

Chapter 3: Systematic reviews of effectiveness

Catalin Tufanaru, Zachary Munn, Edoardo Aromataris, Jared Campbell, Lisa Hopp

How to cite:

Tufanaru C, Munn Z, Aromataris E, Campbell J, Hopp L. Chapter 3: Systematic reviews of effectiveness. In: Aromataris E, Munn Z (Editors). *JBIM Manual for Evidence Synthesis*. JBI, 2020. Available from <https://synthesismanual.jbi.global>. <https://doi.org/10.46658/JBIMES-20-04>

Chapter 3: Contents

- 3.1 Introduction to quantitative evidence and evidence-based practice
- 3.2 Development of a protocol for a systematic review of effectiveness evidence
 - 3.2.1 Title of the systematic review protocol
 - 3.2.2 Review question(s)
 - 3.2.3 Introduction
 - 3.2.4 Inclusion criteria
 - 3.2.4.1 Population (types of participants)
 - 3.2.4.2 Intervention (types of interventions)
 - 3.2.4.3 Comparison (types of comparators)
 - 3.2.4.4 Outcomes
 - 3.2.4.5 Types of studies
 - 3.2.5 Search strategy
 - 3.2.6 Selection of studies
 - 3.2.7 Critical appraisal
 - 3.2.8 Data extraction
 - 3.2.9 Data synthesis
- 3.3 Meta-analysis
 - 3.3.1 Objectives of meta-analysis
 - 3.3.2 Statistical models for meta-analysis
 - 3.3.3 Effect sizes
 - 3.3.4 Considerations for the meta-analysis of dichotomous data
 - 3.3.5 Considerations for the meta-analysis of continuous data
 - 3.3.6 Meta-analysis: Statistical Methods
 - 3.3.7 Subgroups in meta-analysis
 - 3.3.8 Sensitivity analysis in meta-analysis
 - 3.3.9 Meta-regression
 - 3.3.10 Heterogeneity
 - 3.3.10.1 Standard chi-squared test (Cochran test)
 - 3.3.10.2 Quantification of the statistical heterogeneity: I squared
 - 3.3.10.3 Tau-squared for random effects model meta-analysis
 - 3.3.11 Publication bias
- 3.4 Systematic review of effectiveness
 - 3.4.1 Title
 - 3.4.2 Abstract
 - 3.4.3 GRADE 'Summary of Findings' table
 - 3.4.4 Introduction
 - 3.4.5 Review question(s)
 - 3.4.6 Inclusion criteria
 - 3.4.7 Methods
 - 3.4.7.1 Search strategy
 - 3.4.7.2 Study screening and selection
 - 3.4.7.3 Critical appraisal
 - 3.4.7.4 Data extraction
 - 3.4.7.5 Data synthesis
 - 3.4.8 Results
 - 3.4.8.1 Study inclusion
 - 3.4.8.2 Methodological quality
 - 3.4.8.3 Characteristics of included studies
 - 3.4.8.4 Results and meta-analysis
 - 3.4.9 Discussion
 - 3.4.10 Conclusions and recommendations
 - 3.4.11 Conflicts and acknowledgements
 - 3.4.12 Review Appendices
- 3.5 Chapter References
- Appendix 3.1: JBI Critical appraisal checklist for randomized controlled trials
- Appendix 3.2: Discussion of JBI appraisal criteria for randomized controlled trials
- Appendix 3.3: JBI Critical appraisal Checklist for Quasi-Experimental Studies (non-randomized experimental studies)
- Appendix 3.4: Discussion of JBI appraisal criteria for Quasi-Experimental Studies (non-randomized experimental studies)

3.1 Introduction to quantitative evidence and evidence-based practice

Quantitative evidence is generated by research based on traditional scientific methods that generate numerical data. The methods associated with quantitative research in healthcare have developed out of the study of natural and social sciences. It was suggested that quantitative evidence in medicine originated in eighteenth century Britain, when surgeons and physicians started using statistical methods to assess the effectiveness of therapies for scurvy, dropsy, fevers, palsies, syphilis, and different methods of amputation and lithotomy (Trohler 2000). Since these beginnings, quantitative research has expanded to encompass aspects other than effectiveness, such as incidence, prevalence, etiology of disease, psychometric properties, and measurement of physical characteristics, quality of life, and satisfaction with care.

JBI quantitative reviews focusing on evidence of effectiveness examine the extent to which an intervention, when used appropriately, achieves the intended effect. Evidence about the effects of interventions may come from three main categories of studies: experimental studies, quasi-experimental studies and observational studies. Ideally, evidence about the effectiveness of interventions should come from good quality randomized controlled trials (RCTs) that explore final clinical end points (or patient important outcomes) such as morbidity, mortality, and quality of life (not surrogate end points which may include laboratory tests for example) (Brignardello-Petersen et al 2015). Good empirical evidence exists to indicate that RCTs that explored final clinical end points frequently contradicted (refuted) clinical studies that explored surrogate end points and also the results of observational studies (Brignardello-Petersen et al 2015). Some authors have claimed that results from RCTs and observational studies provide consistent results. Thus, the issue of the agreement of the results from RCTs and observational studies remains controversial (Brignardello-Petersen et al 2015).

Although high quality RCTs exploring final clinical end points are considered the "reference standard" (Brignardello-Petersen et al 2015), reviewers should be aware that results from any single RCT cannot be considered as "final" because results from new RCTs may contradict results from previous RCTs (Brignardello-Petersen et al 2015).

Reviewers should be aware that there is no unique universally accepted terminology for the quantitative study designs. Also, there is no unique comprehensive set of descriptions for the different study designs considered here.

Experimental studies meet three conditions: manipulation, control and random assignment. Specifically, the researchers manipulate the intervention of interest and the control condition and they randomly allocate the participants to the intervention or control group (Shadish et al 2002). Random allocation refers to an authentically random process such as the toss of a coin or use of a table of random numbers (Shadish et al 2002). Randomized controlled trials with different designs (parallel design, cross-over design, cluster design) are examples of experimental studies. There are also existing experimental studies (the intervention of interest and the control condition are manipulated by the researchers) where the allocation may not use an authentically random process. For example, if investigators use alternate group allocation like even and odd dates, they cannot ensure that each participant has an equal chance of landing in either group. Experimental studies without authentic random allocation but using systematic alternate group allocation methods mentioned above are experimental studies with pseudo-randomization, or pseudo-RCTs. Quasi-experimental studies are studies where the intervention of interest and the control condition are controlled (manipulated) by the researchers, however, the allocation of participants is not a random, systematic or pseudo-random allocation (Shadish et al 2002). Frequently, participants self-select into groups or the researchers decide which persons should get the intervention and which persons should get the control (Shadish et al 2002).

Observational studies are studies where the intervention of interest and the control condition are not controlled (manipulated) by the researchers and where researchers only observe the presence or absence of the intervention of interest and of the outcome of interest. There are diverse types of observational studies, which can be broadly categorized into analytical observational studies (cohort studies, case-control studies, and analytical cross-sectional studies) and descriptive observational studies (case reports and case series). In a cohort study, investigators select participants based on presence or absence of exposure to an intervention of interest and compare prospectively for the occurrence of the outcome of interest. In a case-control study, researchers select "case" participants or those with the outcome of interest and "control" participants, without the outcome of interest, to compare groups for past exposure or absence of exposure to the intervention. In an analytical cross-sectional study, investigators select participants without reference to the intervention or the presence of the outcome of interest. They then simultaneously examine the groups for the presence or absence of exposure to the intervention of interest and the presence or absence of the outcome of interest. In case reports and case series researchers simply describe the characteristics of participants and the outcomes of interventions.

3.2 Development of a protocol for a systematic review of effectiveness evidence

An *a priori* systematic review protocol is important because it pre-defines the objectives and methods of the systematic review. A review protocol provides the plan or proposal for the systematic review. Any deviations from the review protocol should be discussed in the systematic review report.

The review protocol describes:

- the context and rationale for the review, including what is already known and uncertainties,
- the study selection criteria (inclusion/exclusion criteria),
- the outcome measures, interventions, and comparisons considered,
- the proposed search strategy for identifying relevant studies,
- the procedures for study selection,
- the critical appraisal process and instruments,
- the data extraction process and instruments,
- the process for resolving disagreement between reviewers in study selection, data extraction, and critical appraisal decisions, and
- the proposed approaches to synthesis

3.2.1 Title of the systematic review protocol

A clear, descriptive title is important to allow readers and users to readily identify the scope and relevance of the review. 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. The protocol title should accurately describe and reflect the content of the review protocol and include relevant information with regards the types of participants, types of interventions and comparators and the outcomes considered in the review. The title should be concise and should not be phrased as a question. The title of the review protocol should explicitly identify the publication as a protocol for a systematic review. The following convention is recommended: 'a protocol for a systematic review'. Following the guidance mentioned, for systematic reviews of effectiveness we recommend the following convention: *'The effectiveness of [intervention] compared to [comparator] on [outcome]: a protocol for a systematic review'*.

3.2.2 Review question(s)

The review protocol should provide an explicit and clear statement of the review questions addressed in the review. The review questions should specify the focus of the review (effectiveness), the types of participants, types of interventions and comparators, and the types of outcomes considered. Usually, reviewers use the PICO mnemonic (population, intervention, comparator and outcome) to construct a clear and meaningful review objective/question regarding the quantitative evidence on effectiveness of interventions.

Examples of review questions: *'In community dwelling patients with stable, moderate-to-severe chronic obstructive pulmonary disease'*

1. *What is the effect of inspiratory muscle training versus no specific training on dyspnea and functional ability?*
2. *What is the effect of inspiratory muscle training versus no specific training on inspiratory muscle strength and endurance?*
3. *What is the effect of inspiratory muscle training on hypoxemia and discomfort?*

There should be consistency between the review title and the review questions in terms of the focus of the review. Review authors are encouraged to read the article by Stern et al (2014) regarding the review questions and the inclusion criteria.

3.2.3 Introduction

The introduction of the review protocol should provide explicit and comprehensive information regarding the justification (rationale) for the conduct of the review in the context of what is already known. The introduction should be of sufficient length to discuss all of the elements of the proposed plan for the review; usually all the relevant information may be provided in approximately 1000 words. This section should be written in simple prose for non-expert readers. Usually, a systematic review is informed by international research and is conducted for an international readership, therefore, reviewers should include relevant international literature in this introductory section. There are exceptions, for example, where systematic reviews are conducted on a question relevant to a single country (for example, Australia or UK) or region (Africa) specific issues. However, with the exception of these reviews that use strict limitations on the inclusion criteria, a systematic review should include all relevant international literature. The introduction should provide sufficient details to justify the conduct of the review and the choice of inclusion criteria for the review (types of participants, types of interventions and comparators, the types of outcomes, and types of studies). The review protocol should provide all conceptual and operational definitions that are relevant for the review. It is the responsibility of the reviewers to ensure that their review is not a duplicate of an existing review. It is recommended that reviewers search major electronic databases to determine that there have been no recently published systematic reviews on the same topic. A search of the *JBIR Evidence Synthesis*, Cochrane Database, MEDLINE, DARE, PROSPERO, EPISTEMONIKOS, and ACCESSSS will assist to establish whether or not a recent review exists on the topic of interest. Reviewers should report in the background section the details of this preliminary search. If systematic reviews on the topic of interest have already been conducted, reviewers should explain the differences between the existing reviews and the new proposal and provide an explicit justification for the need to conduct a new systematic review.

The introduction should conclude with an overarching review objective that captures and aligns with the core elements/mnemonic of the inclusion criteria (e.g. PICO). The stated objective should clearly indicate what the review project is trying to achieve. Example of a review objective: *'To synthesize the best available evidence related to using inspiratory muscle training to improve dyspnoea in patients with chronic obstructive pulmonary disease.'* This broad statement provides the general scope but must be further clarified with focused review questions.

The background section of the review protocol should provide information regarding:

- the importance of the topic (prevalence, incidence, morbidity, mortality, impact on quality of life; economic burden),
- concerns expressed by consumers, healthcare professionals, policy-makers,
- the specifics of diverse groups of patients (age, gender, ethnicity, severity of the disease, co-existing diseases) and settings,
- the intervention of interest and how it works,
- any uncertainties and conflicting reports regarding the effectiveness of the intervention of interest,
- other existing interventions with which the intervention of interest may be compared,
- the importance of different outcomes,
- how outcomes are measured (approaches, measurement instruments),
- the relevance of different research study designs in the examination of the topic of interest,
- relevant existing primary research studies,
- what is already known, including details about the existing systematic reviews, including meta-analyses, and
- the justification for the need for a new review and the objectives of the review project.

3.2.4 Inclusion criteria

The review protocol should provide explicit, unambiguous, inclusion criteria for the review. Inclusion criteria should be reasonable, sound (based on scientific arguments), and justified. These criteria will be used in the selection process, when it is decided if a study will be included or not in the review. Usually, it is enough to provide explicit inclusion criteria without specifying explicit exclusion criteria; it is implicitly assumed that exclusion is based on the criteria that are the opposite of those specified as inclusion criteria. However, sometimes, for clarity, in order to avoid any potential ambiguity, it is recommended to provide explicit exclusion criteria. Inclusion criteria for a review are not intended to be considered in isolation; in this regard they should be articulated so as to be as mutually exclusive as possible and not repeat information relevant to other aspects of the PICO.

Two categories of inclusion criteria should be considered: *inclusion criteria based on study characteristics*, and *inclusion criteria based on publication characteristics*. *Inclusion criteria based on study characteristics* are those related to the types of participants and settings, types of interventions, comparators, types and measurement of outcomes, and types of studies. *Inclusion criteria based on publication characteristics* are those related to publication date, language of publication, type of publication (published in commercial scientific databases; documents not published in commercial databases, for example, trials documents). Usually, reviewers use the PICO mnemonic (participants, intervention, comparator and outcome) to construct a clear and meaningful review objective/question regarding the quantitative evidence on effectiveness of interventions. The reviewer uses the same PICO framework to develop inclusion criteria based on study characteristics. The inclusion criteria must provide adequate details about the conceptual and operational definitions of each element to enable reviewers to make reliable decisions when making decisions to include studies.

3.2.4.1 Population (types of participants)

This section should specify the details about types of participants considered for the review, for example, age; gender; ethnicity; diagnosis; diagnostic criteria; stage or severity of the disease; co-existing diseases. What are the most important characteristics of the population? (e.g., age, disease/condition, severity of illness, setting, gender, etc.).

Consider the following example regarding COPD, describe the population (*patients with COPD*), the severity of illness (*moderate-to-severe*), trajectory of the disease (*stable*), with a specific setting (*community dwelling*). Diagnostic criteria should be made clear to allow inclusion and exclusion; if reviewers anticipate subgroup analysis related to population characteristics, these subgroups should be reflected in the population inclusion criteria. For example, '*COPD includes patients with chronic bronchitis and emphysema but not asthma (fixed airway obstruction with forced expiratory volume in one second [FEV₁] less than <80% of predicted). According to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) and the American Thoracic/European Respiratory Society Guidelines (ATS/ERS), the description of the severity of disease is as follows: stage II or moderate disease is an FEV₁ of 50-80% predicted; stage III or severe is an FEV₁ of 30-50% predicted and stage IV or very severe is an FEV₁ <30% predicted. Patients with reversible airway disease (improvement in FEV₁ >20% with fast acting bronchodilator) will be excluded because their response to training may relate more to changes in their airway obstruction than a training effect.*' 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.

3.2.4.2 Intervention (types of interventions)

What is the intervention? This section should specify the details about the intervention of interest for the review, for example, the nature of intervention, frequency, intensity, timing, and details about those administering the intervention. The same kind of information should be specified for all comparators considered in the review. Where possible, the intervention should be described in detail, particularly if it is multifaceted. A more detailed analytical framework can be used to refer to these complexities. If the review is examining a class or group of interventions, a comprehensive list of identified examples should be provided for the reader. Reviewers should plan any subgroup analysis based on different modes, timing, etc. of the intervention during the protocol stage and account for them in the inclusion criteria. For example, '*inspiratory muscle training includes any mode (threshold loading, resistive, hyperpneic,) practiced at least daily for no less than 4 weeks*' allows the reviewers to consider different types of training but specifies the minimum training period.

3.2.4.3 Comparison (types of comparators)

What is the intervention being compared with? (e.g., placebo, standard care, another therapy or no treatment). This section should detail what the intervention of interest is being compared with. The reviewer may wish to examine the comparative effectiveness of two treatments with a specific, head-to-head comparison. In the example (See Section 3.2.4.3), the reviewers may have specified inspiratory muscle training compared to cardiovascular conditioning. This level of detail is important in determining study selection once searching is complete. Systematic reviews of effectiveness based on the inclusive definition of evidence adopted by the JBI often seek to answer broader questions about multifaceted interventions and comparing the intervention of interest with all existing alternative interventions (comparators).

3.2.4.4 Outcomes

The review protocol should list all the outcomes considered. There is an international initiative known as The COMET (Core Outcome Measures in Effectiveness Trials) initiative, involved in the development and application of agreed standardized sets of outcomes for trials on specific conditions. Details are provided on the COMET website (<http://www.comet-initiative.org/>). Reviewers are encouraged to check the available standardized sets of outcomes for trials relevant for their reviews.

Outcomes should be measurable and appropriate to the review objectives and questions. Usually, only a limited number of primary outcomes and a limited number of secondary outcomes are considered for a review. Sometimes, if justified, it is acceptable to include multiple primary and secondary outcomes. However, the appropriateness of the number and scope of outcomes depend on the specifics of the review objectives and review questions (Aromataris 2015). The relevance of each outcome to the review objective/questions should be justified in the background section. Both beneficial outcomes (positive effects) and harms (negative effects, such as adverse effects or side effects) should be considered as outcomes (Aromataris 2015). Essentially, primary outcomes are those outcomes that are the most important outcomes informing the review questions and the conclusions about the beneficial and harmful effects of the intervention of interest for a review (Aromataris 2015). Secondary outcomes are all other outcomes not specified as primary outcomes. A fundamental distinction is that between true endpoints and surrogate outcomes; true endpoints reflect the effects of treatment on aspects of patients' status considered the most important in terms of mortality and morbidity; surrogate outcomes are measured as "surrogates" for true endpoints, for reasons related to complexity, time, and costs of measurement of true endpoints (Tufanaru 2016). Examples of true endpoints are survival time in cancer and bone fractures in osteoporosis; examples of surrogate outcomes are time to progress from one stage to another stage in cancer and bone mineral density in osteoporosis (Tufanaru 2016).

It is recommended that whenever possible true endpoints should be used as primary outcomes, and that if surrogate outcomes are used as primary outcomes then an explicit justification should be provided for the use of a surrogate outcomes instead of true endpoints (Tufanaru 2016). It is expected that all outcomes specified *a priori* in the review protocol, will be explicitly addressed in the systematic review report, regardless of the existence or not of data from included studies on these outcomes (Aromataris 2015).

A further critical aspect refers to the measurement of the specified outcomes. It is recommended that reviewers present explicit information on available measurement instruments, including details about the validity and reliability properties of these instruments (Aromataris 2015).

As JBI endorses the use of the GRADE approach known as the 'Summary of findings' table, reviewers should be aware that the most important outcomes, that is, the primary outcomes specified in the review protocol should be addressed in the review report and should be explicitly presented in the GRADE Summary of findings' table. Details are provided in the GRADE Handbook (Schunemann et al. 2013).

3.2.4.5 Types of studies

There are three approaches regarding choices for inclusion of studies based on their design in JBI systematic reviews. The first option is to clearly state in the protocol what study designs will be included (for example RCTs), and include only studies that are of this design in the review. This approach is transparent and at low risk of subjectivity during selection of studies. However, it runs the risk of leading to an empty review or a review with few included studies.

The second option is to consider using the hierarchy of study designs for including and excluding studies in the review. In this approach, authors may include other study designs if their preferential study designs are not located. If this is the case, there should be a statement about the primary study design of interest and the other types of studies that will be considered if primary study design of interest is not found. It is common to provide a statement that RCTs will be sought, and that in the absence of RCTs, other study designs will be included, such as quasi-experimental studies and observational studies. This is a pragmatic approach with the aim to include the best available evidence within a review.

The third option is to simply include all quantitative study designs (or all study designs up to a point of the hierarchy of evidence - for example experimental studies and cohort studies, both prospective and retrospective).. This inclusive approach is acceptable as it allows for examination of the totality of empirical evidence and may provide invaluable insights regarding the agreement or disagreement of the results from different study designs. Where feasible, JBI prefers and suggests reviewers consider option 3, the most inclusive approach. However, for many topics, this will present a great deal of information which may not be of use to best inform effectiveness.

3.2.5 Search strategy

This section of a review protocol should provide explicit and clear information regarding two different aspects of locating studies: *all information sources* that will be searched for the review, and the *strategies used for searching*. The aim of a systematic review is to identify all relevant studies, published or not, on a given topic. Searching should be based on the principle of comprehensiveness, with the widest reasonable collection of information sources that are considered appropriate to the review.

A systematic review of effectiveness aims to identify, at a minimum (see Section 3.2.4.5) all data derived from experimental trials (published or not) performed on a specific topic. Two recent international initiatives, one called 'All Trials' (<http://www.alltrials.net/>), and the other one called Restoring invisible and Abandoned Trials abbreviated RIAT (<http://www.bmj.com/content/346/bmj.f2865>) are fundamental in this regard.

The review protocol should list all information sources that will be used in the review: electronic bibliographic databases; search engines; trials registers; specific relevant journals; websites of relevant organizations; direct contact with researchers; direct contact with sponsors and funders of clinical trials; contact with regulatory agencies (for example, US FDA). The review protocol, ideally, should specify all the details (a line-by-line description) of the proposed search strategy used for each electronic bibliographic database considered for the review. As a minimum, all the details of the proposed search strategy for at least one major electronic bibliographic database (such as PubMed) should be provided in an appendix. The review protocol should specify the timeframe for search, and any language and date restrictions, with appropriate justifications. The reviewers should consider the potential consequences of language and date search restrictions. If possible, authors should always seek the advice of an expert research librarian when developing a search strategy. Involvement of a research librarian in the development of a search strategy should be acknowledged. For JBI systematic reviews, the search strategy is often described as a three-phase process beginning with the identification of initial key words that are used in a limited number of databases (for example, PubMed and CINAHL); followed by an analysis of the text words contained in the title, abstract and index terms used to describe relevant articles. The second phase consists of the use of database-specific searches for each database specified in the review protocol. The third phase includes the examination of the reference lists of all studies already retrieved with the explicit aim to identify additional relevant studies. The list of all databases that will be considered for database-specific searches should be provided. Usually, a comprehensive search for a review of effectiveness includes a search of relevant multiple bibliographic databases (for example, PubMed, CINAHL, EMBASE etc.), a search of trial registers, a search of relevant grey literature sources, and a hand-search of relevant journals. Reviewers should provide enough information in order to persuade readers that the sources of information considered are relevant and comprehensive and the search strategy is comprehensive and sound. Reviewers are encouraged to read the article by Aromataris and Riitano (2014) regarding searching for evidence.

3.2.6 Selection of studies

This section should describe the process of study inclusion for all stages of selection (based on title and abstract examination; based on full text examination) and the procedures for solving disagreements between reviewers. The software used for the management of the results of the search should be specified (e.g. Covidence, Endnote). Selection is performed based on inclusion criteria (See Section 3.2.4) pre-specified in the review protocol. In a systematic review study selection (both at title/abstract screening and full text screening) should be performed by two or more reviewers, independently. Any disagreements are solved by consensus or by the decision of a third reviewer. JBI reviewers are encouraged to read the article by Porritt et al (2014) regarding study selection and critical appraisal.

3.2.7 Critical appraisal

This section should describe the critical appraisal process and instruments that will be used in the review process and the procedures for solving disagreements between reviewers.

The goal of critical appraisal (assessment of risk of bias) is to assess the methodological quality of a study and to determine the extent to which a study has excluded or minimized the possibility of bias in its design, conduct and analysis. Bias refers to systematic errors in the design, conduct and analysis of quantitative studies that may impact the validity of inferences from these studies. Critical appraisal of the studies included in a systematic review is performed with the explicit goal of identifying the risk of diverse biases in these studies. JBI uses standardized critical appraisal tools for the assessment of risk of diverse biases encountered in quantitative studies. There are JBI standardized appraisal tools based on study design appropriate for JBI reviews of effectiveness (see Appendix 3.2 regarding the JBI standardized appraisal tools). JBI systematic reviews are required to use these JBI standardized appraisal tools. Reviewers should refer in the review protocol to the JBI standardized critical appraisal checklists and provide references for these checklists. It is not necessary to provide these checklists in appendices of the review protocol. If non-JBI appraisal tools are proposed then these tools should be briefly described and correctly referenced. In this case, an explicit justification for the use of non-JBI appraisal tools should be provided in the review protocol.

Two reviewers should perform independent critical appraisal of retrieved studies using the standardized critical appraisal checklists developed by JBI. The protocol should specify that any disagreements are solved by consensus or by the decision of a third reviewer. In experimental studies (randomized experimental studies and quasi-experimental studies) the most important biases are: selection bias, performance bias, attrition bias, detection bias, and reporting bias. In observational studies the most important biases are: selection bias, information bias, and confounding. The review protocol should specify that reviewers plan to report in narrative form and in tables the results of risk of bias (methodological quality) assessments for each aspect of methodological quality (randomization; blinding; measurement; statistical analysis etc.) for each individual study and the overall risk of bias of the entire set of included studies. The critical appraisal phase of the review should not be treated as a rapid 'box ticking exercise' on checklists, but rather as a complex, profound, critical, systematic, thorough examination of the risk of bias of each included study, a solid foundation for an appropriate synthesis of the results.

The review protocol should specify if and how the results of critical appraisal will be used for the exclusion of studies from the review. For example, if studies judged of low methodological quality will be excluded from the review, the details of the circumstances under which such decisions will be made and the explicit criteria or decision rules should be explicitly provided, including explanations for what is considered low methodological quality by reviewers. It is the decision of the review team if they want to exclude from the review studies judged of low methodological quality. Reviewers should explain and justify their criteria and decision rules. The decision as to whether or not to include a study can be made based on meeting a predetermined proportion of all criteria, or on certain criteria being met. It is also possible to weight the different criteria differently. The decisions about the scoring system and the cut-off for inclusion of a study in the review should be made in advance and be agreed upon by all participating reviewers before critical appraisal commences. The review protocol should specify if and how the results of critical appraisal will be used in the synthesis (narrative synthesis or meta-analysis) of the results. It is recommended that the results of critical appraisal should be used in the synthesis phase of the review, for the critical examination of the impact of methodological quality of studies on results (including subgroup analysis or sensitivity analysis). JBI reviewers are encouraged to read the article by Porritt et al (2014) regarding study selection and critical appraisal.

3.2.8 Data extraction

This section of the review protocol should specify the data extraction process and instruments that will be used in the review process, as well as the procedures for solving disagreements between reviewers. Complete and accurate data extraction is essential for a good quality systematic review. Reviewers should carefully consider all the relevant data that should be extracted for the review given the focus of the review, the review objectives/questions, and the inclusion criteria. Details regarding the publication and the study, the participants, settings, the interventions, the comparators, the outcome measures, study design, statistical analysis and results, and all other relevant data (funding; conflict of interest etc.) should be carefully and accurately extracted from all included studies. In a review assessing effectiveness, thorough extraction of details of the intervention is essential to allow for reproducibility of an intervention that is found to be effective (Munn et al. 2014). In a JBI systematic review data extraction is performed by two or more reviewers, independently, using the standardized data extraction form developed by JBI. Any disagreements about data extraction are solved by consensus or by the decision of a third reviewer. If non-JBI data extraction forms are used these should be briefly described and the justification for their use should be explicitly indicated. The review protocol should specify if authors of studies will be contacted by reviewers in order to clarify existing data, to request missing data or additional data. The review protocol should specify the pre-planned approach for the situations when there are multiple reports (publications) for the same study, and for missing data and for data conversion /transformation.

3.2.9 Data synthesis

This section should describe how the data will be combined and reported in the systematic review. Essentially, in a systematic review of effectiveness there are two synthesis options: statistical synthesis (meta-analysis) and narrative summary (narrative synthesis). Details of the statistical models and methods and effect estimates that will be calculate and measures of statistical heterogeneity should be included (See Section 3.3). Authors should ensure that the effect estimates that will be calculate correspond to the type of data (dichotomous and/or continuous) they have suggested will be collected in their protocol (see Section 3.2.4.4). The review protocol should also explicitly specify the pre-planned approaches that will be used for the examination of publication bias, including the use of funnel plots and the use of statistical tests for the examination of publication bias (see Section 3.3.11).

The review protocol should explicitly specify that reviewers plan to use the GRADE approach for the reporting of the strength of evidence, including the reporting of the summary of findings table of evidence. The use of GRADE approach is currently endorsed by JBI and JBI reviewers must use it regardless of the synthesis approach employed, meta-analysis or narrative synthesis.

3.3 Meta-analysis

Meta-analysis refers to the statistical synthesis of quantitative results from two or more studies. The review protocol should state that statistical meta-analysis of data will be conducted if appropriate and that if meta-analysis is not possible, narrative synthesis will be conducted as the primary mechanism of data synthesis. Narrative summary should be included to supplement the technical details provided on the process and results even if meta-analysis is performed and to provide synthesis of data not captured in statistical meta-analysis.

Meta-analysis should be reserved for the results of studies that are considered similar enough from a clinical and methodological point of view (homogeneous studies). If studies are heterogeneous from a clinical or methodological point of view, then it is uncertain if it is appropriate to synthesize the respective studies into meta-analysis. Any meta-analysis where studies are heterogeneous from a clinical or methodological point of view will require substantial justification by the authors. Clinical heterogeneity refers to differences between studies with regards the participants, interventions, comparators, settings, and outcomes. Methodological heterogeneity refers to the study design and the methodological quality of the studies (risk of bias). Studies that are similar with regards the participants, interventions, comparators, settings, outcomes, study design, and risk of bias may be combined in meta-analysis. The judgement that studies are homogeneous enough and that it is appropriate to combine the studies in meta-analysis should be based on the understanding of the review question, the characteristics of the studies, and the interpretability of the results. The decision should not be based just on statistical considerations regarding heterogeneity (Sutton et al 2000).

The review protocol should specify the appropriate possible, reasonable details regarding the anticipated (pre-planned) meta-analysis:

- Objectives of the meta-analysis,
- Meta-analysis model (fixed effects model or random effects model) and the justification,
- Effect size to be used (OR, RR, etc.),
- Meta-analysis method (Peto method etc.) and justification,
- Statistical testing procedures used for the exploration of statistical heterogeneity (such as Q Cochran test) and the rules used for the interpretation of the results,
- Statistical indicator used for the quantification of statistical heterogeneity (such as I^2) and the rules used for the interpretation of the results,
- Pre-planned sensitivity analyses and their justification, and
- Pre-planned subgroup analyses and their justification.

3.3.1 Objectives of meta-analysis

The objectives of meta-analysis should be pre-specified in the review protocol. There are different legitimate objectives for a meta-analysis: to improve statistical power to detect a treatment effect, to estimate a summary average effect, to identify subsets of studies (sub-groups) associated with a beneficial effect, and to explore if there are differences in the size or direction of the treatment effect associated with study-specific variables (Normand 1999).

3.3.2 Statistical models for meta-analysis

There are three categories of statistical models for meta-analysis: the fixed effects model, random effects model, and mixed effects models (Hedges 1992). Only the first two models are used in JBI SUMARI for meta-analysis and discussed here. Using the fixed-effect model we assume that the true effect size for all studies is identical and the effect sizes estimated in studies are different only due to errors in estimating the effect size (Borenstein et al 2010). In the random-effects model we assume a distribution of effects, not a common identical effect size, and we assume that the meta-analysis summary effect size is an estimate of the mean of a distribution of true effects, not a common shared effect size identical for all studies (Borenstein et al 2010).

The proposed statistical model for meta-analysis should be explicitly indicated in the review protocol. When considering statistical inference, meta-analysis using the fixed effects model is appropriate if the aim is to draw statistical conclusions only about the studies included in the meta-analysis, and that the random effects model is appropriate whenever statistical generalizations beyond the included studies are considered (Cooper and Hedges 1994). Commonly, review authors want to generalize the conclusions beyond the actual studies included in meta-analysis, therefore we suggest that the default model for meta-analysis in JBI reviews should be the random effects model. However, it has been recommended by statisticians that the fixed effects model is the appropriate model whenever the number of studies is small (less than five studies) (Cooper and Hedges 1994; Murad et al 2015, p.511). Further details about the fixed effects and random effects models for meta-analysis, including a flowchart for the decisions regarding the selection of the meta-analysis model are provided by Tufanaru et al (2015).

3.3.3 Effect sizes

In this section, effect sizes refer to quantitative indicators of the direction and magnitude of the effects of the interventions on outcomes. Common effect sizes reported in meta-analysis include the risk ratio (RR), risk difference (RD), odds ratio (OR), weighted mean difference (WMD), and standardized mean difference (SMD).

3.3.4 Considerations for the meta-analysis of dichotomous data

For meta-analyses, computation of the logarithm (log) of the RR or the log of OR, or the RD from each individual study may be used or the number of events and the total number of participants for each group. RR and RD may be computed for any experimental study (RCT) or quasi-experimental study or cohort studies. Odds ratios may be computed for any study design (experimental, quasi-experimental, cohort, case-control, or analytical cross-sectional studies). Fleiss (1994) discussed the statistical properties of the OR and concluded that the OR is the preferred effect size for the computation phase of the meta-analysis of binary data regardless of the study design of the studies. However, the OR is not easily interpretable. Therefore, reviewers should be careful in providing correct explicit interpretation of the odds ratios computed in meta-analysis. Reviewers should provide the results expressed using both absolute (RD) and relative (RR) effect sizes for meta-analysis of binary data. Reviewers should provide correct explicit interpretation of the computed effect sizes.

3.3.5 Considerations for the meta-analysis of continuous data

For the effect sizes related to differences in continuous data (WMD, SMD), the data regarding the mean response, the standard deviation, and the number of participants in each group are used. The difference in means is the difference between the mean response in the intervention group and the mean response in the control group. This may be the difference in the means between groups at the final measurement of outcomes, or it may be the difference between the means in their changes from baseline. The simple difference in means is also called the mean difference (MD) or the weighted mean difference (WMD). We will use the term the WMD in this chapter. The WMD is used in meta-analysis of continuous data if all studies included in meta-analyses measured the outcome using the same measurement instrument. For meta-analysis computation the difference in means from each individual study are used. The results are expressed in the natural (clinical) units used for the common measurement instrument. If WMD is used, reviewers should provide explanations regarding the interpretation of the results expressed in units used for the common measurement instrument. The minimum score and the maximum score that are possible on the measurement instrument should be specified together with their interpretation. Also, reviewers should specify what change (difference) is considered significant from a practical or clinical point of view. Reviewers should explain the interpretation of a negative or positive difference. The standardized mean difference (SMD) is a difference in means that is standardized by using information on the variability of data (standard deviation). There are three methods (formulas) that are commonly used for the computation of SMD: Cohen's *d*, Hedges' adjusted *g*, and Glass's *delta*. These three formulae use different standard deviations in their computation. Currently, the JBI SUMARI software offers capabilities for the computation of Cohen's *d*. The SMD is used in meta-analysis of continuous data if the studies measured the same outcome but with different measurement instruments. For meta-analysis computation the SMD from each individual study are used. The results are expressed in units of standard deviation. Reviewers should provide explanations regarding the interpretation of the results. In order to facilitate the interpretation of the results it is recommended that reviewer's convert the results into natural (clinical) units by multiplying the results expressed in units of standard deviation with the standard deviation of the scores from a study on a known measurement instrument. The instrument chosen may be the most commonly used instrument or the instrument which has the best psychometric properties. Reviewers should explain the interpretation of differences and justify what is considered a small or medium or large difference; explanations should be provided for negative or positive differences.

3.3.6 Meta-analysis: Statistical Methods

Different statistical methods are available for meta-analysis: Mantel-Haenszel method, Peto's method, DerSimonian and Laird method, and the inverse variance method. The Mantel-Haenszel method, the Peto's method, and the inverse variance method are methods used with the fixed effects model of meta-analysis (Deeks et al 2008). The DerSimonian and Laird method is used with the random effects model of meta-analysis (Deeks et al 2008).

The inverse variance method may be used with all types of ratios and differences for example the log odds ratio, log relative risk, risk difference, mean difference (weighted mean difference) and standardized mean difference (Petitti 2000; Deeks et al 2008). The Mantel-Haenszel method may be used with ratios, typically with odds ratio, but can be applied to rate ratio and risk ratio (Petitti 2000). The Peto's method is used with odds ratios (Petitti 2000). DerSimonian and Laird method may be used with all types of ratios (odds ratio, risk ratio) and difference (weighted mean difference) and standardized mean difference (Petitti 2000; Deeks et al 2008).

There are different statistical methods (formulae) used to compute a standardized mean difference for each study including the Hedges' method, the Cohen's method, and the Glass method. If a fixed effects model is used for meta-analysis of standardized mean differences then the inverse variance method of meta-analysis may be used. If a random effects model is used for meta-analysis of standardized mean differences then the DerSimonian and Laird method may be used.

When deciding what method for meta-analysis to be used statistical considerations are important. When studies have small sample sizes and the number of events is small in these studies the inverse variance method may not be appropriate; in these circumstances, it may be preferable to use the Mantel-Haenszel method (Deeks et al 2008). Peto's method may produce serious under-estimates when the odds ratio is far from unity (large treatment effects) (Sutton et al 2000). If the number of studies to be combined is small, but the within-study sample sizes per study are large, the inverse-weighted method should be used (Sutton et al 2000, p.69). If there are many studies to combine, but the within-study sample size in each study is small, the Mantel-Haenszel method is preferred (Sutton et al 2000).

3.3.7 Subgroups in meta-analysis

Subgroups refer to diverse grouping of studies based on specific characteristics of the studies such as study design. These characteristics may include the types of participants, types of comparators, and the outcomes. For example, it is possible to group all randomized experimental studies in one subgroup and all observational studies in another group; similarly reviewers may wish to group all studies with young participants in one subgroup and all studies with older participants in another subgroup. For these subgroups, it is possible to perform meta-analysis and to report the summary effects computed within subgroups. Also, it is possible to compare the summary effects computed in diverse subgroups. It is recommended that if subgroup analyses are performed these should be limited in number, should be pre-planned in the review protocol, and explanation and justification should be explicitly provided. These analyses should be carefully interpreted.

3.3.8 Sensitivity analysis in meta-analysis

As there are many decisions involved in meta-analyses it is important to perform a sensitivity analysis in order to explore the impact of different decisions on results. For example, one sensitivity analysis may explore the impact of using different meta-analysis models. Another sensitivity analysis may explore the impact of excluding or including studies in meta-analysis based on sample size, methodological quality, or variance. If results remain consistent across the different analyses, the results can be considered robust as even with different decisions they remain the same/similar. If the results differ across sensitivity analyses, this is an indication that the result may need to be interpreted with caution.

3.3.9 Meta-regression

Meta-regression analysis aims to examine if characteristics of studies are associated with the magnitude and direction of the effect in studies included in meta-analysis. However, given the strict statistical circumstances under which it is appropriate to perform meta-aggregation and also the advanced statistical skills required to use meta-regression software, we cannot recommend the common use of these methods in meta-analysis in JBI reviews of effectiveness.

3.3.10 Heterogeneity

There are different statistical approaches for investigating heterogeneity, included the standard chi-squared test, the I square statistic, and Tau squared.

3.3.10.1 Standard chi-squared test (Cochran test)

The standard chi-squared test (Cochran Q test) for statistical heterogeneity tests the statistical hypothesis that the true treatment effects (the effect size parameters) are the same in all the primary studies included in meta-analysis (Sutton et al 2000). This statistical test uses a test statistic Q that has a chi-squared distribution on $k-1$ degrees of freedom (k represents the number of studies) under the statistical hypothesis; the corresponding p-value for the test statistic is examined (Sutton et al 2000). The statistical power of the test is in most cases very low due to the small number of studies; heterogeneity may be present even if the Q statistic is not statistically significant at conventional levels of significance such as 0.05. A cut-off significance level of 0.10 rather than the usual 0.05 has been advocated (Sutton et al 2000). If results of the test are statistically significant ($p < 0.05$) the statistical hypothesis that the true treatments effects (the effect size parameters) are the same in all the primary studies included in meta-analysis (the hypothesis of homogeneity) is rejected, therefore, it is considered that there is statistical heterogeneity. With a small number of studies (< 20), the Q test should be interpreted very cautiously (Huedo-Medina et al 2006). It is not appropriate to decide the meta-analysis model (fixed or random effects model) based on the results of the Chi squared statistical test (Q test) for heterogeneity.

3.3.10.2 Quantification of the statistical heterogeneity: I squared

The I square statistic (I^2) represents the percentage of the variability in effect estimates that is due to heterogeneity (Deeks et al 2008). I^2 is the proportion of observed dispersion of results from different studies included in a meta-analysis that is real, rather than spurious (Borenstein et al 2009). The I^2 index can be interpreted as the percentage of the total variability in a set of effect sizes due to true heterogeneity (between-studies variability) (Huedo-Medina et al 2006). If $I^2 = 0\%$, this indicates that all variability in effect size estimates is due to sampling error within studies. If $I^2 = 50\%$, it indicates that half of the total variability among effect sizes is caused not by sampling error, but by true heterogeneity between studies (Huedo-Medina et al 2006). I^2 is a percentage and its values lie between 0% and 100% (Higgins et al 2003). A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity (Higgins et al 2003). One proposed suggestion was to consider as low, moderate, and high heterogeneity for I^2 values of 25%, 50%, and 75% (Higgins et al 2003). Another guide to interpretation was proposed: 0% to 40% might not be important; 30% to 60% may represent moderate heterogeneity; 50% to 90% may represent substantial heterogeneity; 75% to 100% considerable heterogeneity (Deeks et al 2008). Authors of the guide mention that careful interpretation of the value of I^2 depends on magnitude and direction of effects and strength of evidence for heterogeneity (Deeks et al 2008). With a small number of studies (< 20) and/or average sample size ($N < 80$) the statistical power for I^2 procedures is less than the usually recommended minimum value of 0.8 (Huedo-Medina et al 2006). With a small number of studies (< 20), both the I^2 confidence interval and the Q test should be interpreted very cautiously (Huedo-Medina et al 2006).

3.3.10.3 Tau-squared for random effects model meta-analysis

In random-effects meta-analysis, the extent of variation among the effects observed in different studies (between-study variance) is referred to as tau-squared, τ^2 , or Tau^2 (Deeks et al 2008). τ^2 is the variance of the effect size parameters across the population of studies and it reflects the variance of the true effect sizes. The square root of this number is referred to as tau (T). τ^2 and Tau reflect the amount of true heterogeneity. τ^2 represents the absolute value of the true variance (heterogeneity). τ^2 is the variance of the true effects while tau (T) is the estimated standard deviation of underlying true effects across studies (Deeks et al 2008). The summary meta-analysis effect and T as standard deviation may be reported in random-effects meta-analysis to describe the distribution of true effects (Borenstein et al 2009).

3.3.11 Publication bias

Publication bias occurs when published studies differ systematically from all conducted studies on a topic. Publication bias arises when studies with statistically significant results or positive results in a specific direction are more likely to be published compared to studies without statistically significant results or negative results. Reviewers should make all reasonable efforts to include in their systematic review all or most of all relevant studies, regardless of the nature of reports (published or unpublished). Publication bias can have a detrimental effect on the validity of systematic reviews (Deeks et al 2008). Funnel plots are a method of investigating the located studies in a meta-analysis for publication bias, they are scatter plots in which an effect estimate of each study is plotted against a measure of size or precision (i.e. standard error) (Deeks et al 2008). The largest studies should be closest to the 'true' value, with the smaller studies spread on either side; creating the shape of a funnel if publication bias is not present. If publication bias has had an effect on the studies available (and there are no other confounding factors) then the 'funnel' should be incomplete with an area missing (Deeks et al 2008). Generally the best way to minimise the impact of publication bias on a systematic review is the inclusion of trial registries and unpublished studies or grey literature (Lau et al 2006; Sterne et al 2011). Funnel plots suffer from numerous issues including low power, numerous alternative explanations for asymmetrical distribution of studies, and inaccurate researcher interpretations of plots (Lau et al 2006; Sterne et al 2011). However, they remain a useful and popular way of investigating publication bias (Deeks et al 2008). Potential reasons for funnel plot asymmetry other than publication bias include: poor methodological quality leading to exaggerated effects in smaller studies (which can be the result of poor methodological design, inadequate analysis, or fraud), true heterogeneity, artefactual causes (in some situations sampling variation can lead to an association between the two factors (effect estimate and measure of size or precision)) and chance (Sterne et al 2011). The visual inspection of funnel plots introduces great uncertainty and subjectivity. In a survey utilizing simulated plots, researchers had only 53% accuracy at identifying publication bias (Lau et al 2006). A very liberal minimum number of studies for the performance of a funnel plot to be justified is ten (Lau et al 2006).

Statistical tests for funnel plot asymmetry (also known as tests for publication bias) investigate the association between effect size estimate and measure of study size or precision. The most popular statistical tests for funnel plot asymmetry are Egger test, Begg test, and the Harbord test. These tests were developed based on the following assumptions: large studies are more likely to be published regardless of statistical significance; small studies are at the greatest risk for being lost; in small studies only the large effects are likely to be statistically significant therefore published small studies often show larger effect sizes compared to larger studies; small and unfavorable effects are more likely to be missing; small studies with large effect sizes are likely to be published (Jin et al 2015). Null statistical hypotheses for these tests reflect the hypothesis of symmetry of the plot, that is, the hypothesis of no publication bias. A finding of not statistically significant P-value for the asymmetry test does not exclude bias. These tests are known to have low power.

A statistical test for funnel plot asymmetry investigates whether the association between effect estimate and measure of study size or precision is larger than what can be expected to have occurred by chance (Sterne et al 2011). These tests are known to have low power and consequently a finding of no evidence of asymmetry does not serve to exclude bias (Sterne et al 2011).

The Begg's Test was proposed by Begg and Mazumdar in 1994. It is used for dichotomous outcomes with intervention effects measured as odds ratios. It is an adjusted rank correlation test (Jin et al 2015). It explores the correlation between the effect estimates and their sampling variances (Jin et al 2015). It is a very popular test, however, it has low power; some statisticians do not recommend its use. It is "fairly powerful" for meta-analysis of 75 studies; it has "moderate power" for meta-analysis of 25 studies (Begg and Mazumdar 1994). It is considered that the test has "appropriate" type I error rate (Jin et al 2015).

The Egger's test was proposed by Egger et al in 1997. It is used for continuous outcomes with intervention effects measured as mean differences. It is a "regression test", that is, it uses a linear regression approach (Jin et al 2015). The standard normal deviate (estimated effect size/estimated standard error) is regressed against the estimate's precision. It is a very popular test. It is considered that the test has "inappropriate" type I error rate when heterogeneity is present and the number of included studies is large (Jin et al 2015). The Egger test for funnel asymmetry is the most cited statistical test for publication bias.

The Harbord Test was proposed by Harbord et al in 2006. It is used for dichotomous outcomes with intervention effects measured as odds ratios. The test uses "a weighted regression model" (Jin et al 2015). It is considered that the test has "inappropriate" type I error rate when heterogeneity is present. It was contended that the Harbord Test has better error rate compared to Egger's test in balanced trials with little or no heterogeneity (Jin et al 2015).

3.4 Systematic review of effectiveness

A systematic review report is important because it provides all the details regarding the conduct of the systematic review and the best available evidence to inform the question posed by the review. Essentially, the content of the sections of the review protocol and the review report are conceptually the same, particularly the background and the methods section. The review protocol specified the proposed plan for the review; the review report reports the conduct of the review, what was actually performed and the results of the review undertaking. All deviations from what was pre-planned in the review protocol should be explicitly reported and justified in the review report.

3.4.1 Title

A clear, descriptive title is important to assist readers and users to readily identify the scope and relevance of the review. The review report title should accurately describe and reflect the content of the review, and should not be phrased as a question. The review title should explicitly identify the publication as a report for a finalized systematic review. It is important to indicate in the review title the focus of the review on effectiveness; we recommend the following convention: *'The effectiveness of [intervention] compared to [comparator] on [outcome]: a systematic review'*. The title of the review should be as descriptive as possible and reflect all relevant information. Ideally, the review title should include in a concise way the relevant information with regards to the types of participants, types of interventions and comparators and the types of outcomes considered in the review.

3.4.2 Abstract

This section forms a structured abstract of the main features of the systematic review. It must be no longer than 500 words and should contain no abbreviations or references. The abstract must accurately reflect and summarize the systematic review with the main focus on the results of the review.

The abstract should report the essential elements of the review using the following sub-headings in this order:

- **Objective:** State an overarching review objective structured using the key components of the inclusion criteria (approximately one to two sentences).
- **Background:** Briefly describe what is already known on the topic and what this review will add to the evidence-base (approximately two to three sentences).
- **Inclusion criteria:** Summarize the inclusion criteria as it relates to the type of review being conducted. Present the information in one or two sentences – **NOT** under individual subheadings.
- **Methods:** List the key information sources searched (those that provided the majority of included studies), any limits placed on the scope of the search (e.g. language), and the date range, or the date of the last search. If the recommended JBI approach to critical appraisal, study selection, data extraction and data synthesis was used, simply state it as such (without naming the actual tool). Otherwise, briefly describe any notable deviations to the methodological approach taken (e.g. criteria used to exclude studies on the basis of methodological quality etc.).
- **Results:** The bulk of the abstract should be reserved to convey the main results of the review.
 - As a general rule, report the number and type of included studies and participants, as well as any pertinent study characteristics. Summarize the overall quality of the included studies and notable aspects of risk of bias.
 - Report the results for all main outcomes (not only those that were statistically significant or clinically important). If meta-analyses were conducted report the summary measures (estimated effect) and confidence intervals and ensure statistics are presented in a standard way. If a meta-analysis was proposed but not conducted, report the reason (e.g. clinical or methodological heterogeneity). Where possible, indicate the number of studies and participants for each main outcome. Describe the direction of the effect (e.g. lower, fewer, greater, more, etc.) in a way that is understandable to patients and health care professionals (i.e. which group was favored and the size of the effect) and indicate the measurement scale used, where applicable.
- **Conclusions:** Provide a conclusion based on a general interpretation of the results considering, for example, the methodological quality of the included studies and any limitations of the review. Briefly convey key implications for practice and/or research.

3.4.3 GRADE 'Summary of Findings' table

The use of the GRADE approach is currently endorsed by JBI and JBI reviewers must use it regardless of the synthesis approach employed, meta-analysis or narrative synthesis. The GRADE 'Summary of Findings' table should be presented immediately below the abstract. The GRADE 'Summary of Findings' table can be developed following the guidance in the [GRADE handbook](#) (Schunemann et al. 2013). Links to resources and [support for using GRADE](#) are available via the [JBI Adelaide GRADE Centre](#).

3.4.4 Introduction

The introduction of the review report should provide explicit and comprehensive information regarding the justification (rationale) for the conduct of the review in the context of what was already known. Ideally, this section of the review report should be a revised, expanded, version of the introductory section from the review protocol. See Section 3.2.3 from the review protocol for further information regarding the content of the introduction.

The introduction should conclude with an overarching review objective that captures and aligns with the core elements/mnemonic of the inclusion criteria (e.g. PICO). The stated objective should clearly indicate what the review project is trying to achieve. Vancouver style of referencing should be used throughout the protocol with superscript numbers without brackets, used for in-text citations.

3.4.5 Review question(s)

The review question(s) should be explicitly stated in unambiguous terms. See the Section 3.2.2 of this Chapter for further information regarding the objectives and questions of a review of effectiveness.

3.4.6 Inclusion criteria

This section should describe the inclusion criteria used for the review. Information should be provided regarding the types of participants, types of interventions, comparators, types of outcomes, and types of studies actually considered and included in the review. See Section 3.2.4 for further details regarding specification of inclusion criteria in the systematic review report.

3.4.7 Methods

This section of the review report is reserved for the methods used to conduct the review and should be presented under the relevant subheadings (See Sections 3.4.6.1 to Section 3.4.6.5), including any deviations from the method outlined in the *a priori* protocol. In empty reviews for example, this section should not refer to methods that were not performed.

Directly below the Methods heading provide the following information:

- State and appropriately cite the JBI methodology that was employed in the conduct of the review and synthesis.
- Refer to and cite the *a priori* protocol that was published, or accepted for publication (e.g. 'in press'), in the [JBI Evidence Synthesis](#).
- If the protocol has been registered with PROSPERO, provide registration information including registration number (e.g. PROSPERO CRD42015425226).

3.4.7.1 Search strategy

The search strategy section of a review report should provide explicit and clear information regarding all information sources that were actually used in the review, and the actual strategies used for searching. The review report should provide details regarding all information sources that were used in the review: electronic bibliographic databases; trial registers; relevant journals; websites of relevant organizations; etc. The review report, ideally, should specify all the details (a line-by-line description) of the actual search strategy used for each electronic bibliographic database used for the review and should be provided in an appendix. The review report should specify the timeframe for search, the date of last search for each database, and any language and date restrictions, with appropriate justifications.

3.4.7.2 Study screening and selection

The review report should describe the actual process of study screening and for all stages of selection (based on title and abstract examination; based on full text examination) and the actual procedures used for solving disagreements between reviewers.

3.4.7.3 Critical appraisal

The review report should specify the critical appraisal process and instruments that were actually used in the review process and the procedures for solving disagreements between reviewers. The review report should describe how the results of critical appraisal were used for the exclusion of studies from the review, if this is the case. The details of the decisions processes and criteria used for exclusion of studies based on results of critical appraisal should be explicitly provided. All details about the scoring systems and the cut-off scores for inclusion of studies in the review should be described and justified.

3.4.7.4 Data extraction

The review report should specify the data extraction process and instruments that were used in the review process and the procedures for solving disagreements between reviewers.

3.4.7.5 Data synthesis

The review report should explicitly specify how the data were combined and reported. Essentially, the review report should provide the details about all preformed analyses and their justifications. The synthesis approaches by which studies were combined should be described in as much detail as is reasonably possible and to enable them to be reproduced.

If meta-analysis was performed, the review report should specify the details regarding the performed meta-analyses. The report should specify:

- the objectives of the meta-analysis
- the effect size used (OR, RR, etc.)
- the meta-analysis model (fixed effects model or random effects model) and the justification
- the meta-analysis method (Peto method etc.) and the justification
- the statistical testing procedures used for the exploration of statistical heterogeneity (such as Q Cochran test) and the rules used for the interpretation of the results
- the statistical indicator used for the quantification of statistical heterogeneity (such as I^2) and the rules used for the interpretation of the results
- the performed sensitivity analyses
- the performed subgroup analyses

3.4.8 Results

This section of the review report has distinct sub-sections describing the process of study inclusion, the methodological quality of the eligible studies, detailed characteristics and description of the included studies and, importantly, the findings of the review and results of the synthesis processes.

3.4.8.1 Study inclusion

This section should provide a narrative summary of the search results and selection process and results. The number of papers identified by the search strategy and the number of papers that were included and excluded should be stated.

A complete and accurate report should be provided regarding:

- the number of studies identified by the search in diverse sources;
- the number of studies excluded after the examination of title and abstract against inclusion criteria;
- the number of full text articles retrieved for examination;
- the number of studies excluded after full text examination against inclusion criteria;
- the number of critically appraised studies;
- the number of studies excluded after critical appraisal;
- the final total number of included studies.

A flowchart using the PRISMA template for the reporting of the selection process should be included.

Ideally, a list of all excluded studies, excluded after full text examination and after critical appraisal, with the explicit reasons for exclusion, should be provided in appendices to the review. As a minimum, at least the list of studies excluded after critical appraisal and the reasons for exclusion should be reported. If no studies were excluded after critical appraisal then the list of all studies excluded after full text examination including the explicit reasons for exclusion, should be provided in appendices to the review.

3.4.8.2 Methodological quality

The review report should report in a comprehensive manner, in narrative form and in tables, the results of risk of bias (methodological quality) assessments for each aspect of methodological quality (randomization; blinding; measurement; statistical analysis etc.) for each individual study and the overall risk of bias of the entire set of included studies. This section must provide an overarching statement of the quality of the included studies as a whole (i.e. low, moderate, high, etc.) and a narrative summary of the methodological quality of the included studies against each of the critical appraisal criteria, with a clear indication of the risks of bias present across the included studies (e.g. performance bias, detection bias etc.). Reporting can be supported (optional) by a table showing the results of the critical appraisal (see Table 3.1 for example). Where only few studies are identified, or there are specific items of interest from included studies, these should be addressed in the narrative also, particularly where studies were deficient, or particularly good. Use of 'Unclear' and 'Not Applicable' should also be explained in the text.

Table 3.1. Critical appraisal results for included studies using the JBI-Critical Appraisal Checklist for randomised controlled trials

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Author(s) ^{ref}	Y	Y	Y	N	Y	U	Y	N	Y	U
Author(s) ^{ref}	Y	N	Y	Y	Y	U	Y	N/A	Y	Y

Y - Yes, N - No, U - Unclear, N/A - not applicable

If appraisal tools are not appended to the review report (citation only), the appraisal questions should be added as a footnote/caption to the table (Table 3.1) so readers can clearly interpret the information presented.

3.4.8.3 Characteristics of included studies

This section should include a narrative summary of the details about the design and details of the included studies. Relevant characteristics of the included studies for which data were extracted and are needed to understand and interpret the results of the study should be synthesized in narrative. This includes the descriptive and demographic features (e.g. the country and setting of the study) of the included studies, as well as the main clinical characteristics, as they relate to the review objective and the inclusion criteria (e.g. PICO). For example, in a review of effects, synthesize characteristics of the population, intervention, comparator, outcomes, and study design. Information on interventions should include treatment modalities and the amount, duration, frequency and intensity of the intervention any details related to the follow-up of the participants. Population characteristics should include the number of participants (i.e. study size) and demographic information such as age, gender and any information relevant to the specific review question (e.g. past medical history, diagnosis, co-morbidities).

Reviewers should provide an appendix of the review report summarized details of the included studies. The examination of the table of included studies should suffice to convince the readers that there is good match between the included studies and the inclusion criteria.

3.4.8.4 Results and meta-analysis

This section should be organized in a meaningful way based on the review objectives and questions and types of interventions, comparators, outcomes and types of studies. This section should provide comprehensive information regarding the results of all performed meta-analyses and additional analyses such as sensitivity analysis and sub-group analysis. Point estimates and interval estimates (confidence intervals) should be reported. Before presenting any meta-analysis results, the conduct of meta-analyses should be justified; reviewers should explicitly provide commentaries regarding the clinical, methodological, and statistical heterogeneity of the studies included in meta-analyses and the appropriateness of conducting meta-analyses. Summary results from meta-analyses should be reported as summary point estimates and interval estimates. The meta-analysis forest plots for all performed meta-analyses should be presented in this section. A narrative summary should complement the forest plots and provide additional commentaries and explanations for all performed meta-analyses (Munn et al 2014).

Reviewers should report the funnel plot for publication bias if such assessment was appropriate and performed. Reviewers should include the results of assessment of risk of publication bias, including the results of statistical tests for publication bias, if such tests were used.

Even if meta-analysis is performed, a narrative summary should be included to supplement the technical details provided on the process and results of meta-analysis and to provide synthesis of data not captured in statistical meta-analysis.

If meta-analysis is not performed, a narrative summary should be included. The narrative summary should provide an overall summary of the findings of the included studies and their biases, strengths and limitations. The essence of narrative summary is that the results are summarized in words and in tables without any statistical meta-analysis. Textual commentaries and tables are used in order to summarize the results from the included studies and to provide context information for these results, thus facilitating understanding of the summarized results.

3.4.9 Discussion

The aim of this section is to briefly summarize the main findings and then focus on the discussion of these results. Results should be discussed, compared and contrasted with what was already known from other sources, other than the review, usually at a minimum the literature mentioned in the background section, however, additional external literature may be discussed here in order to facilitate the understanding and positioning of the review results in a broader research and practice context. The applicability and generalizability of the review results should be discussed. The significance of the results should be discussed for individual studies and for meta-analyses. It is not enough to discuss the statistical significance of the results; the practical/clinical significance of the results should be discussed regardless of the statistical significance of the results. The minimum and maximum values for the scales of measurement or measurement instruments should be discussed and the values that are considered to represent the minimum important change from a clinical/practical point of view.

This section should provide a presentation of the limitations of included studies and the limitations of the review process. Limitations of each included study (limitations in the design and conduct of the research, including risk of bias) should be discussed. Also, the limitations of entire set of included studies should be discussed in terms of common limitations (including risk of bias). All limitations, issues and problems noted in the review process related to the search, selection of study, critical appraisal, data extraction, and data synthesis, should be discussed. The impact of the limitations of the studies and of the review process on the applicability and generalizability of the results should be considered.

3.4.10 Conclusions and recommendations

This section should include the overall conclusions of the review. The conclusions should provide direct answers to the review objectives/questions. These conclusions should be based only on the results of the review and directly inferred from the review results.

Recommendations for practice

This sub-section of Conclusions section should include the recommendations for practice inferred from the results of the review and inferred also based on the discussion of the generalizability of the results and the potential factors that may affect the applicability of results. Recommendations should be assigned a JBI Grade of Recommendation. Refer for the editorial by Munn 2015 for further discussion regarding the appropriateness of making recommendations in systematic reviews.

Recommendations for research

This sub-section of Conclusions should include the recommendations for future research inferred from the results of the review, specifically, inferred from the limitations, issues and problems noted in the review process related to the search, selection of study, critical appraisal, data extraction, and data synthesis.

3.4.11 Conflicts and acknowledgements

Details of requirements in these sections are described in Section 1.6. of this Manual.

Conflicts of interest

A statement which either declares the absence of any conflicts of interest or which describes a specified or potential conflict of interest should be made by the reviewers in this section.

Funding

Authors should provide details regarding any sources of funding for the review project. The role of all funders in the review process, if any, should be explicitly described. If the review is funded, then any potential conflicts of interest or intellectual bias of the funders should be specified in the review.

Acknowledgements

Any acknowledgements should be made in this section e.g. sources of external funding or the contribution of colleagues or institutions. It should also be noted if the systematic review is to count towards a degree award.

3.4.12 Review Appendices

There are several required appendices for a JBI review:

Appendix 1: Search strategy

- A detailed and complete search strategy for all of the major databases and other sites and sources searched must be appended. Major databases that were searched must be identified, including the search platform used where necessary. All search filters with logic employed should be displayed, including the number of records returned.

Appendix 2: Data extraction instrument

- The data extraction instrument used must be appended i.e JBI SUMARI Data Extraction Form.

Appendix 3: List of excluded studies

- Studies excluded following examination of the full-text should be listed along with their reason for exclusion at that stage (i.e. a mismatch with the inclusion criteria). This may be as a separate appendix or itemized in some fashion within the one appendix with those studies excluded at the critical appraisal stage. Reasons for exclusion following appraisal should be provided for each study (these reasons should relate to the methodological quality of the study, not study eligibility).

Appendix 4: Table of included study characteristics

- A table of included studies is required to provide quick reference to important details extracted from of the studies included in the review.

3.5 Chapter References

- Aromataris E. Ins and outcomes. JBI Database System Rev Implement Rep. 2015; 13(4): 1-2.
- Aromataris E, Rittano D. Constructing a search strategy and searching for evidence. A guide to the literature search for a systematic review. Am J Nurs. 2014; 114(5): 49-56.
- Begg CB, Mazumdar M. Operating Characteristics of a Rank Correlation Test for Publication Bias. Biometrics. 1994; 50(4): 1088-1101.
- Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010; 1: 97–111.
- Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Wiley, 2009.
- Brignardello-Petersen R, Ioannidis JPA, Tomlinson G, Guyatt G. Surprising results of randomized trials. In Guyatt G, Rennie D, Meade MO, Cook DJ (editors). Users' Guide to the medical literature. A manual for evidence-based clinical practice. 3rd edition. New York: McGraw-Hill, 2015.
- Cooper H, Hedges LV. Potentials and limitations of research synthesis. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation, 1994.
- Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Chichester (UK): John Wiley & Sons, 2008.
- Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997; 315:629–34.
- Fleiss JL. Measures of effect size for categorical data. In: Harris Cooper Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation, 1994.
- Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Statist. Med. 2006; 25:3443–3457.
- Hedges LV. Meta-analysis. J Educ Behav Stat 1992; 17: 279– 96.
- Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I² index? Psychol Methods. 2006; 11(2): 193-206.
- Jin ZC, Zhou XH, He J. Statistical methods for dealing with publication bias in meta-analysis. Stat Med. 2015; 34(2): 343-60.
- Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. Bmj. 2006; 333 (7568): 597-600.
- Munn Z, Tufanaru C, Aromataris E. JBI's systematic reviews: data extraction and synthesis. Am J Nurs. 2014; 114(7):49-54.
- Munn Z. Implications for Practice: should recommendations be recommended in systematic reviews? JBI database of systematic reviews and implementation reports. 2015 Jan 1;13(7):1-3.
- Murad MH, Montori VM, Ioannidis JPA, et al. Fixed-effects and random-effects models. In: Guyatt G, Rennie D, Meade MO, Cook DJ, editors. Users' guide to the medical literature.
- A manual for evidence-based clinical practice. 3rd ed. New York: McGraw-Hill, 2015.
- Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 1999; 18: 321–59.
- Pettiti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine. 2nd ed. New York: Oxford University Press, 2000.
- Porritt K, Gomersall J, Lockwood C. JBI's Systematic Reviews: Study selection and critical appraisal. Am J Nurs. 2014; 114(6):47-52.
- Schünemann H, Broek J, Guyatt G, Oxman A (editors). GRADE Handbook. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. Updated October 2013.
- Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin Company, 2002.
- Stern C, Jordan Z, McArthur A. Developing the review question and inclusion criteria. Am J Nurs. 2014; 114(4): 53-6.
- Sterne JA, Sutton AJ, Ioannidis JP et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. Bmj. 2011; 343: d4002.

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical research. New York: Wiley, 2000.

Trohler U. The 18th century British Origins of a critical approach. Edinburgh: Royal College of Physicians, 2000.

Tufanaru C. Surrogate outcomes. JBI Database System Rev Implement Rep. 2016; 14(11): 1-2.

Tufanaru C, Munn Z, Stephenson M, Aromataris E. Fixed or random effects meta-analysis? Common methodological issues in systematic reviews of effectiveness. Int J Evid Based Healthc. 2015; 13(3):196-207.

Appendix 3.1: JBI Critical appraisal checklist for randomized controlled trials

JBI Critical Appraisal Checklist for Randomized Controlled Trials

Reviewer_____Date_____

Author _____Year_____Record Number _____

	Y es	No	Uncl ear	NA
1. Was true randomization used for assignment of participants to treatment groups?				
2. Was allocation to treatment groups concealed?				
3. Were treatment groups similar at the baseline?				
4. Were participants blind to treatment assignment?				
5. Were those delivering treatment blind to treatment assignment?				
6. Were outcomes assessors blind to treatment assignment?				
7. Were treatment groups treated identically other than the intervention of interest?				
8. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analyzed?				
9. Were participants analyzed in the groups to which they were randomized?				
10. Were outcomes measured in the same way for treatment groups?				
11. Were outcomes measured in a reliable way?				
12. Was appropriate statistical analysis used?				
13. Was the trial design appropriate, and any deviations from the standard RCT design (individual randomization, parallel groups) accounted for in the conduct and analysis of the trial?				

Overall appraisal: Include Exclude Seek further info

Comments (Including reason for exclusion)

Appendix 3.2: Discussion of JBI appraisal criteria for randomized controlled trials

Critical Appraisal Tool for RCTs (individual participants in parallel groups)

Answers: Yes, No, Unclear or Not Applicable

1. 1. Was true randomization used for assignment of participants to treatment groups?

The differences between participants included in compared groups constitutes a threat to the internal validity of a study exploring causal relationships. If participants are not allocated to treatment and control groups by random assignment there is a risk that the allocation is influenced by the known characteristics of the participants and these differences between the groups may distort the comparability of the groups. A true random assignment of participants to the groups means that a procedure is used that allocates the participants to groups purely based on chance, not influenced by the known characteristics of the participants. Check the details about the randomization procedure used for allocation of the participants to study groups. Was a true chance (random) procedure used? For example, was a list of random numbers used? Was a computer-generated list of random numbers used?

1. 2. Was allocation to groups concealed?

If those allocating participants to the compared groups are aware of which group is next in the allocation process, that is, treatment or control, there is a risk that they may deliberately and purposefully intervene in the allocation of patients by preferentially allocating patients to the treatment group or to the control group and therefore this may distort the implementation of allocation process indicated by the randomization and therefore the results of the study may be distorted. Concealment of allocation (allocation concealment) refers to procedures that prevent those allocating patients from knowing before allocation which treatment or control is next in the allocation process. Check the details about the procedure used for allocation concealment. Was an appropriate allocation concealment procedure used? For example, was central randomization used? Were sequentially numbered, opaque and sealed envelopes used? Were coded drug packs used?

1. 3. Were treatment groups similar at the baseline?

The differences between participants included in compared groups constitute a threat to the internal validity of a study exploring causal relationships. If there are differences between participants included in compared groups there is a risk of selection bias. If there are differences between participants included in the compared groups maybe the 'effect' cannot be attributed to the potential 'cause' (the examined intervention or treatment), as maybe it is plausible that the 'effect' may be explained by the differences between participants, that is, by selection bias. Check the characteristics reported for participants. Are the participants from the compared groups similar with regards to the characteristics that may explain the effect even in the absence of the 'cause', for example, age, severity of the disease, stage of the disease, co-existing conditions and so on? Check the proportions of participants with specific relevant characteristics in the compared groups. Check the means of relevant measurements in the compared groups (pain scores; anxiety scores; etc.). *[Note: Do NOT only consider the P-value for the statistical testing of the differences between groups with regards to the baseline characteristics.]*

1. 4. Were participants blind to treatment assignment?

If participants are aware of their allocation to the treatment group or to the control group there is the risk that they may behave differently and respond or react differently to the intervention of interest or to the control intervention respectively compared to the situations when they are not aware of treatment allocation and therefore the results of the study may be distorted. Blinding of participants is used in order to minimize this risk. Blinding of the participants refers to procedures that prevent participants from knowing which group they are allocated. If blinding of participants is used, participants are not aware if they are in the group receiving the treatment of interest or if they are in any other group receiving the control interventions. Check the details reported in the article about the blinding of participants with regards to treatment assignment. Was an appropriate blinding procedure used? For example, were identical capsules or syringes used? Were identical devices used? Be aware of different terms used, blinding is sometimes also called masking.

1. 5. Were those delivering treatment blind to treatment assignment?

If those delivering treatment are aware of participants' allocation to the treatment group or to the control group there is the risk that they may behave differently with the participants from the treatment group and the participants from the control group, or that they may treat them differently, compared to the situations when they are not aware of treatment allocation and this may influence the implementation of the compared treatments and the results of the study may be distorted. Blinding of those delivering treatment is used in order to minimize this risk. Blinding of those delivering treatment refers to procedures that prevent those delivering treatment from knowing which group they are treating, that is those delivering treatment are not aware if they are treating the group receiving the treatment of interest or if they are treating any other group receiving the control interventions. Check the details reported in the article about the blinding of those delivering treatment with regards to treatment assignment. Is there any information in the article about those delivering the treatment? Were those delivering the treatment unaware of the assignments of participants to the compared groups?

1. 6. Were outcomes assessors blind to treatment assignment?

If those assessing the outcomes are aware of participants' allocation to the treatment group or to the control group there is the risk that they may behave differently with the participants from the treatment group and the participants from the control group compared to the situations when they are not aware of treatment allocation and therefore there is the risk that the measurement of the outcomes may be distorted and the results of the study may be distorted. Blinding of outcomes assessors is used in order to minimize this risk. Check the details reported in the article about the blinding of outcomes assessors with regards to treatment assignment. Is there any information in the article about outcomes assessors? Were those assessing the treatment's effects on outcomes unaware of the assignments of participants to the compared groups?

1. 7. Were treatment groups treated identically other than the intervention of interest?

In order to attribute the 'effect' to the 'cause' (the treatment or intervention of interest), assuming that there is no selection bias, there should be no other difference between the groups in terms of treatment or care received, other than the manipulated 'cause' (the treatment or intervention controlled by the researchers). If there are other exposures or treatments occurring at the same time with the 'cause' (the treatment or intervention of interest), other than the 'cause', then potentially the 'effect' cannot be attributed to the examined 'cause' (the investigated treatment), as it is plausible that the 'effect' may be explained by other exposures or treatments occurring at the same time with the 'cause' (the treatment of interest). Check the reported exposures or interventions received by the compared groups. Are there other exposures or treatments occurring at the same time with the 'cause'? Is it plausible that the 'effect' may be explained by other exposures or treatments occurring at the same time with the 'cause'? Is it clear that there is no other difference between the groups in terms of treatment or care received, other than the treatment or intervention of interest?

1.8. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analyzed?

For this question, follow up refers to the time period from the moment of random allocation (random assignment or randomization) to compared groups to the end time of the trial. This critical appraisal question asks if there is complete knowledge (measurements, observations etc.) for the entire duration of the trial as previously defined (that is, from the moment of random allocation to the end time of the trial), for all randomly allocated participants. If there is incomplete follow up, that is incomplete knowledge about all randomly allocated participants, this is known in the methodological literature as the post-assignment attrition. As RCTs are not perfect, there is almost always post-assignment attrition, and the focus of this question is on the appropriate exploration of post-assignment attrition (description of loss to follow up, description of the reasons for loss to follow up, the estimation of the impact of loss to follow up on the effects etc.). If there are differences with regards to the loss to follow up between the compared groups in an RCT, these differences represent a threat to the internal validity of a randomized experimental study exploring causal effects, as these differences may provide a plausible alternative explanation for the observed 'effect' even in the absence of the 'cause' (the treatment or intervention of interest). When appraising an RCT, check if there were differences with regards to the loss to follow up between the compared groups. If follow up was incomplete (that is, there is incomplete information on all participants), examine the reported details about the strategies used in order to address incomplete follow up, such as descriptions of loss to follow up (absolute numbers; proportions; reasons for loss to follow up) and impact analyses (the analyses of the impact of loss to follow up on results). Was there a description of the incomplete follow up (number of participants and the specific reasons for loss to follow up)? It is important to note that with regards to loss to follow up, it is not enough to know the number of participants and the proportions of participants with incomplete data; the reasons for loss to follow up are essential in the analysis of risk of bias; even if the numbers and proportions of participants with incomplete data are similar or identical in compared groups, if the patterns of reasons for loss to follow up are different (for example, side effects caused by the intervention of interest, lost contact etc.), these may impose a risk of bias if not appropriately explored and considered in the analysis. If there are differences between groups with regards to the loss to follow up (numbers/proportions and reasons), was there an analysis of patterns of loss to follow up? If there are differences between the groups with regards to the loss to follow up, was there an analysis of the impact of the loss to follow up on the results? [Note: Question 8 is NOT about intention-to-treat (ITT) analysis; question 9 is about ITT analysis.]

1.9. Were participants analyzed in the groups to which they were randomized?

This question is about the intention-to-treat (ITT) analysis. There are different statistical analysis strategies available for the analysis of data from randomized controlled trials, such as intention-to-treat analysis (known also as intent to treat; abbreviated, ITT), per-protocol analysis, and as-treated analysis. In the ITT analysis the participants are analyzed in the groups to which they were randomized, regardless of whether they actually participated or not in those groups for the entire duration of the trial, received the experimental intervention or control intervention as planned or whether they were compliant or not with the planned experimental intervention or control intervention. The ITT analysis compares the outcomes for participants from the initial groups created by the initial random allocation of participants to those groups. Check if ITT was reported; check the details of the ITT. Were participants analyzed in the groups to which they were initially randomized, regardless of whether they actually participated in those groups, and regardless of whether they actually received the planned interventions? [Note: The ITT analysis is a type of statistical analysis recommended in the Consolidated Standards of Reporting Trials (CONSORT) statement on best practices in trials reporting, and it is considered a marker of good methodological quality of the analysis of results of a randomized trial. The ITT is estimating the effect of offering the intervention, that is, the effect of instructing the participants to use or take the intervention; the ITT is not estimating the effect of actually receiving the intervention of interest.]

10. Were outcomes measured in the same way for treatment groups?

If the outcome (the 'effect') is not measured in the same way in the compared groups there is a threat to the internal validity of a study exploring a causal relationship as the differences in outcome measurements may be confused with an effect of the treatment (the 'cause'). Check if the outcomes were measured in the same way. Same instrument or scale used? Same measurement timing? Same measurement procedures and instructions?

11. Were outcomes measured in a reliable way?

Unreliability of outcome measurements is one threat that weakens the validity of inferences about the statistical relationship between the 'cause' and the 'effect' estimated in a study exploring causal effects. Unreliability of outcome measurements is one of the different plausible explanations for errors of statistical inference with regards to the existence and the magnitude of the effect determined by the treatment ('cause'). Check the details about the reliability of measurement such as the number of raters, training of raters, the intra-rater reliability, and the inter-raters reliability within the study (not as reported in external sources). This question is about the reliability of the measurement performed in the study, it is not about the validity of the measurement instruments/scales used in the study. *[Note: Two other important threats that weaken the validity of inferences about the statistical relationship between the 'cause' and the 'effect' are low statistical power and the violation of the assumptions of statistical tests. These other two threats are explored within Question 12].*

12. Was appropriate statistical analysis used?

Inappropriate statistical analysis may cause errors of statistical inference with regards to the existence and the magnitude of the effect determined by the treatment ('cause'). Low statistical power and the violation of the assumptions of statistical tests are two important threats that weaken the validity of inferences about the statistical relationship between the 'cause' and the 'effect'. Check the following aspects: if the assumptions of statistical tests were respected; if appropriate statistical power analysis was performed; if appropriate effect sizes were used; if appropriate statistical procedures or methods were used given the number and type of dependent and independent variables, the number of study groups, the nature of the relationship between the groups (independent or dependent groups), and the objectives of statistical analysis (association between variables; prediction; survival analysis etc.).

13. Was the trial design appropriate for the topic, and any deviations from the standard RCT design accounted for in the conduct and analysis?

Certain RCT designs, such as the crossover RCT, should only be conducted when appropriate. Alternative designs may also present additional risks of bias if not accounted for in the design and analysis.

Crossover trials should only be conducted in people with a chronic, stable condition, where the intervention produces a short term effect (i.e. relief in symptoms). Crossover trials should ensure there is an appropriate period of washout between treatments.

Cluster RCTs randomize groups of individuals, forming 'clusters.' When we are assessing outcomes on an individual level in cluster trials, there are unit-of-analysis issues, as individuals within a cluster are correlated. This should be taken into account by the study authors when conducting analysis, and ideally authors will report the intra-cluster correlation coefficient.

Stepped-wedge RCTs may be appropriate when it is expected the intervention will do more good than harm, or due to logistical, practical or financial considerations in the roll out of a new treatment /intervention. Data analysis in these trials should be conducted appropriately, taking into account the effects of time.

Appendix 3.3: JBI Critical appraisal Checklist for Quasi-Experimental Studies (non-randomized experimental studies)

JBI Critical Appraisal Checklist for Quasi-Experimental Studies
(non-randomized experimental studies)

Reviewer _____ Date _____

-

Author _____ Year _____ Record Number _____

	Yes	No	Unclear	Not applicable
1. Is it clear in the study what is the 'cause' and what is the 'effect' (i.e. there is no confusion about which variable comes first)?				
2. Were the participants included in any comparisons similar?				
3. Were the participants included in any comparisons receiving similar treatment/care, other than the exposure or intervention of interest?				
4. Was there a control group?				
5. Were there multiple measurements of the outcome both pre and post the intervention/exposure?				
6. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analyzed?				
7. Were the outcomes of participants included in any comparisons measured in the same way?				
8. Were outcomes measured in a reliable way?				
9. Was appropriate statistical analysis used?				

Overall appraisal: Include Exclude Seek further info

Comments (Including reason for exclusion)

Appendix 3.4: Discussion of JBI appraisal criteria for Quasi-Experimental Studies (non-randomized experimental studies)

Explanation for the critical appraisal tool for Quasi-Experimental Studies (experimental studies without random allocation)

Critical Appraisal Tool for Quasi-Experimental Studies (experimental studies without random allocation)

Answers: Yes, No, Unclear or Not Applicable

1. Is it clear in the study what is the 'cause' and what is the 'effect' (i.e. there is no confusion about which variable comes first)?

Ambiguity with regards to the temporal relationship of variables constitutes a threat to the internal validity of a study exploring causal relationships. The 'cause' (the independent variable, that is, the treatment or intervention of interest) should occur in time before the explored 'effect' (the dependent variable, which is the effect or outcome of interest). Check if it is clear which variable is manipulated as a potential cause. Check if it is clear which variable is measured as the effect of the potential cause. Is it clear that the 'cause' was manipulated before the occurrence of the 'effect'?

2. Were the participants included in any comparisons similar?

The differences between participants included in compared groups constitute a threat to the internal validity of a study exploring causal relationships. If there are differences between participants included in compared groups there is a risk of selection bias. If there are differences between participants included in the compared groups maybe the 'effect' cannot be attributed to the potential 'cause', as maybe it is plausible that the 'effect' may be explained by the differences between participants, that is, by selection bias. Check the characteristics reported for participants. Are the participants from the compared groups similar with regards to the characteristics that may explain the effect even in the absence of the 'cause', for example, age, severity of the disease, stage of the disease, co-existing conditions and so on? *[NOTE: In one single group pre-test/post-test studies where the patients are the same (the same one group) in any pre-post comparisons, the answer to this question should be 'yes.']*

3. Were the participants included in any comparisons receiving similar treatment/care, other than the exposure or intervention of interest?

In order to attribute the 'effect' to the 'cause' (the exposure or intervention of interest), assuming that there is no selection bias, there should be no other difference between the groups in terms of treatments or care received, other than the manipulated 'cause' (the intervention of interest). If there are other exposures or treatments occurring in the same time with the 'cause', other than the intervention of interest, then potentially the 'effect' cannot be attributed to the intervention of interest, as it is plausible that the 'effect' may be explained by other exposures or treatments, other than the intervention of interest, occurring in the same time with the intervention of interest. Check the reported exposures or interventions received by the compared groups. Are there other exposures or treatments occurring in the same time with the intervention of interest? Is it plausible that the 'effect' may be explained by other exposures or treatments occurring in the same time with the intervention of interest?

4. Was there a control group?

Control groups offer the conditions to explore what would have happened with groups exposed to other different treatments, other than to the potential 'cause' (the intervention of interest). The comparison of the treated group (the group exposed to the examined 'cause', that is, the group receiving the intervention of interest) with such other groups strengthens the examination of the causal plausibility. The validity of causal inferences is strengthened in studies with at least one independent control group compared to studies without an independent control group. Check if there are independent, separate groups, used as control groups in the study. *[Note: The control group should be an independent, separate control group, not the pre-test group in a single group pre-test post-test design.]*

5. Were there multiple measurements of the outcome both pre and post the intervention /exposure?

In order to show that there is a change in the outcome (the 'effect') as a result of the intervention/treatment (the 'cause') it is necessary to compare the results of measurement before and after the intervention/treatment. If there is no measurement before the treatment and only measurement after the treatment is available it is not known if there is a change after the treatment compared to before the treatment. If multiple measurements are collected before the intervention/treatment is implemented then it is possible to explore the plausibility of alternative explanations other than the proposed 'cause' (the intervention of interest) for the observed 'effect', such as the naturally occurring changes in the absence of the 'cause', and changes of high (or low) scores towards less extreme values even in the absence of the 'cause' (sometimes called regression to the mean). If multiple measurements are collected after the intervention/treatment is implemented it is possible to explore the changes of the 'effect' in time in each group and to compare these changes across the groups. Check if measurements were collected before the intervention of interest was implemented. Were there multiple pre-test measurements? Check if measurements were collected after the intervention of interest was implemented. Were there multiple post-test measurements?

6. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analyzed?

If there are differences with regards to the loss to follow up between the compared groups these differences represent a threat to the internal validity of a study exploring causal effects as these differences may provide a plausible alternative explanation for the observed 'effect' even in the absence of the 'cause' (the treatment or exposure of interest). Check if there were differences with regards to the loss to follow up between the compared groups. If follow up was incomplete (that is, there is incomplete information on all participants), examine the reported details about the strategies used in order to address incomplete follow up, such as descriptions of loss to follow up (absolute numbers; proportions; reasons for loss to follow up; patterns of loss to follow up) and impact analyses (the analyses of the impact of loss to follow up on results). Was there a description of the incomplete follow up (number of participants and the specific reasons for loss to follow up)? If there are differences between groups with regards to the loss to follow up, was there an analysis of patterns of loss to follow up? If there are differences between the groups with regards to the loss to follow up, was there an analysis of the impact of the loss to follow up on the results?

7. Were the outcomes of participants included in any comparisons measured in the same way?

If the outcome (the 'effect') is not measured in the same way in the compared groups there is a threat to the internal validity of a study exploring a causal relationship as the differences in outcome measurements may be confused with an effect of the treatment or intervention of interest (the 'cause'). Check if the outcomes were measured in the same way. Same instrument or scale used? Same measurement timing? Same measurement procedures and instructions?

8. Were outcomes measured in a reliable way?

Unreliability of outcome measurements is one threat that weakens the validity of inferences about the statistical relationship between the 'cause' and the 'effect' estimated in a study exploring causal effects. Unreliability of outcome measurements is one of different plausible explanations for errors of statistical inference with regards to the existence and the magnitude of the effect determined by the treatment ('cause'). Check the details about the reliability of measurement such as the number of raters, training of raters, the intra-rater reliability, and the inter-raters reliability within the study (not to external sources). This question is about the reliability of the measurement performed in the study, it is not about the validity of the measurement instruments/scales used in the study. *[Note: Two other important threats that weaken the validity of inferences about the statistical relationship between the 'cause' and the 'effect' are low statistical power and the violation of the assumptions of statistical tests. These other threats are not explored within Question 8, these are explored within Question 9.]*

9. Was appropriate statistical analysis used?

Inappropriate statistical analysis may cause errors of statistical inference with regards to the existence and the magnitude of the effect determined by the treatment ('cause'). Low statistical power and the violation of the assumptions of statistical tests are two important threats that weakens the validity of inferences about the statistical relationship between the 'cause' and the 'effect'. Check the following aspects: if the assumptions of statistical tests were respected; if appropriate statistical power analysis was performed; if appropriate effect sizes were used; if appropriate statistical procedures or methods were used given the number and type of dependent and independent variables, the number of study groups, the nature of the relationship between the groups (independent or dependent groups), and the objectives of statistical analysis (association between variables; prediction; survival analysis etc.).