical world which were previously mysterious. This is especially evident in the area of medical science. Thus, it should not be devalued or dismissed in areas of study that are amenable to control and measurement and where causality can reasonably be identified. The positivist paradigm has been subjected to vigorous critique for over a century, yet it still remains dominant in western thought.

Critiquing the positivist paradigm

Critiquing positivism has become almost passé—it has been a part of intellectual debate for centuries, yet its credibility is not at risk, although in some forms the critique continues. The two central problems of the positivist tradition which have concerned its critics are:

- its reduction of the universe to mechanistic systems and structures
- its assumption that objectivity is reality.

As Murphy (Murphy 1975) suggests, the positivist paradigm has as its goal the separation of reality from subjectivity. 'Truth' is seen to have an 'identity that is immune to interpretation' and 'valid knowledge' is assumed to be 'untrammeled by situational contingencies' (pg 601, 603). Structures are seen to be objective and able to be studied without recourse to the personal experiences of the people who are perceived to be regulated by the structures. In this way structures have been reified and divorced from the actions of people, and researchers are 'neutral observers' (Munhall and Oiler 1993) (pg603).

Munhall and Oiler (1993) suggest the social world is perceived as being orderly, two dimensional, and as acting on the individual—regulating actions and ensuring stability. Behavior and verbal responses have been studied in isolation from interpretation and meaning. Research informed by this tradition is concerned with prediction, control and generalizability. It tends to be theory testing and theory results from deductive reasoning. Such research relies upon quantitative methods and results are usually reported in numerical form following statistical analysis.

The question of objectivity

The issue of objectivity is crucial to any discussion on the nature of 'good' research. Critics of positivism argue that the central assumption that truth can only be pursued from a position of objectivity is its fundamental flaw. They suggest that the objective position is essentially an illusion and that scientists working within this paradigm have deceived themselves in believing that they are being objective. At worst, they argue, adherents to a strictly applied positivist tradition have been involved in a monumental power play built upon deception and gate keeping.

Positivism and the origins of quantitative systematic review

The science of systematic reviews evolved within the positivist paradigm, and while aspects and fine detail may be debated, there is broad consensus that a systematic review can be identified by a particular set of characteristics. These, as Triccoa et al (2008) suggest in their analysis of published systematic reviews tend to be focused on minimizing the risk of bias in the following 5 domains. The development and adherence to an a-priori protocol to reduce risk of 'researcher influence particularly in relation to the results; methods for the identification of literature to be assessed for inclusion (publication and citation bias); for how studies are selected for retrieval (selection bias); and how the quality of identified studies is rated or appraised in determining whether they should be included or not (risk of assessment bias).

These accepted conventions sit well within the positivist paradigm as they are objective measures with known impact on reducing the risk of bias. Crotty identified these distinctions in his foundational text on research in the social sciences by highlighting that the attributes of positivism are associated with objectivity, what we study from this perspective has meaning of its own, and this meaning can be understood if our methods ensure the researcher and the researched do not cross contaminate, use empirical methods of measurement, and if the line of inquiry is one that seeks to discover meaning rather than ascribe meaning (Crotty 1998). In this way, Crotty draws out the distinguishing features of quantitative research, and the strengths of research methods that focus on objectivity.

Chapter 2:

The Review of Quantitative Evidence

JBI follows the methods developed by the Cochrane Collaboration for reviews of effects. Systematic reviews published by JBI are normally completed by Collaborating Centers and Evidence Synthesis Groups associated with Collaborating Centers. Based on the approach of the NHS Centre for Reviews and Dissemination at the University of York, a JBI review is undertaken by a designated reviewer who works with a review panel and the review process is confined to a period of twelve months. The Collaboration Support Unit in JBI coordinates the update of reviews.

Quantitative Evidence and Evidence-Based Practice

Quantitative research is defined as "the numerical representation and manipulation of observations for the purpose of describing and explaining the phenomena that those observations reflect" (Babbie 1992). The methods associated with quantitative research developed out of the study of natural phenomena (Pearson, Field et al. 2007). A review of Ulrich Tröhler's 'To Improve the Evidence of Medicine': The 18th century British Origins of a Critical Approach (2000), suggests that quantitative evidence in medicine originated in eighteenth century Britain, when surgeons and physicians started using statistical methods to assess therapies for scurvy, dropsy, fevers, palsies, syphilis, and different methods of amputation and lithotomy (Trohler 2000).

Quantitative research designs in healthcare research attempt to control as many "unknowns" or potential sources of explained findings/bias as possible. The strength of quantitative research lies in its reliability (i.e. repeatability) – the same measurements should yield the same results or answers time after time (Greenhalgh and Taylor 1995; Moher 2003; Bastin 2004; Pearson and Jordan 2007).

Experimental study design

The ideal research design is an experimental design; however for many practical reasons (including ethics) this may not be possible. Randomized controlled trials (RCTs) are considered the best source of experimental evidence as they aim to control as many variables as possible so that any real differences in outcome are due to the intervention alone. RCTs provide robust evidence on whether or not a casual relationship exists between a stated intervention and a specific, measurable outcome, as well as the direction and strength of that outcome. In an ideal RCT design, participants are randomly selected from the population and then randomly allocated to an arm of the experiment. Many tests of statistical significance are based on the above assumptions and this is one of the reasons critical appraisal checklists contain items on random sampling, random allocation and use of appropriate statistical methods.

In reviews of effectiveness, it is common to begin with a statement that randomized controlled trials will be sought, but in the absence of RCTs other experimental study designs will be included. Other study designs may be listed in hierarchical form, giving preference to those

designs which aim to minimize risk of bias (e.g. have some form of randomization or control, or blinding), and end with those most at risk of bias (e.g. descriptive studies with no randomization, control or blinding), or which are most appropriate to the nature of the question.

In addition to risk of bias, study selection may be based on the scope of the research question. The hierarchy of study designs is reasonably consistent internationally, with widespread acceptance that RCTs provide the most robust experimental evidence but it should be noted that the RCT design may not be appropriate for all studies of effectiveness (Joanna Briggs Institute 2008).

Observational study designs

Experimental studies are often not feasible due to a variety of reasons including: ethical issues, financial costs and/or difficulties in recruiting participants. The observational study design provides an alternative way of collecting information and is a much used study design in healthcare research. This type of study has no experimental features and aims to summarize associations between variables in order to generate (rather than to test) hypotheses. They are solely based on observing what happens or what has happened. Observational studies can broadly be described as being either Correlational or Descriptive (Joanna Briggs Institute 2008).

Correlational studies

A Correlational study aims to summarize associations between variables but is unable to make direct inferences about cause and effect as there are too many unknown factors that could potentially influence the data. This type of study design is often useful where it is unethical to deliberately expose participants to harm. The most commonly used Correlational study designs are Cohort and Case-control (Joanna Briggs Institute 2008).

Cohort study

A cohort study is a type of longitudinal study that is commonly used to study exposure-disease associations. A cohort is a group of participants who share a particular characteristic such as an exposure to a drug or being born in the same year for example. Cohort studies can either be prospective or retrospective. Prospective cohort studies collect data after the cohort has been identified and before the appearance of the disease/condition of interest. The appearance of the disease/condition is then counted as an event (eg new case of cancer). In theory all of the individuals within the cohort have the same chance of developing the event of interest over time. A major advantage of this study design is that data is collected on the same participants over time, reducing inter-participant variability. However this type of study is expensive to conduct and can take a long time to generate useful data. Retrospective cohort studies are much less expensive to conduct as they utilize already collected data in the form of medical records. Effectively in a retrospective cohort design, the exposure, latent period and development of the disease/condition have already occurred – the records of the cohort are audited backwards in time to identify particular risk factors for a disease/condition. A disadvantage of this study design is that the data was collected for purposes other than research so information relevant to the study may not have been recorded. Statistically, the prospective cohort study should be summarized by calculating relative risk and retrospective

cohort studies should be summarized by calculating odds ratio (Joanna Briggs Institute 2008).

Case-control study

The case control study also uses a retrospective study design – examining data that has already been collected, such as medical records. "Cases" are those participants who have a particular disease/condition and the "Controls" are those who do not. The records of each are examined and compared to identify characteristics that differ and may be associated with the disease/condition of interest. One recognized disadvantage of this study design is that is does not provide any indication of the absolute risk associated with the disease of interest (Joanna Briggs Institute 2008).

Descriptive studies

Descriptive studies aim to provide basic information such as the prevalence of a disease within a population and generally do not aim to determine relationships between variables. This type of study design is prone to biases such as selection and confounding bias due to the absence of a comparison or control. Case series and case reports are types of descriptive studies (Joanna Briggs Institute 2008).

Case Report/Case Series

A case report provides a detailed description of an individual participant or case. Several case reports can be brought together as a case series. A case series provides detailed descriptions of the exposures and outcomes of participants with a particular disease/condition of interest. This design has been very useful in identifying new diseases and rare reactions or conditions. A case series can be either prospective or retrospective, depending on when the data was collected relative to the exposure. Case report/series lack a comparator or control group but are effective as a question generating study design (Joanna Briggs Institute 2008).

Expert Opinion

The JBI regards the results of well designed research studies grounded in any methodological position as providing more credible evidence that anecdotes or personal opinion; however, in situations where no research evidence exists, expert opinion can be seen to represent the "best available" evidence (Joanna Briggs Institute 2008).

Hierarchy of study designs

Study designs that include fewer controls (and therefore include a greater number of unknown factors or potential sources of bias) are considered to be lower quality of evidence – hence a hierarchy of evidence is created on the basis of the amount of associated bias and therefore certainty of an effect. Many JBI reviews will consider a hierarchy of study studies for inclusion and a protocol should be a statement about the primary study design of interest and the range of studies that will be considered if primary studies of that study design are not found.

- Experimental e.g. randomized controlled trials (RCT)
- Quasi experimental e.g. non-randomized controlled trials

- Observational (Correlational) – e.g. cohort, case control studies
- Observational (Descriptive) – e.g. case series and case study
- Expert opinion

Study design is an important factor when considering when to combine study results in meta analysis, as discussed in later sections.

Hierarchy of Quantitative Evidence – JBI Levels of Evidence

The hierarchy of study designs has lead to a more sophisticated hierarchy or levels of evidence, on the basis of best available evidence. Several international organizations generate levels of evidence and they are reasonably consistent (Cluett; Gutatt 1995). Each systematic review will have levels of evidence associated with its findings. The JBI levels of evidence are discussed in a later section in more detail, they are described briefly in Table 1.

In this case quantitative evidence is ranked in terms of research findings most likely to provide valid information on the effectiveness of a treatment/care option. Such hierarchies usually have systematic review with meta-analysis at the top, followed closely by RCTs. There are several other hierarchies of evidence for assessing studies that provide evidence on diagnosis, prevention and economic evaluations (Cluett): their focus remains quantitative. The major disadvantage in this is that while some health topics may concentrate on treatment/management effectiveness, their themes are very possibly not addressed in RCTs. For example, Kotaska suggests that vaginal breech birth is too complex and multifaceted to be appropriately considered within trials alone (Kotaska 2004). It has been reported how one RCT on breech birth has changed practice (Cluett). The reasons for this are likely to be complicated and involved underlying professional beliefs as well as the evidence. The emphasis, however, on trials as the apogee of hierarchy of evidence may be viewed as only encouraging an acceptance of this as the 'gold standard' in all circumstances, rather than reflecting on whether a specific subject or topic is best considered from a different perspective, using different research approaches. It must be acknowledged that quantitative studies alone cannot explore or address all the complexities of the more social aspects of human life (Cluett; Gray 2004). For example,

Table 1. JBI levels of evidence (brief)

Level 1 (strongest evidence)	Meta-analysis (with homogeneity) of experimental studies (eg RCT with concealed randomization) OR One or more large experimental studies with narrow confidence intervals
Level 2	One or more smaller RCTs with wider confidence intervals OR Quasi-experimental studies (without randomization)
Level 3	a. Cohort studies (with control group) b. Case-controlled c. Observational studies (without control group)
Level 4	Expert opinion, or physiology bench research, or consensus

in midwifery this would include themes such as experience of birth, parenthood, or topics regarding social support, transition to parenthood, uptake of antenatal screening, education, views on lifestyle such as smoking, etc (Aslam 2000). These are more appropriately explored though qualitative research approaches that seek to explore and understand the dynamics of human nature, what makes them believe, think and act as they do (Mays and Pope 1996; Casebeer and Verhoef 1997; Pearson, Field et al. 2007; Pearson and Jordan 2007).

SECTION 2

Developing a protocol for an effects review of quantitative evidence

Chapter 3:
The Systematic Review Protocol

A protocol is important because it pre-defines the objectives and methods of the systematic review. It details on what basis the reviewers will include and exclude studies. What data is important and how it will be extracted and synthesized is also described. A protocol provides the plan for the systematic review and as such can help restrict the likelihood of reporting bias. Any deviations between the protocol and systematic review report should be explained in the systematic review report.

That a protocol is one of the features of a systematic review that sets it apart from traditional literature reviews with their associated risk of bias is broadly speaking accepted as an accurate observation (there are a number of implicit assumptions that will impact on the benefits of an a-priori protocol). The purposes of protocols have been described as relating to transparency, avoiding reviewer "chasing" of ad hoc outcomes, auditability, and avoidance of using the literature to support a particular line of argument, providing a clearly objective analysis of the literature; all of which can be summed up as an attempt to decrease the risk of methods of convenience influencing what is done, and hence what a review finds. Much has been written confirming these distinctions (Cooper, Hedges et al. 2009; Krainovich-Miller, Haber et al. 2009). However, Dixon-Woods perhaps has said it most clearly, suggesting that a protocol is an attempt to minimize arbitrariness by making explicit the review process, so that, in principle, another reviewer with access to the same resources could undertake the review and reach broadly the same conclusions' (Dixon-Woods, Booth et al. 1997).

As with other international organizations, JBI advocates for, and expects standardization in systematic review development as part of its mission to enhance the quality and reliability of reviews being developed across an international collaboration. To facilitate this process, JBI have developed computer software.

The System for the Unified Management, Assessment and Review of Information (SUMARI) is the Joanna Briggs Institutes premier software for the systematic review of literature. It is designed to assist researchers and practitioners in fields such as health, social sciences and humanities to conduct systematic reviews of evidence of feasibility, appropriateness, meaningfulness, effectiveness and to conduct economic evaluations of activities and interventions.

SUMARI includes the Comprehensive Review Management System (CReMS) software, designed to assist reviewers to manage and document a review by incorporating the review protocol, search results and findings. Reviewers are required to undertake systematic reviews using CReMS software.

CReMS links to four analytic modules of SUMARI:

- JBI Qualitative Assessment and Review Instrument (QARI)
- JBI Meta Analysis of Statistics Assessment and Review Instrument (MAStARI)
- JBI Narrative, Opinion and Text Assessment and Review Instrument (NOTARI)
- JBI Analysis of Cost, Technology and Utilization Assessment and Review Instrument (ACTUARI)

JBI quantitative reviews are conducted through the MAStARI module. Before reviewers are able to use CReMS or any of the SUMARI modules, they need to register through the JBI website and obtain a username and password. This process is free of charge.

Chapter 4:
Developing a Protocol

There are a number of variations in the specific style and layout of a systematic review protocol but the approaches of the Cochrane Collaboration; the Campbell Collaboration; the Joanna Briggs Institute; and the Committee on Standards for Systematic Reviews of Comparative Effectiveness Research of the Institute of Medicine (IOM) have much in common and represent the core standards for protocols. In this book, we describe the development of a quantitative protocol using the Joanna Briggs Institute approach, linked to the web-based systematic review software suite JBI-SUMARI. The SUMARI user guide is a recommended reference for technical aspects of creating a JBI review (Joanna Briggs Institute 2007).

Review title

The title of the protocol should be as descriptive as is reasonable and reflect relevant information. If the review aims to examine clinical effectiveness this should be stated in the title. If specific interventions and/or patient outcomes are to be examined these should also be included in the title. Where possible the setting and target population should also be stated (Joanna Briggs Institute 2008).

For example: "*The clinical effectiveness of smoking cessation strategies for adults in acute care mental health facilities: a systematic review*".

This example provides readers with an indication of the population, the interventions, and the outcome of interest, as well as the fact that it is a systematic review. Ensuring the relevant fields of the PICO mnemonic are incorporated in the title assists peer reviewers as well as end users to identify the scope and relevance of the review. This particular title contains less information than might be provided in a classical review of effects where a particular intervention is compared with a particular control among a specific population to address. A more specific example that was focused on a limited number of interventions might read as thus:

> "*The effectiveness of 2mg Nicotine Gum compared with 4mg Nicotine Gum among adult smokers aged 18 or above with a greater than 10pack history on abstinence at 6 months from program entry*".

In this example, you can see that the increased detail has focused the review in much more detailed and definitive terms, it will be a narrower review, rather than looking at a broader range of interventions among a wider population. Either approach is acceptable, with many reviews of effectiveness seeking to address wider ranges of interventions with broader applicability, while it is equally valid to focus a review on detailed interventions, outcomes and participants. You will have noted that neither example provides a duration or intensity of therapy, and these are characteristics that would need to be described in the inclusion criteria given they are not stated in the title. The review title should provide as much detail as possible to allow

effective cataloguing on electronic databases. The clearer and more specific a title is, the more readily a reader will be able to make decisions about the potential relevance of the systematic review.

Question/Objectives development

Once a topic has been identified, a focused, answerable question is developed. This question is reflected in the review title and is specified in detail in the review objective section of the protocol. As with any research, the most important decision in preparing a review is to determine its focus (Higgins and Green 2008). Clearly framed questions are essential for determining the structure of a systematic review or meta-analysis (Hedges 1994). In essence, the properly formulated question will guide much of the review process, including strategies for locating and selecting studies or data, for critically appraising their relevance and validity, and for analyzing variations among their results. A properly formulated question also provides relevance for the initial assessment in the review process. Therefore, it is important that you take your time over it, and discuss it with your co-reviewers.

A range of mnemonics is available to guide the structuring of systematic review questions, the most common for quantitative reviews being PICO. The PICO mnemonic begins with identification of the:

Population, the Intervention being investigated and its Comparator and ends with a specific Outcome(s) of interest to the review. A specific mnemonic for qualitative reviews has also been developed which identifies the Population, the Phenomena of Interest, and the Context. A more generic mnemonic that can be used across quantitative and qualitative reviews is the SPICE mnemonic, where the Setting, Perspective, Intervention, Comparison and (method of) Evaluation are described.

The level of detail incorporated into each aspect of a mnemonic will vary, and consideration of the following will assist reviewers to determine the appropriate level of detail for their review. The population may be the primary focus of interest (for example, in reviews examining gender-based phenomena such as smoking or alcohol use among women) and may further specify an age group of interest or a particular exposure to a disease or intervention.

The intervention(s) under consideration need to be transparently reported and may be expressed as a broad statement such as "The Management of . . .", or framed as a statement of "intervention" and "outcome" of interest. Interventions should be clearly described as there are many types of randomized trials with control groups and blinding (Manchikanti, Hirsch et al. 2008). Explanatory trials test whether an intervention is efficacious; that is whether it can have a beneficial effect in an ideal situation. Pragmatic trials measure effectiveness; the degree of beneficial effect in real clinical practice. Thus, the explanatory trial seeks to maximize the internal validity by issuing rigorous control of all variables other than the intervention, and the pragmatic trial seeks to maximize external validity to ensure that the results can be generalized.

Comparators may include placebos and/or alternative treatments. In qualitative reviews, the interest relates to the experience of a particular phenomenon (for example, men's perceptions of healthy living). There may be one or a range of outcomes of interest depending on the nature

of the topic and planned scope of the review. Comparators (or controls) should be clearly described. It is important to know what the intervention is being compared with. Examples include: no control, placebo or alternative treatments.

Outcomes should be measurable and chosen for their relevance to the review topic and research question. They allow interpretation of the validity and generalizability of the review findings. Examples of outcomes include: morbidity, mortality, quality of life. Reviewers should avoid the temptation of being too vague when determining review outcomes. In identifying which outcomes will be specified, it is useful to consider the interests of the target audience of the review findings, the impact that having a large number of outcomes may have on the scope and progress of the review, the resources (including time) to be committed to the review and the measurability of each specified outcome (Higgins and Green 2008).

While it is important to utilize primary outcomes and secondary outcomes such as functional status, trivial outcomes should not be included as they only overwhelm and confuse the readers by including data that is of little or no importance alongside the data that is important. Consequently, explicit criteria for establishing the presence of appropriate outcomes and if necessary, their combinations must be specified (Higgins and Green 2008). It is always beneficial to list the outcomes of interest and give consideration as to:

- how the outcomes might be measured;
- when the outcomes should be measured;
- which are the most important outcomes;
- bad outcomes as well as good outcomes (note this is not the same as positively or negatively skewed outcome statements).

The objectives of the review should provide a clear statement of the questions being addressed with reference to participants, interventions, comparators and outcomes. Clear objectives and specificity in the review questions assist in focusing the protocol, allow the protocol (and final report) to be more effectively identified, as well as to provide a structure for the development of the full review report. The review objectives should be stated in full. Conventionally, a statement of the overall objective is made and elements of the review are then listed as review questions. For example: "To systematically review the evidence to determine the best available evidence related to the post harvest management of Split Thickness Skin Graft donor sites." This broad statement is then detailed in relation to the specific questions of interest that will guide the development of the review criteria, such as:

Among adults in the acute postoperative phase (5 days) following skin grafting, what dressings used in the management of the STSG donor site are most effective;

- in reducing time to healing,
- in reducing rates of infection, and
- in reducing pain levels and promoting comfort?

What interventions/dressings are most effective in managing delayed healing/infection in the split skin graft donor site?

What interventions are most effective in managing the healed split skin donor site?

Background

The Joanna Briggs Institute places significant emphasis on a comprehensive, clear and meaningful background section to every systematic review (Joanna Briggs Institute 2008). The background should be approximately 1000 words in length and describe the issue under review including the target population, intervention(s) and outcome(s) that are documented in the literature. The background should provide sufficient detail to justify the conduct of the review and the choice of interventions and outcomes. Where complex or multifaceted interventions are being described, it may be important to detail the whole of the intervention for an international readership. Any topic-specific jargon or terms and specific operational definitions should also be explained. In describing the background literature value statements about the effects of interventions should be avoided. The background should avoid making statements about effectiveness unless they are specific to papers that illustrate the need for a systematic review of the body of literature related to the topic. (For example: "Use of acupuncture is effective in increasing smoking cessation rates in hospitalized patients". This is what the review will determine. If this type of statement is made it should be clear that it is not the reviewer's conclusion but that of a third party, such as "Smith indicates that acupuncture is effective in increasing smoking cessation rates in hospitalized patients". Such statements in the background need to be balanced by other points of view, emphasizing the need for the synthesis of potentially diverse bodies of literature.)

Criteria for inclusion/exclusion

Population

In the above example, the PICO mnemonic describes the population (adults) within a specific setting (acute care) within a specific time frame (5 days). There are no subgroups or exclusions described; hence all patients meeting the described criteria would be included in the analysis for each outcome. Specific reference to population characteristics, either for inclusion or exclusion should be based on a clear, scientific justification rather than based on unsubstantiated clinical, theoretical or personal reasoning (Joanna Briggs Institute 2008).

Intervention

In the above example, there is no single intervention of interest, rather the term "dressings" is used to indicate that the review will consider all wound dressing products. Where possible, the intervention should be described in detail, particularly if it is multifaceted. Consideration should also be given to whether there is risk of exposure to the intervention in comparator groups in the included primary studies (Joanna Briggs Institute 2008).

Comparison

The protocol should detail what the intervention of interest is being compared with. This can be as focused as one comparison eg comparing "dressing X with dressing Y' or as broad as "what dressings" from the example above. This level of detail is important in determining inclusion and exclusion once searching and appraisal is complete. Within reviews of effects, the comparator is the one element of the PICO mnemonic that can be either left out of the

question/s, or posited as a generalized statement. Reviews of effects based on the inclusive definition of evidence adopted by The Joanna Briggs Institute often seek to answer broader questions about multifaceted interventions. This contrasts with reviews of effects conducted via The Cochrane Collaboration, where the comparator often needs to be described in detail (Joanna Briggs Institute 2008).

Outcome(s)

The protocol should include a list of all the outcome measures being considered. The relevance of each outcome to the review objective should be apparent from the background section. Outcomes should be measurable and appropriate to the review objective. Outcomes might be classified as being of primary or secondary interest in relation to the review objective. It is useful to list outcomes and identify them as either primary or secondary, short-term or absolute and discuss which ones will be included (Joanna Briggs Institute 2008).

Types of studies

Generally, JBI systematic reviews consider primary research studies, however where appropriate, a systematic review can draw on other systematic reviews as a source of evidence. Authors should justify the need for their review if there are already published systematic reviews on the topic. The appropriate JBI critical appraisal tool should be used and all details should be transparent. As with different study designs, it is inappropriate to combine systematic review data with that of primary research, and data should be treated separately (Joanna Briggs Institute 2008).

This section should flow naturally from the criteria that have been established to this point, and particularly from the review objective and questions. The review question will determine the methodological approach and therefore the most appropriate study designs to include in the review. Quantitative research designs in healthcare research attempt to control as many "unknowns" or potential sources of explained findings/bias as possible. The ideal research design is an experimental design, however for many practical reasons (including ethics) this may not be possible.

As mentioned previously, many JBI reviews will consider a hierarchy of study studies for inclusion. If this is to be the case, there should be a statement about the primary study design of interest and the range of studies that will be considered if primary studies with that design are not found. In reviews of effectiveness, it is common to begin with a statement that randomized controlled trials will be sought, but in the absence of RCTs other experimental study designs will be included. Other study designs may be listed in hierarchical form, giving preference to those designs which aim to minimize risk of bias (e.g. have some form of randomization or control, or blinding), and end with those most at risk of bias (e.g. descriptive studies with no randomization, control or blinding), or which are most appropriate to the nature of the question. In addition to risk of bias, study selection may be based on the scope of the question. The hierarchy of study designs is reasonably consistent internationally, with widespread acceptance that RCTs provide the most robust evidence of effects.

The JBI levels of evidence should be used to describe all included study designs and can be accessed from the JBI website. The differences between these study designs are discussed in detail in JBI Comprehensive Systematic Review training. It is important to use the critical

appraisal tool appropriate to the study design when determining methodological quality of a study for inclusion into a review. The types of studies included in a JBI quantitative review is standardized in CReMS, and consists of the following statement:

"The review will consider any randomized controlled trials; in the absence of RCTs other research designs, such as non-randomized controlled trials and before and after studies, will be considered for inclusion in a narrative summary to enable the identification of current best evidence."

Search strategy

Study selection and data collection are skills that are greatly enhanced by having a working knowledge of how to search the scientific literature. Where the skills and knowledge to search the literature are not readily available, it is strongly recommended that you seek the assistance of an Information Scientist.

Constructing your search

The search process can be described as a three-phase process in the protocol. In summary, these stages are to analyze the question and identify types of key words and related similes that can be used. These exploratory terms are applied to each database that will be included in the comprehensive, exhaustive search of the literature. Studies identified via the text words and similes will reference particular subject heading terms, and may also reference additional text words that are free text terms. A sufficient number of studies should be looked at until you are no longer finding additional subject heading terms, or new text words that apply to your clinical question. These terms are collated and applied on a per database basis, ensuring that the subject heading words are specific to the database they are being applied to, and that for each database, a further search for additional text words is conducted.

The application of these collated terms within the relevant database is the second phase. This phase is fully recorded and the terms used, their sequence, and numbers of results should all be reported in the relevant appendices. Phase two is where traditional databases, grey or gray literature sources are searched, and where experts, peak bodies, non government organizations and researchers of standing in the field and other potential sources of relevant studies are contacted to request any studies or to indicate where such studies may be located outside of the traditional databases.

Phase three of the search strategy, sometimes known as pearling is to review the reference lists of the citations that have been retrieved. If new studies are identified, these are screened for any additional text or subject heading words as a validation test of your search strategy. Once the reference lists are returning only pre-identified studies, you can be more confident that the search strategy has captured the terms appropriate to your review questions. Further basic principals associated with searching are listed below:

- Break up the search question
- Do not over complicate the search by including too many terms
- Do not search across major databases simultaneously (i.e. both CINAHL & EMBASE, as descriptors are not the same in all databases)
- Test search queries

- Are you getting the results you require? What is the specificity of your results? If you are not finding what you want – refine your searches
- Allow for English and American spelling. Use a wildcard character, in most databases this is a '?' (i.e. Colo?r results = colour or color randomi?ed results = randomized or randomized)
- Other wildcards like '$' are unlimited for example: organi$ = organising or organizing or organized or organization
- Variable number of characters: You are able to limit the truncation dog$2 will find dogma (i.e. two letters after dog)
- Boolean searches AND/OR/NOT can lead to either too many results with little relevancy or not enough results as the correct keywords have been missed
- Often drop down boxes or the ability to assign limits will give you date ranges or types of studies.

Use keywords, also referred to as text words or free-text searching to find more colloquial terms for your clinical search. Consider alternatives such as the word Chemist used to describe Pharmacy in an Australian database. Think laterally and consider the search and all its parameters before embarking on definitive search strategies. In Summary, effective searching will include the following types of characteristics:

- using synonyms, wildcards ($ or *), exploded medical subject headings (MeSH)
- Using multiple databases, multiple languages (if applicable)
- Identifying relevant references from the retrieved references
- Using filters – year of publication, population (adults/children), study type
- Hand searching of journals, contacting relevant experts and identifying 'grey literature' (Hand searching involves manual searching of key journals in a particular topic area).

The Institute consider a range of study designs to identify the best available evidence to appropriately answer the identified clinical question (Figure 1).

JBI Levels of Evidence

Levels of evidence are assigned according to the research design being included. The JBI levels of evidence[1] addressing evidence relating to studies of FAME (Feasibility, Appropriateness, Meaningfulness and Effectiveness) (Table 2).

Building the search strategy

Searching for the evidence is a three-step process:

(i) Exploratory searching
(ii) Implementing a standardized tested search strategy within each selected database
(iii) Reviewing the reference list of retrieved studies

Prior to commencing a systematic review it is important to search existing systematic review libraries to ensure that the review you are planning has not already been conducted or currently

---

[1] *http://www.joannabriggs.edu.au/pubs/approach.php*

Figure 1: Study types to identify the best available evidence

Systematic Reviews
Randomized Controlled Trials (RCTs)
Quasi-Experimental Studies
Cohort Studies
Case Control Studies
Cross-Sectional Studies
Case Series
Case Report
Expert Opinion
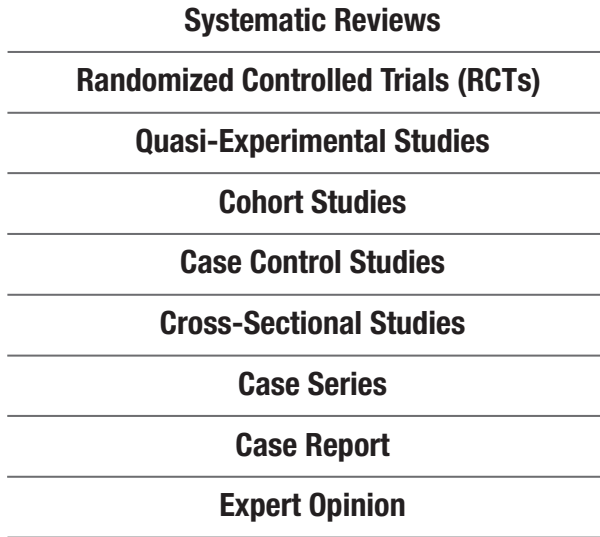
being updated. (e.g. The Cochrane Library, JBI Library of Systematic Reviews and the Centre for Reviews and Dissemination (DARE database)).

The search strategy should also describe any limitations to the scope of searching in terms of dates, resources accessed or languages; each of these may vary depending on the nature of the topic being reviewed, or the resources available to each center. Limiting by date may be used where the focus of the review is on a more recent intervention or innovation. However, date limiting may exclude seminal early studies in the field and should thus be used with caution; the decision preferably being endorsed by topic experts and justified in the protocol. Similarly, restricting study inclusion on the basis of language will have an impact on the comprehensiveness and completeness of the review findings. Where possible, reviewers should seek collaborative agreements with other JBI entities to ensure minimal language restrictions are placed on the identification and inclusion of primary studies

The comprehensiveness of searching and documenting the databases searched is a core component of the systematic review's credibility. In addition to databases of published re-search, there are several online sources of grey or unpublished literature that should be considered. Grey literature is a term that refers to papers, reports, technical notes or other documents produced and published by governmental agencies, academic institutions and other groups that are not distributed or indexed by commercial publishers. Many of these documents are difficult to locate and obtain. Rather than compete with the published liter-ature, grey literature has the potential to complement and communicate findings to a wider audience, as well as to reduce publication bias.

Constructing MeSH/EMTREE/keyword lists

- Conduct basic search to ascertain what you are looking for i.e. Terminology
- Construct MeSH/EMTREE/Subject headings, keywords or text words for your search

Table 2. Joanna Briggs Institute Levels of Evidence

Levels of Evidence	Feasibility F (1-4)	Appropriateness A (1-4)	Meaningfulness M (1-4)	Effectiveness E (1-4)	Economic Evidence
1	Metasynthesis of research with unequivocal synthesized findings	Metasynthesis of research with unequivocal synthesized findings	Metasynthesis of research with unequivocal synthesized findings	Meta-analysis (with homogeneity) of experimental studies (e.g. RCT with concealed randomization) OR One or more large experimental studies with narrow confidence intervals	Metasynthesis (with homogeneity) of evaluations of important alternative interventions comparing all clinically relevant outcomes against appropriate cost measurement, and including a clinically sensible sensitivity analysis
2	Metasynthesis of research with credible synthesized findings	Metasynthesis of research with credible synthesized findings	Metasynthesis of research with credible synthesized findings	One or more smaller RCTs with wider confidence intervals OR Quasi-experimental studies (without randomization)	Evaluations of important alternative interventions comparing all clinically relevant outcomes against appropriate cost measurement, and including a clinically sensible sensitivity analysis
3	a. Metasynthesis of text/opinion with credible synthesized findings b. One or more single research studies of high quality	a. Metasynthesis of text/opinion with credible synthesized findings b. One or more single research studies of high quality	a. Metasynthesis of text/opinion with credible synthesized findings b. One or more single research studies of high quality	a. Cohort studies (with control group) b. Case-controlled c. Observational studies (without control group)	Evaluations of important alternative interventions comparing a limited number of appropriate cost measurement, without a clinically sensible sensitivity analysis
4	Expert opinion	Expert opinion	Expert opinion	Expert opinion, or physiology bench research, or consensus	Expert opinion, or based on economic theory

Why do databases index?

Indexing or subject listing of information is very important. It is a way to control data, to give it commonality, consistency and decrease the instances of spelling and cultural differences. It insures that you are able to retrieve all of the information on a given topic. Other disciplines use similar methods, for example pharmaceutical or chemistry names will vary from product name country to country, company to company or compound name. MEDLINE and EMBASE and several other databases have standardized subject terms as a controlled vocabulary or thesaurus. This is to minimise missing information because different words have been used to describe the same concept. However, it is worth noting that an indexer still uses his or her discretion and either extensive or limited knowledge about a subject to assign a heading to a concept.

MEDLINE and EMBASE have different approaches to indexing and it can be said that EMBASE is more technical particularly when referring to the pharmacological or pharmaceutical areas. The search fields are also different in each database. It is not always advisable to use the 'explode' subject terms initially, so as to include more specific terms automatically in the search. However, be aware that this can skew your search away from the specific search that you are looking for. Other indexing terms include Chemical Abstracts Service use Registry numbers to index their compounds or drug names.

For example:

To look for Atacand you could use all of the following terms, candesartan cilexetil, Kenzen, Antihypertensive Agents,1-(cyclohexylocarbonyloxy)ethyl-2-ethoxy-1-(2′-(1H-tetrazol-5-yl) biphenyl-4-yl)-1H-benzimidazole-7-carboxylate or the better way would be to look for the indexing term, the CAS Registry number 145040-37-5

MeSH terms display hierarchically by category, with more specific terms arranged beneath broader terms.

Many databases have shortcuts or abbreviations to help you efficiently search. For example abbreviated subject headings:

- cystic fibrosis/th (for therapy)
- asthma/dt,pc (for drug therapy or prevention and control)
- carcinoma/et (etiology)

When gathering your initial information to help you structure your search strategy and find all the terms to be used in the search it is often useful to draw out a form of concept map. Sometimes it is easier to see a diagram of what you are looking for to help you visually group words in and map your PICO with your MeSH and keywords. Another great thing is to remind yourself of spelling differences and any other clues that you have found to be of use.

What follows is an *example* of a search strategy in OVID Medline, broken down into sections with each section searched separately and the combined at the end. This search strategy was used for a systematic review commissioned by the NHMRC and undertaken at JBI.

Setting
1. hospital.mp. or exp Hospitals/
2. sub acute care.mp. or exp Sub acute Care/

3. residential facility.mp. or exp Residential Facilities/

4. intensive care unit.mp. or exp Intensive Care Units/

5. intensive care neonatal.mp. or exp Intensive Care, Neonatal/

6. intensive care units pediatric.mp. or exp Intensive Care Units, Pediatric/

7. exp Burn Units/ or burns unit.mp.

8. residential facilities.mp. or exp Residential Facilities/

9. exp Community Health Centers/ or exp Family Practice/ or community health centres.mp. or exp Primary Health Care/ or exp Rural Health Services/

10. exp Dental Care/

11. exp Long-Term Care/ or long.mp.

12. homes for the aged.mp. or exp Homes for the Aged/

13. nursing homes.mp. or exp Nursing Homes/

14. health services for the aged.mp. or exp Health Services for the Aged/

15. community health nursing.mp. or exp Community Health Nursing/

16. 6 or 11 or 3 or 7 or 9 or 12 or 2 or 15 or 14 or 8 or 1 or 4 or 13 or 10 or 5

Bacteria

17. staphylococcus aureus.mp. or exp Staphylococcus aureus/

18. staphylococcal infections.mp. or exp Staphylococcal Infections/

19. (staphylococc* and Infectio*).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

20. (staphylococc* and (bacteremia or bacteraemia)).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

21. 17 or 18 or 20 or 19

Drug Resistance

22. methicillin resistance.mp. or exp Methicillin Resistance/

23. (methicillin* and resistan*).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

24. multi drug resistan*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

25. antibiotic resistan*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

26. mrsa.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

27. ((meticillin or metacillin) and resistan*).mp. [mp=title, original title, abstract, name of substance word, subject heading word]

28. 22 or 23 or 24 or 25 or 26 or 27

Personal/protective equipment

29. protective clothing.mp. or Protective Clothing/

30. glove*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

31. exp Gloves, Protective/

32. exp Gloves, Surgical/ or gloves.mp.

33. gown*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

34. apron*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

35. Masks/ or masks.mp.

36. mask*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

37. barrier.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

38. contact precaution*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

39. universal precaution*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

40. droplet precaution*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

41. airborne precaution*.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

42. 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41

Study Design

43. clinical trial.mp. or exp Clinical Trial/

44. randomized.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

45. placebo.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

46. randomized.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

47. 44 or 46

48. randomly.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

49. trial.mp. [mp=title, original title, abstract, name of substance word, subject heading word]

50. 46 or 45 or 42 or 40 or 44

51. Humans/

52. 48 and 47

Combining

53. 21 AND 28 (Bacterium AND Resistance)

54. 16 AND 53 AND 42 AND 52

(Setting AND (Bacterium AND Resistance) AND Personal Protective Equipment AND Study Design)

Assessment of Methodological Quality

The basis for inclusion (and exclusion) of studies in a systematic review needs to be transparent and clearly documented in the protocol. A systematic review aims to synthesis the best available evidence; therefore the review should aim to include the highest quality of evidence possible. Methodological quality is determined by critical appraisal using validated tools. There are a variety of checklists and tools available to assess the validity of studies. Most of these use a series of criteria that can be scored as being met, not met or unclear. The decision as to whether or not to include a study can be made based on meeting a pre-determined proportion of all criteria, or on certain criteria being met. It is also possible to weight the different criteria differently, for example blinding of assessors (to prevent detection bias) may be considered to be twice as important as blinding the caregivers (to prevent performance bias). It is important that appraisal tools are appropriate for the design of the study; this is so that the questions of the tool are specific to that study design (Joanna Briggs Institute 2008).

The decisions about the scoring system and the cut-off for inclusion should be made in advance, and be agreed upon by all participating reviewers before critical appraisal commences. It is JBI policy that all study types must be critically appraised using the standard critical appraisal instruments for specific study designs, built into the analytical modules of the JBI SUMARI software. The protocol must therefore describe how the methodological quality/validity of primary studies will be assessed; any exclusion criteria based on quality considerations (Joanna Briggs Institute 2008).

The main object of critical appraisal is to assess the methodological quality of a study and to determine the extent to which a study has addressed the possibility of bias in its design, conduct and analysis. If a study has not excluded the possibility of bias, then its results are questionable and could well be invalid. Therefore, part of the systematic review process is to evaluate how well the potential for bias has been excluded from a study, with the aim of only including high quality studies in the resulting systematic review. A secondary although no less strategic benefit of critical appraisal is to take the opportunity to ensure each retrieved study has included the population, intervention and outcomes of interest specified in the review.

The best study design in terms of excluding bias is the double blinded randomized placebo controlled trial (RCT) (Higgins and Green 2008; Joanna Briggs Institute 2008). Nevertheless, there are four main forms of bias that can affect any RCT: selection bias, performance bias, attrition bias and detection bias:

- Selection bias refers chiefly to whether or not the assignment of participants to either treatment or control groups (e.g. in a comparison of only two groups) has been made so that all potential participants have an equal chance of being assigned to either group, and that the assignment of participants is concealed from the researchers, at least until the treatment has been allocated.
- Performance bias refers to differences in care provided to patients if the caregiver is aware of whether a patient is in a control or treatment group.
- Attrition bias refers to differences between control and treatment groups in terms of patients dropping out of a study, or not being followed up as diligently.
- Detection bias occurs if an assessor evaluates an outcome differently for patients depending on whether they are in the control or treatment group.

Critical appraisal tools are included in the MAStARI program, and can be completed electronically for RCTs, case-control/cohort studies and descriptive/case series studies. JBI-MAStARI has been designed with the intention that there will be at least two reviewers (a primary and a secondary) independently conducting the critical appraisal. The secondary reviewer can only conduct their appraisal after the primary reviewer has completed theirs; the secondary reviewer is blinded to the findings of the primary reviewer. Once the secondary reviewer has completed their appraisal, the primary reviewer compares the two appraisals. The two reviewers should discuss cases where there is a lack of consensus in terms of whether a study should be included; it is appropriate to seek assistance from a third reviewer as required (Higgins and Green 2008; Joanna Briggs Institute 2008).

Data extraction

Data extraction refers to the process of identifying and recording relevant details from original (eg primary) research studies that will be included in the systematic review. A standardized extraction tool is used to minimise the risk of error when extracting data. Other error-minimising strategies include; ensuring that both reviewers have practiced using the extraction tool and can apply the tool consistently. It is also recommended that reviewers extract data independently before conferring. These strategies aim to facilitate accurate and reliable data entry in to JBI-CReMS for analysis.

Details regarding the participants, the intervention, the outcome measures and the results are to be extracted from included studies. It is JBI policy that data extraction for all study types must be carried out using the standard data extraction instruments for specific study designs, built into the analytical modules of the JBI SUMARI software. The protocol must therefore describe how data will be extracted and include the appropriate JBI data extraction instruments as appendices to the protocol.

Studies may include several outcomes; however the review should focus on extracting information related to the research questions and outcomes of interest. Information that may impact upon the generalizability of the review findings such as study method, setting and population characteristics should also be extracted and reported. Population characteristics include factors such as age, past medical history, co-morbidities, complications or other potential confounders.

The data extracted will vary depending on the review question; however it will generally either be dichotomous or continuous in nature. Dichotomous data will include the number of participants with the exposure/intervention (n) and the total sample (N) for both control and treatment groups. Classically, this is stated as n/N; therefore there will be two columns of data for each outcome of interest.

For continuous data, the mean and standard deviation (SD), plus sample size are extracted for each specified outcome for both the control and intervention (or exposure) group. Typically, this is expressed as mean (SD)/n where n = the sample size for the particular group. If the standard error (SE) only is reported, the SD can be calculated from the SE, as long as the sample size (n) is known. The equation for converting from SE to SD is simply: $\mathbf{SD = SE \times \sqrt{n}}$

In some cases it may not be possible to extract all necessary raw data from an included study for a systematic review, as sometimes only aggregated data are reported, or perhaps

data from two different patient populations have been combined in the data analysis, and your review is focussed on only one of the patient populations. In these circumstances, the standard approach is to make contact with the authors of the publication and seek their assistance in providing the raw data. Most researchers are obliging when it comes to these requests providing that records are still available. If the study authors do not respond or if the data is unavailable, this should be noted in the report.

In addition to the data, conclusions that study authors have drawn based on the data are also extracted. It is useful to identify the study authors' conclusions and establish whether there is agreement with conclusions made by the reviewer authors.

Data synthesis

The protocol should also detail how the data will be combined and reported. A synthesis can either be descriptive (narrative synthesis) or statistical (meta analysis). Statistical combination of study data provides a summary estimate, using transparent rules specific in advance (Borestein, Hedges et al. 2009). This allows an overall effect of a treatment/intervention to be determined. Whilst the ultimate aim for a quantitative systematic review is to combine study data in meta analysis, this is not always appropriate or possible. Data from two or more separate studies are required to generate a synthesis.

The three main areas surrounding data that should be considered when deciding whether or not to combine data are:

- Clinical – are the patient characteristics similar? (such as age, diagnoses, co-morbidities, treatments).
- Methodological – do the studies use the same study design and measure the same outcomes?
- Statistical – were outcomes measured in the same way, at the same time points, using comparable scales?

It is important to combine the studies in an appropriate manner using methods appropriate to the specific type and nature of data that has been extracted. In the protocol, the methods by which studies will be combined should be described in as much detail as is reasonably possible. As bullet points below indicate, this may require describing the approaches for both dichotomous and continuous data (Borestein, Hedges et al. 2009).

- which test of statistical heterogeneity is to be used (such as Chi square or I2),
- at which point statistical heterogeneity is considered significant,
- whether fixed or random effects models will be utilized and which specific methods of meta analysis may be used for the anticipated types of data (i.e. continuous or dichotomous).

Where possible study results should be pooled in statistical meta-analysis using (for reviews conducted through a Cochrane Review Group) Review Manager or the Joanna Briggs Institute Meta Analysis of Statistics Assessment and Review Instrument (JBI-MAStARI). All numeric outcome data must be double entered to prevent data entry errors. Odds ratio (for categorical data) and standard or weighted mean differences (for continuous data) and their 95% confidence intervals should be calculated for analysis. Heterogeneity should be assessed using

the standard Chi-square or I2. Where statistical pooling is not possible the findings should be presented in narrative summary.

When used in relation to meta-analysis, the term 'heterogeneity' refers to the amount of variation in the characteristics of included studies. For example, if three studies are to be included in a meta-analysis, do each of the included studies have similar sample demographics, and assess the same intervention? (Note that the method by which the intervention is measured does not need to be identical.) While some variation between studies will always occur due to chance alone, heterogeneity is said to occur if there are significant differences between studies, and under these circumstances meta-analysis is not valid and should not be undertaken. But how does one tell whether or not differences are significant?

Visual inspection of the meta-analysis is the first stage of assessing heterogeneity. JBI-MAStARI plots the results of individual studies and thus indicates the magnitude of any effect between the treatment and control groups. Do the individual studies show a similar direction and magnitude of effect – i.e. are the rectangular symbols at similar positions on the X-axis?

A formal statistical test of the similarity of studies is provided by the test of homogeneity (Higgins 2002). This test calculates a probability (P value) from a Chi-square statistic calculated using estimates of the individual study weight, effect size and the overall effect size. However, note that this test suffers from a lack of power – and will often fail to detect a significant difference when a difference actually exists – especially when there are relatively few studies included in the meta-analysis. Because of this low power, some review authors use a significance level of $P < 0.01$, rather than the conventional 0.05 value, in order to protect against the possibility of falsely stating that there is no heterogeneity present (Hardy and Thompson 1998).

In meta-analysis, the results of similar, individual studies are combined to determine the overall effect of a particular form of health care intervention (the treatment) compared to another standard or control intervention for a specified patient population and outcome. In meta-analysis, the effect size and weight of each study are calculated. The effect size indicates the direction and magnitude of the results of a particular study (i.e. do the results favour the treatment or control, and if so, by how much), while the weight is indicative of how much information a study provides to the overall analysis when all studies are combined together.

While meta-analysis is the ultimate goal of a systematic review of quantitative studies, a number of criteria must first be met before the results of different studies can be validly combined. Studies to be included in meta-analysis should be similar to each other so that generalization of results is valid. The four main criteria that must be considered are:

- patient population (e.g. is it valid to combine the results of studies on different races of people, or different aged people?)
- intervention (e.g. are the interventions being given to the 'treatment' group in each study similar enough to allow meta-analysis?)
- control (e.g. are the control groups in each study receiving treatment similar enough to warrant combination and meta-analysis?)
- outcome (e.g. is it valid to combine studies that have measured pain via a visual analogue scale with those that have used a pain diary?)

The questions raised above can be very difficult to answer and often involve subjective decision-making. Involvement of experienced systematic reviewers and/or researchers with a good understanding of the clinical question being investigated should help in situations where judgement is required. Borenstein et al (2009) also provide a good reference. These situations should be clearly described and discussed in the systematic review report (Borestein, Hedges et al. 2009).

Deeks and Altman (2001) suggest three important criteria for choosing a summary statistic for meta-analysis: (i) consistency of effect across studies, (ii) mathematical properties, and (iii) ease of interpretation (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001).

(i) Consistency of effect is important because the aim of meta-analysis is to bring together the results of several studies into a single result. The available evidence suggests that relative measures of effect such as the odds ratio (OR) and relative risk (RR) are more consistent than absolute measures (absolute measures of effect include the risk difference and the number needed to treat – these are not currently included as analytical options in JBI-CReMS/MAStARI and thus will not be discussed further) (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001). There is little difference between the RR and OR in terms of consistency between studies.

(ii) The main mathematical property required by summary statistics is the availability of a reliable variance estimate, a feature of both OR and RR. Consensus about the other two mathematical properties (reliance on which of the two outcome states [e.g. mortality/survival] is coded as the event, and the OR being the only statistic which is unbounded) has not yet been reached (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001).

(iii) Ease of interpretation does vary between OR and RR. Most clinicians and lay readers can intuitively grasp the concept of being at risk of an outcome more easily than the odds of an outcome occurring. Misinterpretation of an OR as an RR will usually result in an overestimation of the effect size, suggesting that the treatment is better or worse than it actually is (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001). For this reason, RR are preferred; however, meta-analysis of OR is still valid. When meta-analysis of OR is conducted, reviewers should be careful to explain how odds ratios should be interpreted, and differences between OR and RR when outcomes are common.

Statistical assumptions in meta-analysis

Effect size

The effect size has been described as being the "currency" of the systematic review, it statistically describes the relationship between two variables (Borestein, Hedges et al. 2009). It is calculated for each included study and then a summary effect size is calculated to examine the relationship across the studies. The effect size could be a single number such as a prevalence or a ration such as a risk ratio.

Meta-analysis can be based on either of two statistical assumptions – fixed or random effects. The fixed effect model assumes that there is one true effect underlying the studies in the

analysis and that all differences in the data are due to sampling error or chance and that there is no heterogeneity between the studies (Borestein, Hedges et al. 2009). A fixed effect model is statistically stringent and should be used when there is little heterogeneity, as determined by Chi square (or I2).

A random effects model allows more flexibility, assuming that there may be other factors influencing the data than error or chance. For example, the effect size may be influenced in studies where the participants are more educated, older or healthier or if a more intense intervention is being used (Borestein, Hedges et al. 2009). The effect size is assumed to follow a normal distribution and consequently has a mean and variance.

Essentially, the test for homogeneity is asking the statistical question "is the variance around the estimate of the effect size zero or non zero?" If the variance around the estimate of the effect size is zero, then there is no heterogeneity present, and the results of the fixed and random effects models will be similar.

There is no consensus about whether fixed or random effects models should be used in meta-analysis. In many cases when heterogeneity is absent, the two methods will give similar overall results. When heterogeneity is present, the random effects estimate provides a more conservative estimate of the overall effect size, and is less likely to detect significant differences. For this reason, random effects models are sometimes employed when heterogeneity is not severe; however, the random effects model does not actually analyse the heterogeneity away and should not be considered as a substitute for a thorough investigation into the reasons for the heterogeneity (Hardy and Thompson 1998). Additionally, random effects models give relatively more weight to the results of smaller studies – this may not be desirable because smaller studies are typically more prone to bias and of lower quality than larger studies (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001).

There are a number of meta-analytical techniques available. The selection of a particular technique is governed by three things: the study type, nature of the data extracted and assumptions underlying the meta-analysis. The following paragraphs will introduce the tests that are available in JBI-MAStARI and when it is appropriate to use each of the tests.

When the outcomes of included studies are dichotomous, JBI-MAStARI can be used to generate two overall effect sizes: odds ratios (OR) and relative risks (also known as risk ratios, RR). The choice of whether OR or RR are calculated is important and should be carefully considered with due reference to three criteria.

Dichotomous data – methods of meta-analysis

There are several different methods available to pool results of dichotomous data, depending on the data type and whether a random or fixed effects model is required: Mantel-Haenszel, Peto's and DerSimonian and Laird.

Mantel-Haenszel is the default meta-analytical method for dichotomous data using a fixed effects model. Both OR and RR can be pooled using Mantel-Haenszel methods; the calculation of study weights and effect sizes, and overall effect sizes differs slightly between OR and RR. The Mantel-Haenszel method is generally preferred in meta-analysis to another method (inverse variance) because it has been shown to be more robust when data are sparse (in

terms of event rates being low and/or the number of trials being small) (Deeks, Higgins et al. 2001).

Peto's odds ratio is an alternative method for meta-analysis of OR using a fixed effects method. It employs an approximation that can be inaccurate if treatment affects are very large, and when the sample sizes between treatment and control groups are unbalanced. However, the method is appropriate when event rates are very low and effect sizes are not overly large (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001).

DerSimonian and Laird methods are used in the meta-analysis of OR and RR using a random effects model. Although the study effect sizes and heterogeneity statistics are calculated as for the fixed effects model, the study weights and overall effect sizes in DerSimonian and Laird random effects models are calculated slightly differently to fixed models.

Meta-analysis of continuous data

When the outcomes of included studies are continuous, JBI-MAStARI can be used to generate two overall effect size calculations using weighted mean differences (WMD) or standardized mean differences (SMD). The WMD measures the difference in means of each study when all outcome measurements are made using the same scale. It then calculates an overall difference in mean for all studies (this is equivalent to the effect size) based on a weighted average of all studies, which is, in turn related to the SD. JBI-MAStARI uses the inverse variance method of calculating WMD for fixed effects models and the DerSimonian and Laird method for random effects models.

Alternatively, different studies may measure the same outcome using different scales. For example, pain can be measured on a range of different scales including non-verbal scales (e.g. visual analogue scale) and verbal scales (e.g. 5 point categorical scale). These studies can be combined in a meta analysis that incorporates SMD. If the measurement scales operate in the same direction (e.g. an increase in pain is measured as an increase in on both scales), then using SMD is straightforward. However, if two measurement scales operate in a different direction – for example a score of 10 is the worst pain imaginable on one scale but a score of 1 is the worst pain imaginable on another scale – then data from one scale need to be reversed. This is relatively simply achieved by multiplying the mean data from one scale (for both treatment and control groups) by -1. Standard deviations do not need to be modified.

There are two relatively common options for calculation of the SMD using fixed effects: Cohen's SMD and Hedges' SMD. Both options produce a similar result, although Hedges' SMD is generally preferred as it includes an adjustment to correct for small sample size bias (Deeks, Higgins et al. 2006). As per WMD, the DerSimonian and Laird method is used for random effects models calculations for SMD.

Narrative Summary

Although the focus of this section has been on describing and explaining the types of meta-analysis, where meta-analysis is not possible the protocol should describe a process for narrative summary. Narrative summary should draw upon the data extraction, with an emphasis on textual summary of study characteristics as well as data relevant to the specified outcomes.

Conflict of Interest

A statement should be included in every review protocol that either declares the absence of any conflict of interest, or describes a specified or potential conflict of interest. Conflict of interest statements should adhere to the guidelines of the International Committee of Medical Journal Editors (ICMJE) for individual authors and project support (http://www.icmje.org/ethical_4conflicts.html). Additionally, the Committee on Publication Ethics (COPE) have extensive guidelines for conflict of interest statements that are intended to protect the authors as well as the readers, and review authors should ensure they are familiar with and adhere to the principals described within the COPE framework (http://www.publicationethics.org/).

Acknowledgements

The source of financial grants and other funding must be acknowledged, including the reviewers' commercial links and affiliations. The contribution of colleagues or Institutions should also be acknowledged.

References

Protocols most often use Vancouver style referencing. References should be numbered in the order in which they appear with superscript Arabic numerals in the order in which they appear in text. Full reference details should be listed in numerical order in the reference section.

More information about the Vancouver style is detailed in the International Committee of Medical Journal Editors' revised 'Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication', and can be found at http://www.ICMJE.org/

Appendices

Appendices should be placed at the end of the protocol and be numbered with Roman numerals in the order in which they appear in text. At a minimum this will include database specific search strategies, critical appraisal and data extraction tools as well as tables of included and excluded studies (with the rationale for exclusion on a per paper basis).

3

Conducting a Systematic Review of Quantitative Evidence

Chapter 5:

Searching for Quantitative Evidence

Searching in systematic reviews is characterized by particular criteria, which, of themselves are significant factors in the scientific validity of systematic reviews. These include requirements that searches be comprehensive, exhaustive, and transparent, or auditable. Search strategies should also be developed in accordance with the particular structure of each individual database that is being included in the review.

There are a huge number of such strategies freely available online. Most are based on the same principals – giving preference to study designs that either are at least risk of bias, or in the case of reviews of questions broader than clinical effectiveness, they focus on designs that respond most fully to the nature or scope of the question. Search strategies are a balance between sensitivity, and specificity, where sensitivity returns a high number of relevant studies, but also a higher number of studies not relevant to your question. Conversely, specificity results in a small number of total studies, but these are more relevant to your particular question. There are formula for calculating the sensitivity and specificity of search strategies, although it is worth discussing the relevance of formula with a librarian or information scientist before embarking on the process of calculating these values to be included in your search strategy reporting. Given numerous organizations have undertaken the testing of methodological search filters, the pragmatic reviewer would look first to compare the findings across key organizations rather than attempt this work alone.

The aim of a systematic review is to identify all relevant international research on a given topic. This is done by utilising a well-designed search strategy across a breadth of resources. There is insufficient evidence to suggest a particular number of databases, or that even whether particular databases provide sufficient topic coverage, therefore, literature searching should be based on the principal of inclusiveness - with the widest reasonable range of databases included that are considered appropriate to the focus of the review. If possible, authors should seek the advice of a research librarian to aid construction of a search strategy.

The protocol should provide a detailed strategy including the search terms to be used and the resources (e.g. electronic databases and specific journals, websites, experts, etc) to be

searched. Within systematic reviews, the search strategy is often described as a three-phase process beginning with the identification of initial key words followed by analysis of the text words contained in the title and abstract, and of the index terms used to describe relevant articles. The second phase is to construct database-specific searches for each database included in protocol, and the third phase is to review the reference lists of all studies that are retrieved for appraisal to search for additional studies.

Subject Qualifiers that may be helpful in constructing a search strategy

Subject qualifiers are a type of hierarchy in MeSH. Qualifiers are used to refine your search by subject heading. For example, if your topic of interest was wound care, and you added the subject qualifier <adverse events> to the subject heading Wound Care, you will retrieve records of documents that discuss adverse events in relation to wound care. Key subject qualifiers used in searching vary from database to database, and should only be used within the particular database, rather than across multiple databases, examples from the National Library of Medicine's MEDLINE database include:

- ab = words in abstract
- exp = before an index term indicates that the term was exploded
- hw = word in subject heading
- mp = free text search for a term
- pt = publication type · *sh = subject heading
- ti = words in title
- tw = text words in title/abstract
- ? = in middle of term indicates use of a wildcard
- / = MeSH subject heading (and includes all subheadings being selected)
- $ = truncation symbol
- adj = two terms where they appear adjacent to one another (so adj4, for example, is within four words)

Search filters for methodology

Search filters are pre-tested strategies that identify articles based on criteria such as specified words in the title, abstract and keywords. They can be of use to restrict the number of articles identified by a search from the vast amounts of literature indexed in the major medical databases. Search filters look for sources of evidence based on matching specific criteria – such as certain predefined words in the title or abstract of an article. Search filters have strengths and weaknesses:

- Strengths: they are easy to implement and can be pre-stored or developed as an interface
- Limitations: database-specific; platform-specific; time-specific; not all empirically tested and therefore not reproducible; assume that articles are appropriately indexed by authors and databases.

Search filters tend to be based on particular aspects of study design, and these forms of filter have been investigated and tested over time with regard to their sensitivity and specificity. The filters for quantitative evidence are generally based on the hierarchy of evidence and therefore, preference the randomized controlled trial and other controlled designs over lower levels of evidence that are at increased risk of bias.

The example provided below has been extracted from the Scottish Intercollegiate Guidelines Network (date accessed 3rd July, 2011) website (http://www.sign.ac.uk/methodology/filters.html). The inclusion in this text is not a particular endorsement of this methodological search filter in preference to any other search filter. It is recommended that you read widely, particularly with regard to the sensitivity and precision of search filters before choosing one for your review. The search filter used by SIGN to retrieve randomized controlled trials has been adapted from the first two sections of strategy designed by the Cochrane Collaboration identifying RCTs for systematic review.

Medline

1	Randomized Controlled Trials as Topic/
2	randomized controlled trial/
3	Random Allocation/
4	Double Blind Method/
5	Single Blind Method/
6	clinical trial/
7	clinical trial, phase i.pt
8	clinical trial, phase ii.pt
9	clinical trial, phase iii.pt
10	clinical trial, phase iv.pt
11	controlled clinical trial.pt
12	randomized controlled trial.pt
13	multicenter study.pt
14	clinical trial.pt
15	exp Clinical Trials as topic/
16	or/1-15
17	(clinical adj trial$).tw
18	((singl$ or doubl$ or treb$ or tripl$) adj (blind$3 or mask$3)).tw
19	PLACEBOS/
20	placebo$.tw
21	randomly allocated.tw
22	(allocated adj2 random$).tw
23	or/17-22
24	16 or 23
25	case report.tw
26	letter/
27	historical article/
28	or/25-27
29	24 not 28

By way of contrast, the Cochrane search strategy for RCTs using the OVID platform is much shorter. It has higher sensitivity than precision (accessed 3rd July, 2011; http://www.cochrane-handbook.org/).

1   randomized controlled trial.pt.
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

Ovid search syntax

.pt. denotes a Publication Type term;

.ab. denotes a word in the abstract;

.fs. denotes a 'floating' subheading;

.sh. denotes a Medical Subject Heading (MeSH) term;

.ti. denotes a word in the title.

These terms and filters are specific to randomized controlled trials in humans. As JBI reviews are of broader evidence than randomized controlled trials, examples of methodological filters for observational studies have also been included. It is worth noting and being aware of the differences between major databases in how words are categorized and used. These differences highlight the value in performing a full and detailed search within each individual database rather than across 2 or more databases at the same time.

The Observational Studies search filter used by SIGN was developed in-house to retrieve studies most likely to meet SIGN's methodological criteria.

Medline

1   Epidemiologic studies/
2   Exp case control studies/
3   Exp cohort studies/
4   Case control.tw.
5   (cohort adj (study or studies)).tw.
6   Cohort analy$.tw.
7   (Follow up adj (study or studies)).tw.
8   (observational adj (study or studies)).tw.
9   Longitudinal.tw.
10 Retrospective.tw.
11 Cross sectional.tw.
12 Cross-sectional studies/
13 Or/1-12

Embase
1   Clinical study/
2   Case control study
3   Family study/
4   Longitudinal study/
5   Retrospective study/
6   Prospective study/
7   Randomized controlled trials/
8   6 not 7
9   Cohort analysis/
10   (Cohort adj (study or studies)).mp.
11   (Case control adj (study or studies)).tw.
12   (follow up adj (study or studies)).tw.
13   (observational adj (study or studies)).tw.
14   (epidemiologic$ adj (study or studies)).tw.
15   (cross sectional adj (study or studies)).tw.
16   Or/1-5,8-15

Chapter 6:
Selecting and critically appraising studies

When the search for evidence is complete (or as the search progresses in some cases) reviewers decide which papers found should be retrieved and then subjected to critical appraisal. This initial process is referred to as the selection of papers for appraisal. All selected papers are then subjected to critical appraisal to determine methodological quality.

Selecting studies

Study selection is an initial assessment that occurs following the review search addressing the simple question: "should the paper be retrieved?" Studies in a review will also undergo another 'round' of selection in the next systematic step in the review process (Joanna Briggs Institute 2008). This second round of assessment asks a different question: "should the study be included in the review?" - this is critical appraisal. Study selection is performed with the aim of selecting only those studies that address the review question and that match the inclusion criteria documented in the protocol of your review. Two assessors, to limit the risk of error, should perform the process. Both assessors will scan the lists of titles, and if necessary abstracts, to determine if the full text of the reference should be retrieved. Sometimes it will be difficult or impossible to determine if the reference matches the inclusion criteria of the review on the basis of the title or abstract alone; in this case the full text should be retrieved for further clarification. It is best to err on the side of caution in this process (Joanna Briggs Institute 2008). It is better to spend a bit more time here, in careful consideration, rather than risk missing important and relevant evidence related to the review question. The entire process must be transparent and clear so that if an independent person were to apply the same inclusion criteria to the same list of citations, they would arrive at the same result of included studies.

Assessment of methodological quality/critical appraisal

A description of how methodological assessment was determined should be included, with reference to the JBI critical appraisal tool(s) used. A copy of the tool(s) should be included in the appendix section. As discussed in the section on protocol development, it is JBI policy that all study types must be critically appraised using the critical appraisal instruments for specific study designs incorporated in to the analytical modules of the JBI SUMARI software (Joanna Briggs Institute 2008). The primary and secondary reviewer should discuss each item of appraisal for each study design included in their review.

In particular, discussions should focus on what is considered acceptable to the needs of the review in terms of the specific study characteristics such as randomization or blinding in RCTs (Joanna Briggs Institute 2008). The reviewers should be clear on what constitutes acceptable levels of information to allocate a positive appraisal compared with a negative, or response of "unclear". This discussion should take place before independently conducting the appraisal.

The JBI-MAStARI Approach to Critical Appraisal of trials and observational studies

RCTs and quasi (pseudo) RCTs/CCTs provide the most robust form of evidence for effects because they provide evidence related to whether or not a causal relationship exists between a stated intervention, and a specific, measurable outcome, and the direction and strength of the relationship. Properly performed RCTs reduce bias, confounding factors, and results by chance. They have three essential elements

- Randomization (where possible);
- Researcher-controlled manipulation of the independent variable; and
- Researcher control of the experimental situation

RCTs are often used to evaluate how effective a new treatment/therapy/intervention is for patients with a certain condition. Individuals (or other units) are randomly allocated to a treatment group. Randomization is essential as this ensures that all treatment groups are comparable at the beginning. Confounding factors (variables), which may somehow impact upon the results of the study such as age, gender, etc will be spread evenly across groups to ensure treatment arms are as comparable as possible prior to receiving the intervention. Properly designed and performed randomized controlled trials reduce the risk of bias, confounding factors, and results by chance. However, poorly conducted randomized controlled trials are susceptible to bias and may produce misleading information or exaggerated treatment effects (Altman, Schulz et al. 2001; Moher, Schulz et al. 2001; Kao, Tyson et al. 2008).

The Critical Appraisal Criteria for Randomized Controlled Trials

There are 10 questions to guide the appraisal of randomized and quasi-randomized controlled trials.

1. Was the assignment to treatment groups truly random?

There are three broad types of randomization within trials, randomization, quasi- (or pseudo) and stratified randomization. True randomization occurs when every patient has a truly equal chance of being in any group included in the trial. This may involve using computer generated allocation methods to ensure allocation is truly random. The Consort criteria for randomization (http://www.consort-statement.org/index.aspx?o=1025, accessed 21st Sept, 2007) state that:

> With random allocation, each participant has a known probability of receiving each treatment before one is assigned, but the actual treatment is determined by a chance process and cannot be predicted. (Higgins and Green 2008)

Consort emphasizes that true randomization will minimize selection bias, thus identification of the method of randomization provides reviewers with a good indication of study quality. In the presence of true randomization, the sample is said to be representative of the population of interest, with homogeneity of characteristics at baseline. Hence any variation between groups in the trial would be expected to reflect similar differences in the relevant population.

In quasi-randomization, allocation is not truly random, being based on a sequential method of allocation such as birth date, medical record number, or order of entry in to the study

(alternate allocation). These methods may not conceal allocation effectively; hence there is an increased risk of selection bias associated with their usage.

The third type of randomization commonly utilized in randomized trials is stratification. Stratification may be used where a confounding factor (a characteristic that is considered likely to influence the study results, i.e. medications or co-morbidities) needs to be evenly distributed across groups.

Whichever approach to randomization was used, it should be described with sufficient detail to enable reviewers to determine whether the method used was sufficient to minimize selection bias. Authors of primary studies have competing interests in describing their methods, the need to be descriptive at times conflicts with the need to fit within word limits. However, brevity in the methods often leaves reviewers unable to determine the actual method of randomization. Generalist phrases such as "random", "random allocation" or "randomization" are not sufficient detail for a reviewer to conclude randomization was "truly random", it is then up to the reviewer to determine how to rank such papers. This should be raised in initial discussion between the primary and secondary reviewers before they commence their independent critical appraisal.

2. Were participants blinded to treatment allocation?

Blinding of participants is considered optimal as patients who know which arm of a study they have been allocated to may inadvertently influence the study by developing anxiety or conversely, being overly optimistic, attempting to "please" the researchers. This means under- or over-reporting outcomes such as pain or analgesic usage; lack of blinding may also increase loss to follow-up depending on the nature of the intervention being investigated.

3. Was allocation to treatment groups concealed from the allocator?

Allocation is the process by which individuals (or groups if stratified allocation was used) are entered in to one of the study arms following randomization. The Cochrane Systematic Review handbook states:

> When assessing a potential participant's eligibility for a trial, those who are recruiting participants . . . should remain unaware of the next assignment in the sequence until after the decision about eligibility has been made. Then, after assignment has been revealed, they should not be able to alter the assignment or the decision about eligibility. The ideal is for the process to be impervious to any influence by the individuals making the allocation. (Higgins and Green 2008)

Allocator concealment of group allocation is intended to reduce the risk of selection bias. Selection bias is a risk where the allocator may influence the specific treatment arm an individual is allocated to, thus optimally, trials will report the allocator was unaware of which group all study participants were randomized to, and had no subsequent influence on any changes in allocation.

4. Were the outcomes of people who withdrew described and included in the analysis?

Commonly intention to treat analysis is utilized where losses to follow-up are included in the analysis. Intention to treat (ITT) analysis may reduce bias due to changes in the characteristics between control and treatment groups that can occur if people either drop out, or if there

is a significant level of mortality in one particular group. The Cochrane Systematic Review handbook identifies two related criteria for ITT analysis, although it is equally clear that how these criteria are applied remains an issue of debate:

Trial participants should be analyzed in the groups to which they were randomized regardless of which (or how much) treatment they actually received, and regardless of other protocol irregularities, such as ineligibility

All participants should be included regardless of whether their outcomes were actually collected (Higgins and Green 2008).

5. Were those assessing the outcomes blind to the treatment allocation?

In randomized controlled trials, allocation by a third party not otherwise directly involved in the implementation of the study is preferred. Where these resources are not available, electronic assignment systems may be described in trials. Inadequate blinding of allocation is associated with more favorable outcomes for the primary intervention of interest in RCTs (Kjaergard, Vilumsen et al. 1999; Higgins and Green 2008).

Reviewers should seek to establish whether those assessing outcomes were truly blinded to allocation. Some sources suggest blinded assessment reduces the risk of detection bias. Note that studies reporting multiple outcomes may be at risk of detection bias for some outcomes within a study, but not others. Therefore, attempts should be made to establish if outcomes assessors were blinded to all outcomes of interest to the review.

6. Were the control and treatment groups comparable at entry?

Homogeneity or comparability at entry is related to the method of allocation. If allocation was truly random, groups are more likely to be comparable as characteristics are considered to be randomly distributed across both groups. However, randomization does not guarantee comparability. Primary studies should report on the baseline characteristics of all groups, with an emphasis on any differences between groups that reach statistical probability.

7. Were groups treated identically other than for the named intervention?

Studies need to be read carefully to determine if there were any differences in how the groups were treated – other than the intervention of interest. If there was a difference in how the groups were treated that arose from flaws in the trial design, or conduct, this is known as a systematic difference and is a form of bias which will skew study results away from the accuracy the primary authors would otherwise have intended. Randomization, blinding and allocation concealment are intended to reduce the effects of unintentional differences in treatment between groups (Joanna Briggs Institute 2008).

8. Were outcomes measured in the same way for all groups?

In identifying how robust the outcomes for a study are, the definitions, scales and their values as well as methods of implementation of scales needs to be the same for all groups. This question should include consideration of the assessors, were they the same people or trained in the same way, or were there differences such as different type of health professionals involved in measurement of group outcomes (Joanna Briggs Institute 2008).

9. Were outcomes measured in a reliable way?

Were the instruments used to measure outcomes adequately described, and had they been previously validated, or piloted within the trial? These types of questions inform reviewers

of this risk to detection bias. Give consideration to the quality of reporting of findings. If an RCT reports percentage of change but gave no baseline data, it is not possible to determine the relevance of the reported value between groups (or within a single group). If a P value is reported but no confidence interval given, the significance has been established, but the degree of certainty in the finding has not (Joanna Briggs Institute 2008).

10. Was appropriate statistical analysis used?

As with any consideration of statistical analysis, consideration should be given to whether there was a more appropriate alternate statistical method that could have been used. Advice from a statistician may be needed to establish if the methods of analysis were appropriate.

The Critical Appraisal Criteria for Cohort Studies

Cohort (with control)/Case-controlled studies

Cohort studies compare outcomes in groups that did and did not receive an intervention or have an exposure. However, the method of group allocation in Cohort or Case-controlled studies is not random. Case-control or Cohort studies can be used to identify if the benefits observed in randomized trials translate into effectiveness across broader populations in clinical settings and provide information on adverse events and risks (Mamdani 2005; Normand 2005; Rochon 2005).

1. Is the sample representative of patients in the population as a whole?

This question relies upon knowledge of the broader characteristics of the population of interest. If the study is of women undergoing chemotherapy for breast cancer knowledge of at least the characteristics, demographics, medical history is needed. The term population as a whole should not be taken to infer every individual from everywhere subject to a similar intervention or with similar disease or exposure characteristics. Instead, give consideration to specific population characteristics in the study, including age range, gender, morbidities, medications, and other potentially influential factors.

2. Are the patients at a similar point in the course of their condition/illness?

Check the paper carefully for descriptions of diagnosis and prognosis to determine if patients within and across groups have similar characteristics in relation to disease or exposure, for example tobacco use.

3. Has bias been minimized in relation to selection of cases and controls?

It is useful to determine if patients were included in the study based on either a specified diagnosis or definition. This is more likely to decrease the risk of bias. Characteristics are another useful approach to matching groups, and studies that did not use specified diagnostic methods or definitions should provide evidence on matching by key characteristics.

4. Are confounding factors identified and strategies to deal with them stated?

Confounding has occurred where the estimated intervention effect is biased by the presence of some difference between the comparison groups (apart from the intended intervention/s). Typical confounders include baseline characteristics, prognostic factors, or concomitant interventions. A confounder is a difference between the comparison groups and it influences the direction of the study results. A high quality study at the level of cohort or case-control design

will identify the potential confounders and measure them (where possible). This is difficult for studies where behavioral, attitudinal or lifestyle factors may impact on the results.

5. Are outcomes assessed using objective criteria?

Refer back to item three of this appraisal scale and read the methods section of the paper again. If the outcomes were assessed based on existing definitions or diagnostic criteria, then the answer to this question is likely to be yes. If the outcomes were assessed using observer reported, or self reported scales, the risk of over- or under-reporting is increased, and objectivity is compromised. Importantly, determine if the measurement tools used were validated instruments as this has a significant impact on outcome assessment validity.

6. Was follow-up carried out over a sufficient time period?

The appropriate length of time for follow-up will vary with the nature and characteristics of the population of interest and/or the intervention, disease or exposure. To estimate an appropriate duration of follow-up, read across multiple papers and take note of the range for duration of follow-up. The opinions of experts in clinical practice or clinical research may also assist in determining an appropriate duration of follow-up.

7. Were the outcomes of people who withdrew described and included in the analysis?

Commonly intention to treat analysis is utilized where losses to follow-up are included in the analysis. Intention to treat (ITT) analysis may reduce bias due to changes in the characteristics between control and treatment groups that can occur if people either drop out, or if there is a significant level of mortality in one particular group. The Cochrane Systematic Review handbook identifies two related criteria for ITT analysis, although how these criteria are applied remains somewhat contentious (Higgins and Green 2008):

- Trial participants should be analyzed in the groups to which they were randomized regardless of which (or how much) treatment they actually received, and regardless of other protocol irregularities, such as ineligibility
- All participants should be included regardless of whether their outcomes were actually collected

8. Were outcomes measured in a reliable way?

Having established the objectivity of the outcome measurement instrument (see item 5 of this scale), it's important to establish how the measurement was conducted. Were those involved in collecting data trained or educated in the use of the instrument/s? If there was more than one data collector, were they similar in terms of level of education, clinical or research experience, or level of responsibility in the piece of research being appraised?

9. Was appropriate statistical analysis used?

As with any consideration of statistical analysis, consideration should be given to whether there was a more appropriate alternate statistical method that could have been used.

The methods section of cohort or case-control studies should be detailed enough for reviewers to identify which analytical technique was used (in particular, regression or stratification) and how specific confounders were measured.

For studies utilizing regression analysis, it is useful to identify if the study identified which variables were included and how they related to the outcome. If stratification was the analytical approach used, were the strata of analysis defined by the specified variables? Additionally, it is also important to assess the appropriateness of the analytical strategy in terms of the assumptions associated with the approach as differing methods of analysis are based on differing assumptions about the data and how it will respond.

The Critical Appraisal Criteria for Descriptive Studies

Descriptive/Case-series

1. Was the study based on a random or pseudo-random sample?

Recruitment is the calling or advertising strategy for gaining interest in the study, and is not the same as allocation, therefore; seemingly random methods of recruitment such as open advertising should not be considered a method of sampling. Moreover, a descriptive study commonly has a single arm; therefore allocation is not randomized between groups. Studies may report random allocation from a population, and the methods section should report how allocation was performed.

2. Were the criteria for inclusion in the sample clearly defined?

How was the sample recruited? Give consideration to whether responders have potential to differ in some significant way to non-responders. Was inclusion based on clearly defined characteristics or subjective values and opinions such as personal interest of the participants in the topic.

3. Were confounding factors identified and strategies to deal with them stated?

Any confounding factors should be identified, and the study report methods for measuring their potential impact on the study results. Confounding factors do not need to be "controlled" or eliminated from a descriptive study, the results of these studies are useful regardless, but more so if an attempt is made to measure the scope of impact.

4. Were outcomes assessed using objective criteria?

If the outcomes were assessed based on existing definitions or diagnostic criteria, then the answer to this question is likely to be yes. If the outcomes were assessed using observer reported, or self reported scales, the risk of over or under reporting is increased, and objectivity is compromised. Importantly, determine if the measurement tools used were validated instruments as this has a significant impact on outcome assessment validity.

5. If comparisons were being made, was there sufficient description of groups?

This item should focus on any reported characteristics, note that the comparator group in a descriptive study may not be in the primary study, but may be extrapolated from other sources. Regardless of the source, some attempt should have been made to identify and measure the similarities between included groups.

6. Was follow-up carried out over a sufficient time period?

The appropriate length of time for follow-up will vary with the nature and characteristics of the population of interest and/or the intervention, disease or exposure. To estimate an appropriate

duration of follow-up, read across multiple papers and take note of the range for duration of follow-up. The opinions of experts in clinical practice or clinical research may also assist in determining an appropriate duration of follow-up.

7. Were the outcomes of people who withdrew described and included in the analysis?

Commonly intention to treat analysis is utilized where losses to follow-up are included in the analysis. Intention to treat (ITT) analysis may reduce bias due to changes in the characteristics between control and treatment groups that can occur if people either drop out, or if there is a significant level of mortality in one particular group. The Cochrane Systematic Review handbook identifies two related criteria for ITT analysis, although how these criteria are applied remains somewhat contentious:

Trial participants should be analyzed in the groups to which they were randomized regardless of which (or how much) treatment they actually received, and regardless of other protocol irregularities, such as ineligibility.

All participants should be included regardless of whether their outcomes were actually collected (Higgins and Green 2008).

8. Were outcomes measured in a reliable way?

It's important to establish how the measurement was conducted. Were those involved in collecting data trained or educated in the use of the instrument/s? If there was more than one data collector, were they similar in terms of level of education, clinical or research experience, or level of responsibility in the piece of research being appraised? With descriptive studies, caution should be exercised where statistical significance is linked by authors with a causal effect, as this study design does not enable such statements to be validated.

9. Was appropriate statistical analysis used?

Broadly, two principles apply to determining if the statistical analysis was appropriate. Firstly, as with any consideration of statistical analysis, consideration should be given to whether there was a more appropriate alternate statistical method that could have been used for the study design and type of data collected. Secondly, did the authors report baseline data, or change values in addition to endpoint data. For example, reporting an endpoint as a percentage value, but no baseline values means reviewers are unable to determine the magnitude of change.

Chapter 7:
Data Extraction

Once the studies have been included in a review, the relevant results have to be abstracted from the reports. This requires definition of the comparison and outcome to be assessed, and is often a quite complex process. It is also open to subjective influences, and so the extraction of the key data from each study should involve two or more investigators working independently using a data extraction instrument. The data extracted includes:

- Source - citation and contact details
- Eligibility - confirm eligibility for review
- Methods - study design, concerns about bias
- Participants - total number, setting, diagnostic criteria
- Interventions - total number of intervention groups
- Outcomes - outcomes and time points
- Results - for each outcome of interest.

Difficulties related to the extraction of data include different populations outcome measures, interventions administered differently and the reliability of data extraction (i.e. between reviewers). Errors in data extraction can be minimized by using a data extraction form; pilot testing the extraction form prior to commencement of the review; training and assessing data extractors; having two people extract data from each study; and blinding extraction before conferring.

In quantitative studies, the numbers of participants per group are recorded and the interventions per group described. When the interventions have been recorded in detail, the primary study author's conclusions and any reviewer comments are added. The JBI analytical modules are designed to extract data from all quantitative study designs provided that the primary study has actually reported the necessary data in the form of exposures or outcomes as numbers with for treatment and control numbers (commonly recorded as n/N per group) or the sample size, means and standard deviations. The following description of data extraction applies to:

- Randomized controlled trials and pseudo-randomized controlled trials
- Comparable (controlled) cohort or case control studies, and
- Descriptive or case series studies.

Data extraction: Study Details

Method

A method usually describes the process-based approach to how the research was conducted. In a traditional review of effects, this will be "randomized controlled trial", i.e. it is the study design. It is useful to further add any particular characteristics of the design such as

Figure 2: MAStARI Initial Data Extraction

whether it used an intention to treat analysis, whether it was single, double or triple blinded, whether controls were prospective, and whether randomization was truly random, or quasi-randomization.

Setting

This term is used to describe where the research was conducted - the specific location, for example: at home; in a nursing home; in a hospital; in a dementia specific ward in a sub-acute hospital. If the setting was a hospital (which is typical for clinical trials on the effects of health care interventions), what type and level of care was provided through that facility? Were specific wards or areas utilized? Considering these questions will assist in meaningful identification of the type of setting.

Participants

Information entered in this field should be related to the inclusion and exclusion criteria of the research, and include (but not be limited to) descriptions of age, gender, number of included subjects, ethnicity, level of functionality, and cultural background. Included in this section should be definitions of terms used to group people that may be ambiguous or unclear, for example, if the paper includes role definitions. In quantitative studies the aim is to establish whether the participants were statistically similar. However, for readers of the review report, they may also wish to identify how similar or dissimilar their own patient population is from study populations. A detailed description assists both write-up of the review report and end users of the report in their decision-making. It is particularly important that the '# Participants' (i.e. the number of participants) fields are completed, this data is also used in later calculations.

Interventions

When completing the interventions fields, note that in cases where a new or modified intervention (the 'treatment') is being compared to a traditional or existing procedure (the 'control'), it is conventional to include the treatment(s) as intervention A (whose participants are in group A) and the control as intervention B (whose participants are in group B). JBI-MAStARI requires entry of outcomes prior to entry of interventions and data per group for interventions.

Authors Conclusions

These are the conclusions reached by the study author, usually found in or at the end of the discussion section.

Reviewers Conclusions

These are the conclusions reached by the Reviewer, and may address any issue related to the study with regard to its perceived quality, methods, units of analysis, outcomes, how the data was analyzed, or whether the authors conclusions were considered reliable.

Date Extraction: Study Findings

The second phase of data extraction in a quantitative review is the particular numeric values related to the a-priori stated outcomes of interest. The types of data that you will read and come across in published or unpublished studies include the following statistics that are common across the health sciences. It is important to be aware of these types of data, and how to interpret them, however, these are calculations that are based on the raw data, and it is this raw data (discussed later) that forms the basis of phase two extraction in the process of determining whether or not meta analysis will be undertaken as the method of synthesis.

Odds ratio

Odds ratio is the ratio of the odds of an event in the intervention group to the odds of an event in the control group. An odds ratio of one indicates no difference between comparison groups. For undesirable outcomes an odds ratio of less than one indicates that the intervention was effective in reducing the risk of that outcome.

Risk

The risk of an event is the probability that an event will occur within a stated time period (P). This is sometimes referred to as the absolute risk. For example:

- The risk of developing anaemia during pregnancy for a particular group of pregnant women would be the number of women who develop anaemia during pregnancy divided by the total number of pregnant women in the group.
- The risk of a further stroke occurring in the year following an initial stroke would be the number who have another stroke within a year divided by the total number of stroke patients being followed up.

Relative risk

When we use the term 'relative risk', we are referring to the ratio of risk in the intervention group to the risk in the control group. A risk ratio of one indicates no difference between comparison groups. For undesirable outcomes a risk ratio of less than one indicates that the intervention was effective in reducing the risk of that outcome.

Risk difference

Risk differences are the absolute difference in the event rate between two comparison groups. A risk difference of zero indicates no difference between comparison groups. For undesirable outcomes a risk difference that is less than zero indicates that the intervention was effective in reducing the risk of that outcome.

Number Needed to Treat

This is the additional number of people you would need to give a new treatment to in order to cure one extra person compared to the old treatment. Alternatively for a harmful exposure, the number needed to treat becomes the number needed to harm and it is the additional number of individuals who need to be exposed to the risk in order to have one extra person develop the disease, compared to the unexposed group.

Relative Risk and Odds Ratio

The odds ratio can be interpreted as a relative risk when an event is rare and the two are often quoted interchangeably. For case-control studies it is not possible to calculate the RR and thus the odds ratio is used. For cross-sectional and cohort studies both can be derived and if it is not clear which is the causal variable and which is the outcome should use the odds ratio as it is symmetrical, in that it gives the same answer if the causal and outcome variables are swapped. Odds ratios have mathematical properties that make them more often quoted for formal statistical analyses

Mean difference

Mean difference, as the name implies, is the difference between the means (i.e. the average values) of two groups.

Weighted mean difference

Weighted mean difference refers to the situation where studies have measured an outcome on the same scale and the weight given to the mean difference in each study is usually equal to the inverse of the variance.

Standardized differences in mean

Standardized differences in mean refers to the situation where studies have measured an outcome using different scales or units of measurement (for example inches and centimetres) and the mean difference may be divided by an estimate of the within group standard deviation to produce a standardized value without any units.

Figure 3: Example of numeric data extraction fields in JBI-MAStARI for dichotomous data

Data Extraction: Outcomes and Numeric Data

The second phase of data extraction is to extract the data of relevance to the specific outcomes of interest that were stated in the a-priori protocol. For outcomes that are measured dichotomously, the raw values are numbers within sample with the exposure compared to the total number of participants within that arm of the study. Figure 3 illustrates how JBI-MAStARI facilitates the collection of data. The only variation in continuous outcomes is that the mean and standard deviation are collected, the template and process is the same as for dichotomous data. The data extraction is undertaken by a single reviewer, and therefore a double data entry process is required in order to validate and provide confirmation of the reliability and accuracy of the data being entered. This is a common convention in quantitative systematic reviews where the intended end point is meta analysis.

Chapter 8:
Data synthesis

A meta-analysis is performed to calculate a more precise estimate of the outcome of interest. Meta-analysis is a quantitative method of combining the results of independent studies. This is achieved by pooling the results of various studies, in effect increasing the total sample size of the analysis, and improving the precision of the outcome estimate. Meta-analysis of RCTs aims to derive an overall estimate of effect. The principle purpose of the meta-analysis of observational studies is to investigate the reasons for differences in risk estimates between studies and to discover patterns of risk among study results (Joanna Briggs Institute 2008; Borestein, Hedges et al. 2009). During the process of combining studies, you will undoubtedly encounter the problem related to the appropriateness of pooling or combining the results of different studies. This problem arises due to heterogeneity between studies. Heterogeneity may be problematic due to methodological issues or despite similar methodology, different outcomes being measured! Meta-analyses of observational studies, more so than for clinical trials often have the added challenge of incorporating various designs and levels of quality. This issue is a problem when there is more variation between studies than would be expected based on sampling alone. Heterogeneity between studies is often more common and extreme in observational studies than clinical studies (Sutton, Abrams et al. 2000).

Tests of heterogeneity are based on the assumption that all studies in the systematic review are essentially the same, therefore, these tests effectively measure the extent to which the observed study outcomes deviate from the calculated summary outcome. Visual inspection of the meta-analysis Forest plot can be the first stage of assessing heterogeneity. Longer confidence intervals (CI) indicate less certain estimates. Statistically, a $\chi 2$ Test for Homogeneity can be used. This test calculates a P value using an individual studies weight, effect size and overall effect size. The Q-test is also efficient for determining heterogeneity. A funnel plot can be used to visually detect sources of heterogeneity such as publication and selection bias. If the plot appears asymmetrical, it may suggest a heterogeneous sample.

For a meta-analysis to be feasible all outcomes of interest must be similar and measured in the same way, that is, they must be homogeneous. The relative risk (risk ratio), risk difference, and odds ratio are common numerical measures for dichotomous (binary) or "yes-no" outcomes. The hazard ratio is similarly used to present dichotomous survival data. Continuous outcomes, like blood pressure for example will most commonly be presented with the mean difference (effect size) (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001). The effect size is commonly standardized by the pooled estimate of the within-group variance. Where continuous outcomes are skewed, you may encounter transformed data (e.g. logarithmic) or use of the median rather than the mean.

The focus of the remainder of this discussion will be on the more commonly used dichotomous outcomes. A single summary measure of study outcomes is a weighted average of all study outcomes. The weight indicates the 'influence' of the study and in a meta-analysis a study

with a large number of subjects is thus more influential than a study with a small number of subjects. The estimate of the precision of this summary measure is the CI. As a meta-analysis aims to improve the precision of the outcome measure, the CI around your summary measure should ideally shrink when compared to the individual studies included in the meta-analysis - the smaller the better!

There are various statistical methods for the combination of study outcomes, including fixed effects analysis, and random effects analysis. These can be distinguished by their methods used for estimating the CI, or precision of the overall summary outcome.

When using fixed effects analysis for dichotomous outcomes there are various methods available including Woolf's Method, Mantel-Haenszel Method and Peto's Method, which will in most instances yield similar result (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001). Each of these methods are referred to as fixed effects, as all studies are measuring the same parameter and any difference in outcome observed across studies is only due to chance – that is, it is assumed there is no variation inherent in the source population. In essence, each of the studies in the meta-analysis these methods take into account, considers within study variation rather than the between study variation, and hence these methods are not used if there is significant heterogeneity apparent. The CI of the summary measure therefore, will reflect variability between patients within the sample.

Where there is evidence of statistical heterogeneity between studies the fixed effects model will not fit the observed data well and therefore it is more appropriate to use the random effects model. Random effects are often applied to compensate for the heterogeneity apparent in observational studies. In this model, variability in data arises from variability between the patients (or within the sample) and also from the differences between the studies also. It is assumed that all studies are different, and that the outcome of a study will fluctuate around its own true value. It is assumed that each of these true values is drawn randomly from the same normal distribution within the population. The resultant summary outcome is the estimate of the mean of the normal probability distribution of sample outcomes from which our sample of outcomes was randomly drawn. The summary value from a random effects model will often have a wider CI than seen for the fixed effects model. Where there is no heterogeneity present, the results of fixed and random effects models will be similar. Sometimes, when heterogeneity is indicated, it may be an indication that it is not appropriate to proceed with meta-analysis and the results of included studies should be summarized solely as a narrative review.

Meta-analysis is useful if studies report different treatment effects or if studies are too small (insufficient power) to detect meaningful effect. It can be used if studies:

- have the same population;
- use the same intervention administered in the same way;
- measure the same outcomes; and
- studies are homogeneous (i.e. sufficiently similar to estimate an average effect).

Calculating an Average

Most commonly this involves calculating the Odds Ratio for dichotomous data (e.g. the outcome is either present or not). The particular formula (calculated for you in the software) is

Figure 4: Example of a Forest Plot

Study or Subgroup	Exercise Events	Exercise Total	Usual Care Events	Usual Care Total	Weight	Odds Ratio M–H, Fixed, 95% CI	Odds Ratio M–H, Fixed, 95% CI
Back 2008	1	21	0	16	1.3%	2.41 [0.09, 63.25]	
DeBusk 1994	25	293	33	292	75.4%	0.73 [0.42, 1.27]	
Haskell 1994	9	145	3	155	6.8%	3.35 [0.89, 12.64]	
Kovoor 2006	5	72	4	70	9.4%	1.23 [0.32, 4.79]	
Schuler 1992	2	56	3	57	7.1%	0.67 [0.11, 4.15]	
Total (95% CI)		587		590	100.0%	0.97 [0.63, 1.51]	
Total events	42		43				

Heterogeneity: Chi^2 = 4.96, df = 4 (P = 0.29); I^2 = 19%
Test for overall effect: Z = 0.12 (P = 0.91)

0.01   0.1   1   10   100
Favours experimental   Favours control

51/49 = 1.04. For dichotomous data, the point at where there is no statistically significant difference between groups is set at 1; (no difference between groups = 1). Therefore, an Odds Ratio that does not include one will show a significant difference either in favour of the treatment or control intervention.

For continuous data (e.g. Body weight or temperature), the mean difference, or weighted mean difference is calculated. For continuous data, the point at where there is no statistically significant difference between groups is set at zero (0); (no difference between groups = 0). Therefore, a mean difference that does not include zero will show a significant difference either in favour of the treatment or control intervention.

A confidence interval is reported for both dichotomous and continuous data. The confidence interval is the set of values within which it is accepted that the real result lies (for a given degree of certainty such as 90%, 95% or 99%). Confidence intervals are an indication of how precise the findings are. Sample size greatly impacts the CI, e.g. the larger the sample size the smaller the CI, the greater the power and confidence of the estimate. When calculated for Odds Ratio, the CI provides the upper and lower limit of the odds that a treatment may or may not work. If the odds ratio is 1, odds are even and therefore, not significantly different (recall the odds of having a boy).

The results of a meta- analysis are presented in a forest plot – often referred to as a meta-view graph (Figure 4).

Odds ratio (for categorical data) and standard or weighted mean differences (for continuous data) and their 95% confidence intervals are calculated in the meta-view graph.

Heterogeneity

Heterogeneity is assessed using the standard Chi-square. When used in relation to meta-analysis, the term 'heterogeneity' refers to the amount of variation in the characteristics of included studies. For example, if three studies are to be included in a meta-analysis, do each of the included studies have similar sample demographics, and assess the same intervention? (Note that the method by which the intervention is measured does not need to be identical.) While some variation between studies will always occur due to chance alone, heterogeneity is said to occur if there are significant differences between studies, and under these circumstances meta-analysis is not valid and should not be undertaken. But how does one tell whether or not differences are significant? Visual inspection of the meta-analysis is

the first stage of assessing heterogeneity. JBI-MAStARI plots the results of individual studies and thus indicates the magnitude of any effect between the treatment and control groups. Do the individual studies show a similar direction and magnitude of effect – i.e. are the rectangular symbols at similar positions on the X-axis? A formal statistical test of the similarity of studies is provided by the test of homogeneity. This test calculates a probability (P value) from a Chi-square statistic calculated using estimates of the individual study's weight, effect size and the overall effect size. However, note that this test suffers from a lack of power – and will often fail to detect a significant difference when a difference actually exists – especially when there are relatively few studies included in the meta-analysis. Because of this low power, some review authors use a significance level of $P < 0.01$, rather than the conventional 0.05 value, in order to protect against the possibility of falsely stating that there is no heterogeneity present (Joanna Briggs Institute 2008).

In meta-analysis, the results of similar, individual studies are combined to determine the overall effect of a particular form of health care intervention (the treatment) compared to another standard or control intervention for a specified patient population and outcome. In meta-analysis, the effect size and weight of each study are calculated. The effect size indicates the direction and magnitude of the results of a particular study (i.e. do the results favour the treatment or control, and if so, by how much), while the weight is indicative of how much information a study provides to the overall analysis when all studies are combined together.

Fixed Effects and Random Effects

Meta-analysis can be based on either of two assumptions. In a fixed effects model, it is assumed that any differences between treatment and control are the same (or fixed) in each study. Thus any observed differences among the studies' results are due solely to chance and there is no heterogeneity between the studies. However, when there is heterogeneity apparent (for example, the test of homogeneity is significant), the validity of the assumption of a fixed effect is questionable, and thus another approach is to consider that the treatment effects for the individual studies are not identical and, in fact, follow a distribution related to an overall average treatment effect. That is, the effect size is random, and is assumed to follow a Normal distribution and consequently has a mean and variance.

Essentially, the test for homogeneity is asking the statistical question "is the variance around the estimate of the effect size zero or non zero?" If the variance around the estimate of the effect size is zero, then there is no heterogeneity present, and the results of the fixed and random effects models will be similar.

There is no consensus about whether fixed or random effects models should be used in meta-analysis. In many cases when heterogeneity is absent, the two methods will give similar overall results. When heterogeneity is present, the random effects estimate provides a more conservative estimate of the overall effect size, and is less likely to detect significant differences. For this reason, random effects models are sometimes employed when heterogeneity is not severe; however, the random effects model does not actually analyze the heterogeneity away and should not be considered as a substitute for a thorough investigation into the reasons for the heterogeneity. Additionally, random effects models give relatively more weight to the

results of smaller studies – this may not be desirable because smaller studies are typically more prone to bias and of lower quality than larger studies.

Meta-analytical techniques available in MAStARI

There are a number of meta-analytical techniques available. The selection of a particular technique is governed by three things: the study type, nature of the data extracted and assumptions underlying the meta-analysis. Here, we introduce the tests that are available in JBI-MAStARI and discuss when it is appropriate to use each of the tests.

When the outcomes of included studies are dichotomous, JBI-MAStARI can be used to generate two overall effect sizes: odds ratios (OR) and relative risks (also known as risk ratios, RR). The choice of whether OR or RR are calculated is important and should be carefully considered with due reference to three criteria.

Dichotomous data – methods of meta-analysis

There are several different methods available to pool results of dichotomous data, depending on the data type and whether a random or fixed effects model is required: Mantel-Haenszel, Peto's; and DerSimonian and Laird.

Mantel-Haenszel

Mantel-Haenszel is the default meta-analytical method for dichotomous data using a fixed effects model. Both OR and RR can be pooled using Mantel-Haenszel methods; the calculation of study weights and effect sizes, and overall effect sizes differs slightly between OR and RR. The Mantel- Haenszel method is generally preferred in meta-analysis to another method (inverse variance) because it has been shown to be more robust when data are sparse (in terms of event rates being low and/or the number of trials being small).

Peto's odds ratio

Peto's odds ratio is an alternative method for meta-analysis of OR using a fixed effects method. It employs an approximation that can be inaccurate if treatment affects are very large, and when the sample sizes between treatment and control groups are unbalanced. However, the method is appropriate when event rates are very low and effect sizes are not overly large.

DerSimonian and Laird

DerSimonian and Laird methods are used in the meta-analysis of OR and RR using a random effects model. Although the study effect sizes and heterogeneity statistics are calculated as for the fixed effects model, the study weights and overall effect sizes in DerSimonian and Laird random effects models are calculated slightly differently to fixed models.

Meta-analysis of continuous data

When the outcomes of included studies are continuous, JBI-MAStARI can be used to generate two overall effect size calculations using weighted mean differences (WMD) or standardized mean differences (SMD).

The definition of a continuous outcome is that it may be measured on a scale that is continuously variable, e.g. for any two valid continuous measurements there is always one in between. This includes outcomes that are:

- Numerical
- Made up of many ordered categories

Methods for meta-analysis of continuous data assume that the data have a Normal distribution, and revolve around means and standard deviations. A mean is the 'average' (i.e. sum of the observations divided by the number of observations). The standard deviation is a measure of how variable the observations are around the mean. A small standard deviation indicates that the observations are all near the mean; a large standard deviation indicates that the observations vary substantially.

The meta-analysis of differences between means from different trials relies on the outcome being measured in the same units in every trial: we can't combine a difference in mean weight loss in kilograms with a difference in mean weight loss in pounds, you should directly convert all the data to the same units.

However, we can't combine two different psychometric scales even if they both measure depression as the multiplication factor is not known. A way around this is to compare standardized mean differences, rather than actual means. The standardized mean difference is the difference in means divided by a standard deviation. This standard deviation is the pooled standard deviation of participants' outcomes across the whole trial. Note that it is not the standard error of the difference in means (a common confusion).

Weighted mean difference

The WMD measures the difference in means of each study when all outcome measurements are made using the same scale. It then calculates an overall difference in mean for all studies (this is equivalent to the effect size) based on a weighted average of all studies, which is, in turn related to the SD. JBI-MAStARI uses the inverse variance method of calculating WMD for fixed effects models and the DerSimonian and Laird method for random effects models.

Alternatively, different studies may measure the same outcome using different scales. For example, pain can be measured on a range of different scales including non-verbal scales (e.g. visual analogue scale) and verbal scales (e.g. 5 point categorical scale). These studies can be combined in a meta-analysis that incorporates SMD. If the measurement scales operate in the same direction (e.g. An increase in pain is measured as an increase in on both scales), then using SMD is straightforward.

However, if two measurement scales operate in a different direction – for example a score of 10 is the worst pain imaginable on one scale but a score of 1 is the worst pain imaginable on another scale – then data from one scale need to be reversed. This is relatively simply achieved by multiplying the mean data from one scale (for both treatment and control groups) by -1. Standard deviations do not need to be modified.

Standard mean difference

JBI-MAStARI provides two options for calculation of the SMD using fixed effects: Cohen's SMD and Hedges' SMD. Both options produce a similar result, although Hedges' SMD is

Figure 5: MAStARI comparison screen

generally preferred as it includes an adjustment to correct for small sample size bias (Deeks, Higgins et al. 2001; Deeks, Higgins et al. 2001). As per WMD, the DerSimonian and Laird method is used for random effects models calculations for SMD.

Executing a meta analysis

The meta-analysis module is made up of a number of drop down menus that allow the user to specify the comparison required (i.e. which case group (group with outcome of event) is to be compared group without outcome of event), the outcome to be included and the statistical tests to be used (Figure 5).

The user must specify the correct data type (continuous/dichotomous), the required effects model to be used (random/fixed), the statistical method of meta-analysis required and the size of confidence limits to be included in the calculations. The method to be used will depend on the data type.

As discussed previously, where there is variability between studies, it is referred to as hetero-geneity. This classification can be broken down in to a series of sub classifications that reflect accepted types of heterogeneity:
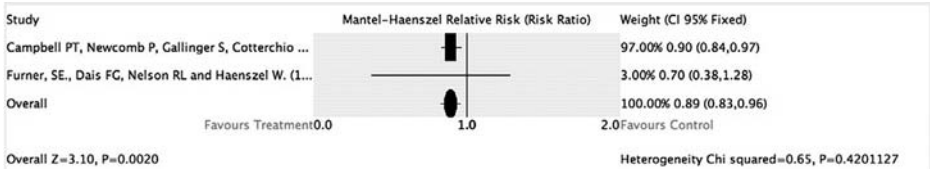
- Clinical heterogeneity: variability in the participants, interventions or outcomes
- Methodological heterogeneity variability in study design and risk of bias
- Statistical heterogeneity in the intervention effects being evaluated in the studies

Identifying and measuring heterogeneity

A number of options are available if (statistical) heterogeneity is identified among a group of studies that are otherwise considered appropriate for a meta-analysis.

- Undertake a confirmatory check of the data to ensure they are accurate,
- Do not do a meta-analysis, consider narrative summary as an appropriate form of synthesis,
- Explore heterogeneity through sub group analysis or regression analysis – this helps to understand, but does not resolve heterogeneity,

Figure 6: Example of a meta view graph from JBI MAStARI.

Study	Mantel–Haenszel Relative Risk (Risk Ratio)	Weight (CI 95% Fixed)
Campbell PT, Newcomb P, Gallinger S, Cotterchio ...		97.00% 0.90 (0.84,0.97)
Furner, SE., Dais FC, Nelson RL and Haenszel W. (1...		3.00% 0.70 (0.38,1.28)
Overall		100.00% 0.89 (0.83,0.96)
	Favours Treatment 0.0 1.0 2.0 Favours Control	
Overall Z=3.10, P=0.0020		Heterogeneity Chi squared=0.65, P=0.4201127

- Ignore the presence of heterogeneity – apply a fixed effects model, although note that the conclusions of a meta analysis in the presence of statistical heterogeneity does not resolve the heterogeneity, and conclusions from such an analysis are likely to be inaccurate,
- Perform a random-effects meta-analysis – random effects models incorporate heterogeneity in their statistical assumptions,
- Change the effect measure e.g. mean difference to standardized mean difference, or vary from odds ratio, relative risk or risk difference.
- Exclude studies from the analysis – this increases the risk of "fishing" for findings, and given it can not be stated with any certainty in an a-priori protocol, such an approach would be considered a significant compromise of the integrity of the review results.

Forest plot of dichotomous data

As in this example, we are examining risk of having bowel cancer (outcome) in relation to exposure to oestrogen (equivalent to intervention), the combined result of the analysis showed combined relative risk of having bowel cancer in oestrogen users versus nonusers.

Saving forest plot/MAStARI- views

The forest plot can be saved to your computer in jpeg (jpg) format using the save graph to disk button of the page. Carefully note the file name and the contents of the meta analysis. The forest plot can then be pasted directly into your systematic review in MS word, or attached to your submission for publication.

Summarising the Findings in Narrative Form

Where meta-analysis is not appropriate, a narrative summary should draw upon the data extraction, with an emphasis on a textual summary of study characteristics as well as data relevant to the specified outcomes.

Reporting findings

There is no standardized international approach to structuring how the findings of reviews will be reported. The audience for the review should be considered when structuring and writing the findings up. Meta-view graphs represent a specific item of analysis that can be incorporated in to the results section of a review (Figure 6). However, the results are more than the meta-view graphs, and whether it is structured based on the intervention of interest, or

Figure 7: Example of a flow chart with studies identified and the audit trail of decision points leading to final numbers of included and excluded studies
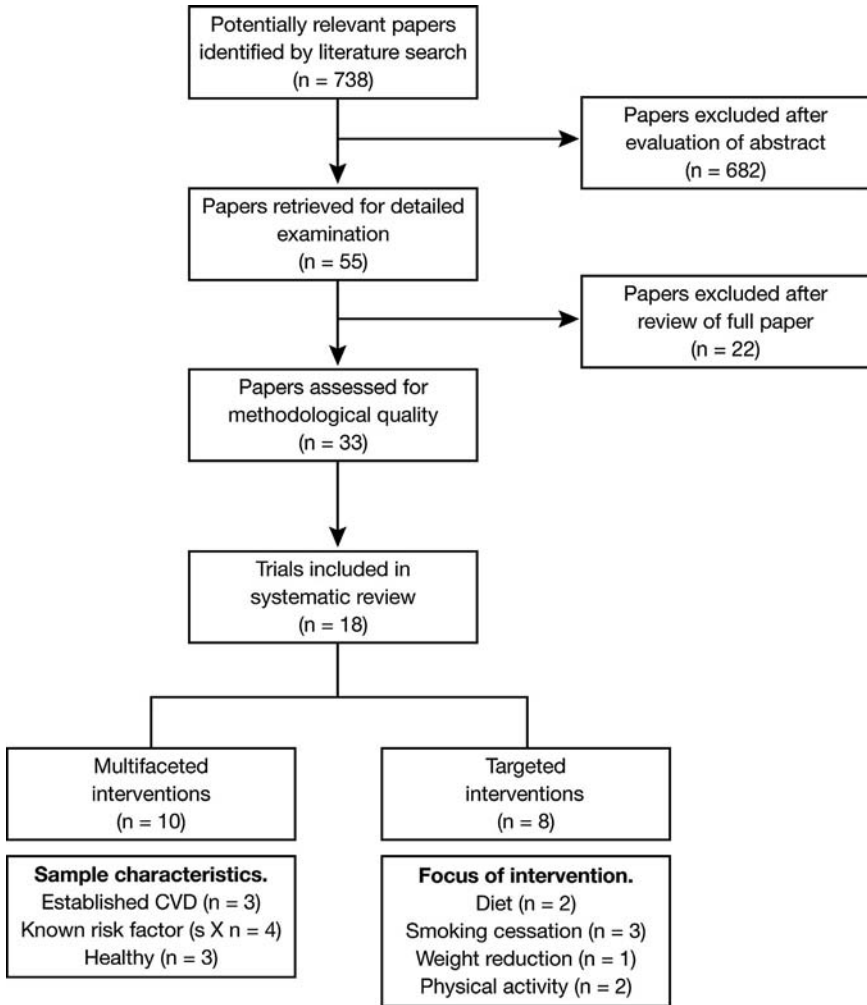
some other structure, the content of this section needs to present the results with clarity using the available tools (meta-view graphs, tables, figures) supported by textual descriptions.

The results section should be framed in such a way that as a minimum, the following fields are described in the protocol as either planned for reporting, or given consideration by the reviewers in preparing their systematic review report as per the following example:

- Numbers of studies identified,
- Numbers of retrieved studies,
- Numbers of studies matching preferred study design (i.e. RCTs),

- Numbers and designs of other types of studies,
- Numbers of appraised studies,
- Numbers of excluded studies and overview of reasons for exclusion,
- Numbers of included studies.

These results are commonly written in narrative style, and also illustrated with a flow diagram (Figure 7).

Chapter 9:
Developing a systematic review report

A systematic review report is the final outcome of a review. To a large extent, the components of the systematic review report will mirror the content of the original protocol. As with the protocol, there should be a comprehensive background that justifies conducting the review, a description of the objectives of the review, an account of the criteria that were used for considering studies for the review, the search strategy used and methods utilized for critical appraisal, extraction and synthesis of data (Joanna Briggs Institute 2008).

The review of results includes both a flow chart describing the search, selection and inclusion process and a description of the studies that includes the type and number of papers identified. The number of papers that were included and excluded should be stated. There should also be a summary of the overall quality of the literature identified (Joanna Briggs Institute 2008). The results section must be organized in a meaningful way based on the objectives of the review and the criteria for considering studies. Particular consideration should be given to the types of interventions and outcomes.

The discussion should include an overview of the results and it should address issues arising from the conduct of the review including limitations and issues arising from the results of the review. Conclusions should center on the implications for practice and for research. These should be detailed and must be based on the documented results, not author opinion. Where evidence is of a sufficient level, appropriate recommendations should also be made. Recommendations must be clear, concise and unambiguous.

Assigning levels of evidence to recommendations

The recommendations drawn from the results of aggregative reviews are each assigned a level of evidence based on the nature of the research used to inform the development of the recommendation.

As Table 2 shows, levels of evidence derived from aggregative reviews relate to the credibility of the findings that lead to a recommendation. Recommendations based on evidence where all of the findings it is derived from are "unequivocal are ranked as "Level 1" evidence; and where the findings are all at least "credible", as "Level 2" evidence. Reviewers are expected to, when drafting recommendations for practice, include a level of evidence congruent with the research design that led to the recommendation.

Appendices

Again, as in the initial protocol, the final review report should include references and appendices. The references should be appropriate in content and volume and include background references and studies from the initial search. The appendices should include:

- Critical appraisal form(s)
- Data extraction form(s)

- Table of included studies
- Table of excluded studies with justification for exclusion

These checklists should reflect the types of studies, settings, participants, interventions, and outcomes for the review question posed. If systematic review reports are of a high enough standard they may be utilized as evidence upon which to base clinical practice guidelines.

Conflict of interest

As per the section on protocol development, a detailed, explicit statement is necessary and should be targeted to independently established standards.

A statement should be included in every review protocol that either declares the absence of any conflict of interest, or describes a specified or potential conflict of interest. Conflict of interest statements should adhere to the guidelines of the International Committee of Medical Journal Editors (ICMJE) for individual authors and project support (http://www.icmje.org/ethical_4conflicts.html). Additionally, the Committee on Publication Ethics (COPE) have extensive guidelines for conflict of interest statements that are intended to protect the authors as well as the readers, and review authors should ensure they are familiar with and adhere to the principals described within the COPE framework (http://www.publicationethics.org/).

Implications for practice

Implications for practice should be detailed and based on the documented results, not reviewer opinion. In qualitative reviews, recommendations are declamatory statements that are steeped in context. Therefore, generalizability occurs between cases rather than across broad populations. Recommendations must be clear, concise and unambiguous.

Implications for research

All implications for research must be derived from the results of the review, based on identified gaps in the literature or on areas of weakness, such as methodological weaknesses. Implications for research should avoid generalized statements calling for further research, but should be linked to specific issues.

Discussion and Conclusions

The stated purpose of this book is to provide an introductory guide to the reliable, transparent and rigorous conduct of systematic reviews of questions related to the effects of health care interventions. The structure has therefore been to pragmatically align with the process for conducting a systematic review.

The first section outlines some understandings of the origins of knowledge that is classified as positivist empirical knowledge (although all research is of itself empirical by nature). The intent of this chapter was to frame our current understandings around the notion of objectivity, and where they come from. It is rare for any book situated in the positivist perspective, particularly a research text to provide any perspective on the nature of knowledge within this paradigm. However, given the statistical processes within meta analysis are based on probabilities that

rely upon these notions, it seemed appropriate to include some information on this basis of positivist knowledge. For readers who are immersed in the positivist perspective, it may serve to highlight that there are some criticisms of this field, and to become somewhat familiar with those lines of argument. From this chapter onwards, the book is very much focused on the practicalities and nuances of preparing for, then undertaking a systematic review of the research evidence.

Therefore, the second section begins with a protocol, with an emphasis on developing a rigorous, detailed, question that is based upon the population, intervention, comparator, outcome mnemonic. The inclusion criteria in this section are demonstrably aligned to the review question and provide the detailed analysis of the PICO question structure within the protocol. This chapter continues with a clear breakdown of the steps and stages involved in developing a protocol, as it is this rigorous, detailed, analytic a-priori protocol that in a substantive way, defines the rigor of the subsequent review and gives systematic reviews there international credibility.

Section three extends on the core concepts introduced in section two with particular emphasis on searching, and strategies for identifying the literature relevant to your question. The section on searching is not an exhaustive resource, but does provide appropriate information to give the context and detailed consideration on how to apply principals of searching. As this is a central process to the validity of the systematic review, we strongly suggest that this section not be used as a definitive resource for searching, but that it be used as an information resource to assist in equipping reviewers with entry level knowledge. Those unfamiliar with searching for evidence using explicit protocol based approaches as is necessary in a systematic review are strongly encouraged to seek the advice and support of an information scientist or librarian who specializes in systematic reviews. The remainder of section three deals with appraisal of study designs relevant to reviews of effects, and data extraction.

This section also includes an extensive discussion and description of the parameters for meta analysis, what the process is, what the necessary considerations are related to ensuring that the meta analysis is conducted appropriately and with transparency. Importantly, it also discusses what the limitations of meta analysis are, and how to interpret and use the results of a meta analysis even where the results are potentially confounded by statistical heterogeneity. These are essential considerations, the new reviewer needs to be aware that a meta view graph is not the end of the task, that the graph needs to be interpreted appropriately, and further investigation may be necessary.

As this is a text for systematic review, this section finishes not with the discussion on methods for meta analysis, but with more pragmatic directions and recommendations for how to write up a completed review report. The guidance in this section is based on the processes developed and used across the Joanna Briggs Institute and its international collaboration. These are highly standardized methods that are associated with high quality syntheses for the new reviewer, the student, or those looking for introductory information on methods. There are numerous other texts on this topic, however, if you are considering this text as a resource or methods guide, it is worth knowing that the content is reflective of methods that have been adopted internationally, and are supported by other resources that complement the information, methods and materials described in this text.

References

Altman, D., K. Schulz, et al. (2001). "The Revized CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration." *Annals of Internal Medicine* 134(8): 663-694.

Aslam, R., Ed. (2000). *Research and evidence in Midwifery*. Linking Research and Practice in Midwifery. Edinburgh, Baillière Tindall.

Babbie, E. (1992). *The Practice of Social Research*. Belmont, Wadsworth Publishing Company.

Bastin, H. (2004). "Personal Views: learning from evidence based mistakes." *British Medical Journal* 329: 7473.

Borestein, M., L. Hedges, et al. (2009). *Computing effect sizes for meta-analysis*. Chichester, John Wiley & Sons Ltd.

Casebeer, A. and M. Verhoef (1997). "Combining qualitative and quantitative methods: considering the possibilities for enhancing the study of chronic diseases." *Chronic Diseases in Canada* 183.

Cluett, E. "Evidence Based Practice." 24.

Cooper, H., L. Hedges, et al., Eds. (2009). *The handbook of research synthesis and meta-analysis*. New York, Russell Sage Foundation.

Crotty, M. (1998). *The Foundations of Social Research: Meaning and Perspective in the research process*. Singapore, Allen&Unwin.

Deeks, J., J. Higgins, et al., Eds. (2001). *Effect measures for meta-analysis of trials with binary outcomes.* Systematic Reviews in Health Care: Meta analysis in context. London, BMJ Publication Group.

Deeks, J., J. Higgins, et al. (2001). Statistical Methods for examining heterogeneity and combining results from several studies in meta-analysis. *Systematic reviews in health care: meta analysis in context*. G. D.-S. M. Egger, and D. Altman. London, BMJ Publication Group.

Deeks, J., J. Higgins, et al., Eds. (2006). *Analysing and presenting results*. Cochrane handbook for systematic reviews of interventions. Chickchester, John Wiley & Sons.

Dixon-Woods, M., A. Booth, et al. (1997). "Synthesising qualitative research: a review of published reports." *Qualitative Research* 7: 375.

Fay, B. (1975). *Social Theory and Political Practice*. London, Allen & Unwin.

Gray, M. (2004). "Evidence based policy making." *British Medical Journal* 329(7473): 988.

Greenhalgh, T. and R. Taylor (1995). "How to read a paper: papers that go beyond numbers (qualitative research)." *British Medical Journal* 315: 7110.

Gutatt, G. (1995). "A method of grading health care recommendations." *Journal of the American Medical Association* 274: 1800-1804.

Hardy, R. and S. Thompson (1998). "Detecting and describing heterogeneity in meta-analysi." *Statistics in Medicine* 17(8): 841-856.

Hedges, L., Ed. (1994). *Statistical Considerations*. The Handbook of Research Synthesis. New York, Russell Sage Foundation.

Higgins, J. (2002). "Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and prac." *Journal of Health Services and Reserch Policy* 7(1): 51-61.

Higgins, J. and S. Green, Eds. (2008). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1*

Joanna Briggs Institute (2007). *Sumari User Guide Version 4.0*. Adelaide, Joanna Briggs Institute.

Joanna Briggs Institute (2008). *Joanna Briggs Institute Reviewers Manual*. Adelaide, Joanna Briggs Institute.

Kao, L., J. Tyson, et al. (2008). "Clinical research methodology I: Introduction to randomized trials." *Journal of the American College of Surgeons* 2: 361-369.

Kjaergard, L., J. Vilumsen, et al. (1999). Quality of randomized clinical trials affects estimates of intervention efficacy,. *7th Cochrane Colloquium*. Italy, Rome.

Kotaska, A. (2004). "Inappropriate use of randomized trials to evaluate complex phenomena: case study of vaginal breech delivery." *BMC Medical Research* 329(7473): 1039-1042.

Krainovich-Miller, B., J. Haber, et al. (2009). "Evidence-based practice challenge: teaching critical appraisal of systematic reviews and clinical practice guidelines to graduate students." *Journal of Nursing Education* 48(4): 186-195.

Mamdani, M. (2005). "Readers guide to critical appraisal of cohort studies: 2. Assessing potential for confounding." *British Medical Journal* 330: 960-962.

Manchikanti, E., J. Hirsch, et al. (2008). "Evidence-based Medicine, systematic reviews and guidelines in interventional pain management: Part 2:

Randomized Controlled Trials." *Pain Physician* 11: 717-773.

Mays, N. and C. Pope, Eds. (1996). *Qualitative Research in health care*. London, BM Publishing Group.

Moher, D. (2003). "The inclusion of reports of randomized controlled trials published in languages other than english." *Journal of the American Medical Association* 268: 2420-2425.

Moher, D., K. Schulz, et al. (2001). "The CONSORT statement: Revized recommendations for improving the quality of reports of parallel-group randomized trials." *Journal of the American Medical Association* 285: 1987-1991.

Munhall, P. and B. Oiler (1993). Nursing Research. A Qualitative Perspective. New York, National League for Nursing.

Murphy, B. (1975). *Social Theory and Political Practice*. London, Allen & Unwin.

Normand, S.-L. (2005). "Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding." *British Medical Journal* 330: 2101-2103.

Pearson, A., J. Field, et al. (2007). *Evidence Based Practice in Nursing and Health Care: Assimilating research, experience and expertise*. Oxford, Blackwell Publishing.

Pearson, A. and Z. Jordan (2007). The Joanna Briggs Institute Approach to "Evidence. Adelaide.

Rochon, P. (2005). "Readers guide to critical appraisal of cohort studies: 1. Role and design." *British Medical Journal* 330: 895-897.

Sutton, A., K. Abrams, et al. (2000). *Methods for Meta-Analysis in Medical Research.* New York, Wiley.

Triccoa, A., J. Tetzlaffa, et al. (2008). "Few systematic reviews exist documenting the extent of bias: a systematic review." *Journal of Clinical Epidemiology* 61: 422-434.

Trohler, U. (2000). To improve the evidence of medicine': The 18th century British Origins of a Critical Approach. Edinburgh, Royal College of Physicians.