3

Defining the criteria for including studies and how they will be grouped for the synthesis

Joanne E McKenzie, Sue E Brennan, Rebecca E Ryan, Hilary J Thomson, Renea V Johnston, James Thomas

KEY POINTS

- The scope of a review is defined by the types of population (participants), types of interventions (and comparisons), and the types of outcomes that are of interest. The acronym PICO (population, interventions, comparators and outcomes) helps to serve as a reminder of these.
- The population, intervention and comparison components of the question, with the additional specification of types of study that will be included, form the basis of the pre-specified eligibility criteria for the review. It is rare to use outcomes as eligibility criteria: studies should be included irrespective of whether they *report* outcome data, but may legitimately be excluded if they do not *measure* outcomes of interest, or if they explicitly aim to prevent a particular outcome.
- Cochrane Reviews should include all outcomes that are likely to be meaningful and not include trivial outcomes. Critical and important outcomes should be limited in number and include adverse as well as beneficial outcomes.
- Review authors should plan at the protocol stage how the different populations, interventions, outcomes and study designs within the scope of the review will be grouped for analysis.

3.1 Introduction

One of the features that distinguishes a systematic review from a narrative review is that systematic review authors should pre-specify criteria for including and excluding studies in the review (eligibility criteria, see MECIR Box 3.2.a).

When developing the protocol, one of the first steps is to determine the elements of the review question (including the population, intervention(s), comparator(s) and

This chapter should be cited as: McKenzie JE, Brennan SE, Ryan RE, Thomson HJ, Johnston RV, Thomas J. Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd Edition. Chichester (UK): John Wiley & Sons, 2019: 33–66.

outcomes, or PICO elements) and how the intervention, in the specified population, produces the expected outcomes (see Chapter 2, Section 2.5.1 and Chapter 17, Section 17.2.1). Eligibility criteria are based on the PICO elements of the review question plus a specification of the types of studies that have addressed these questions. The population, interventions and comparators in the review question usually translate directly into eligibility criteria for the review, though this is not always a straightforward process and requires a thoughtful approach, as this chapter shows. Outcomes usually are not part of the criteria for including studies, and a Cochrane Review would typically seek all sufficiently rigorous studies (most commonly randomized trials) of a particular comparison of interventions in a particular population of participants, irrespective of the outcomes measured or reported. It should be noted that some reviews do legitimately restrict eligibility to specific outcomes. For example, the same intervention may be studied in the same population for different purposes; or a review may specifically address the adverse effects of an intervention used for several conditions (see Chapter 19).

Eligibility criteria do not exist in isolation, but should be specified with the synthesis of the studies they describe in mind. This will involve making plans for how to group variants of the PICO elements for synthesis. This chapter describes the processes by which the structure of the synthesis can be mapped out at the beginning of the review, and the interplay between the review question, considerations for the analysis and their operationalization in terms of eligibility criteria. Decisions about which studies to include (and exclude), and how they will be combined in the review's synthesis, should be documented and justified in the review protocol.

A distinction between three different stages in the review at which the PICO construct might be used is helpful for understanding the decisions that need to be made. In Chapter 2 (Section 2.3) we introduced the ideas of a **review PICO** (on which eligibility of studies is based), the **PICO for each synthesis** (defining the question that each specific synthesis aims to answer) and the **PICO of the included studies** (what was actually investigated in the included studies). In this chapter, we focus on the **review PICO** and the **PICO for each synthesis** as a basis for specifying which studies should be included in the review and planning its syntheses. These PICOs should relate clearly and directly to the questions or hypotheses that are posed when the review is formulated (see Chapter 2) and will involve specifying the population in question, and a set of comparisons between the intervention groups.

An integral part of the process of setting up the review is to specify which characteristics of the interventions (e.g. individual compounds of a drug), populations (e.g. acute and chronic conditions), outcomes (e.g. different depression measurement scales) and study designs, will be grouped together. Such decisions should be made independent of knowing which studies will be included and the methods of synthesis that will be used (e.g. meta-analysis). There may be a need to modify the comparisons and even add new ones at the review stage in light of the data that are collected. For example, important variations in the intervention may be discovered only after data are collected, or modifying the comparison may facilitate the possibility of synthesis when only one or few studies meet the comparison PICO. Planning for the latter scenario at the protocol stage may lead to less post-hoc decision making (Chapter 2, Section 2.5.3) and, of course, any changes made during the conduct of the review should be recorded and documented in the final report.

3.2 Articulating the review and comparison PICO

3.2.1 Defining types of participants: which people and populations?

The criteria for considering types of people included in studies in a review should be sufficiently broad to encompass the likely diversity of studies and the likely scenarios in which the interventions will be used, but sufficiently narrow to ensure that a meaningful answer can be obtained when studies are considered together; they should be specified in advance (see MECIR Box 3.2.a). As discussed in Chapter 2 (Section 2.3.1), the degree of breadth will vary, depending on the question being asked and the analytical approach to be employed. A range of evidence may inform the choice of population characteristics to examine, including theoretical considerations, evidence from other interventions that have a similar mechanism of action, and in vitro or animal studies. Consideration should be given to whether the population characteristic is at the level of the participant (e.g. age, severity of disease) or the study (e.g. care setting, geographical

MECIR Box 3.2.a Relevant expectations for conduct of intervention reviews

C5: Predefining unambiguous criteria for participants (**Mandatory**)

Define in advance the eligibility criteria for participants in the studies.	Predefined, unambiguous eligibility criteria are a fundamental prerequisite for a systematic review. The criteria for considering types of people included in studies in a review should be sufficiently broad to encompass the likely diversity of studies, but sufficiently narrow to ensure that a meaningful answer can be obtained when studies are considered in aggregate. Considerations when specifying participants include setting, diagnosis or definition of condition and demographic factors. Any restrictions to study populations must be based on a sound rationale, since it is important that Cochrane Reviews are widely relevant.

C6: Predefining a strategy for studies with a subset of eligible participants (**Highly desirable**)

Define in advance how studies that include only a subset of relevant participants will be addressed.	Sometimes a study includes some 'eligible' participants and some 'ineligible' participants, for example when an age cut-off is used in the review's eligibility criteria. If data from the eligible participants cannot be retrieved, a mechanism for dealing with this situation should be pre-specified.

location), since this has implications for grouping studies and for the method of synthesis (Chapter 10, Section 10.11.5). It is often helpful to consider the types of people that are of interest in three steps.

First, the **diseases or conditions of interest should be defined** using explicit criteria for establishing their presence (or absence). Criteria that will force the unnecessary exclusion of studies should be avoided. For example, diagnostic criteria that were developed more recently – which may be viewed as the current gold standard for diagnosing the condition of interest – will not have been used in earlier studies. Expensive or recent diagnostic tests may not be available in many countries or settings, and time-consuming tests may not be practical in routine healthcare settings.

Second, the **broad population and setting of interest should be defined**. This involves deciding whether a specific population group is within scope, determined by factors such as age, sex, race, educational status or the presence of a particular condition such as angina or shortness of breath. Interest may focus on a particular setting such as a community, hospital, nursing home, chronic care institution, or outpatient setting. Box 3.2.a outlines some factors to consider when developing population criteria.

Whichever criteria are used for defining the population and setting of interest, it is common to encounter studies that only partially overlap with the review's population. For example, in a review focusing on children, a cut-point of less than 16 years might be desirable, but studies may be identified with participants aged from 12 to 18. Unless the study reports separate data from the eligible section of the population (in which case data from the eligible participants can be included in the review), review authors will need a strategy for dealing with these studies (see MECIR Box 3.2.a). This will involve balancing concerns about reduced applicability by including participants who do not meet the eligibility criteria, against the loss of data when studies are excluded. Arbitrary rules (such as including a study if more than 80% of the participants are under 16) will not be practical if detailed information is not available from the study. A less stringent rule, such as 'the majority of participants are under 16' may be sufficient. Although there is a risk of review authors' biases affecting post-hoc inclusion decisions (which is why many authors endeavour to pre-specify these rules), this may be outweighed by a common-sense strategy in which eligibility decisions keep faith with the objectives of the review rather than with arbitrary rules. Difficult decisions should be documented in the review, checked with the advisory group (if available, see Chapter 1), and

Box 3.2.a Factors to consider when developing criteria for 'Types of participants'

- How is the disease/condition defined?
- What are the most important characteristics that describe these people (participants)?
- Are there any relevant demographic factors (e.g. age, sex, ethnicity)?
- What is the setting (e.g. hospital, community, etc)?
- Who should make the diagnosis?
- Are there other types of people who should be excluded from the review (because they are likely to react to the intervention in a different way)?
- How will studies involving only a subset of relevant participants be handled?

MECIR Box 3.2.b Relevant expectations for conduct of intervention reviews

C13: Changing eligibility criteria (**Mandatory**)

Justify any changes to eligibility criteria or outcomes studied. In particular, post-hoc decisions about inclusion or exclusion of studies should keep faith with the objectives of the review rather than with arbitrary rules.	Following pre-specified eligibility criteria is a fundamental attribute of a systematic review. However, unanticipated issues may arise. Review authors should make sensible post-hoc decisions about exclusion of studies, and these should be documented in the review, possibly accompanied by sensitivity analyses. Changes to the protocol must not be made on the basis of the findings of the studies or the synthesis, as this can introduce bias.

sensitivity analyses can assess the impact of these decisions on the review's findings (see Chapter 10, Section 10.14 and MECIR Box 3.2.b).

Third, there should be consideration of whether there are **population characteristics that might be expected to modify the size of the intervention effects** (e.g. different severities of heart failure). Identifying subpopulations may be important for implementation of the intervention. If relevant subpopulations are identified, two courses of action are possible: limiting the scope of the review to exclude certain subpopulations; or maintaining the breadth of the review and addressing subpopulations in the analysis.

Restricting the review with respect to specific population characteristics or settings should be based on a sound rationale. It is important that Cochrane Reviews are globally relevant, so the rationale for the exclusion of studies based on population characteristics should be justified. For example, focusing a review of the effectiveness of mammographic screening on women between 40 and 50 years old may be justified based on biological plausibility, previously published systematic reviews and existing controversy. On the other hand, focusing a review on a particular subgroup of people on the basis of their age, sex or ethnicity simply because of personal interests, when there is no underlying biologic or sociological justification for doing so, should be avoided, as these reviews will be less useful to decision makers and readers of the review.

Maintaining the breadth of the review may be best when it is uncertain whether there are important differences in effects among various subgroups of people, since this allows investigation of these differences (see Chapter 10, Section 10.11.5). Review authors may combine the results from different subpopulations in the same synthesis, examining whether a given subdivision explains variation (heterogeneity) among the intervention effects. Alternatively, the results may be synthesized in separate comparisons representing different subpopulations. Splitting by subpopulation risks there being too few studies to yield a useful synthesis (see Table 3.2.a and Chapter 2, Section 2.3.2). Consideration needs to be given to the subgroup analysis method,

Table 3.2.a Examples of population attributes and characteristics

Population attributes	Examples of population characteristics (and their subpopulations)	Examples of examination of population characteristics in Cochrane Reviews
Intended recipient of intervention	Patient, carer, healthcare provider (general practitioners, nurses, allied health professionals), health system, policy maker, community	In a review of e-learning programmes for health professionals, a subgroup analysis was planned to examine if the effects were modified by the *type of healthcare provider* (doctors, nurses or physiotherapists). The authors hypothesized that e-learning programmes for doctors would be more effective than for other health professionals, but did not provide a rationale (Vaona et al 2018).
Disease/condition (to be treated or prevented)	Type and severity of a condition	In a review of platelet-rich therapies for musculoskeletal soft tissue injuries, a subgroup analysis was undertaken to examine if the effects of platelet-rich therapies were modified by the *type of condition* (e.g. rotator cuff tear, anterior cruciate ligament reconstruction, chronic Achilles tendinopathy) (Moraes et al 2014). In planning a review of beta-blockers for heart failure, subgroup analyses were specified to examine if the effects of beta-blockers are modified by the *underlying cause of heart failure* (e.g. idiopathic dilated cardiomyopathy, ischaemic heart disease, valvular heart disease, hypertension) and the *severity of heart failure* ('reduced left ventricular ejection fraction (LVEF)' ≤ 40%, 'mid-range LVEF' > 40% and < 50%, 'preserved LVEF' ≥ 50%, mixed, not-specified). Studies have shown that patient characteristics and comorbidities differ by heart failure severity, and that therapies have been shown to reduce morbidity in 'reduced LVEF' patients, but the benefits in the other groups are uncertain (Safi et al 2017).
Participant characteristics	Age (neonate, child, adolescent, adult, older adult) Race/ethnicity Sex/gender PROGRESS-Plus equity characteristics (e.g. place of residence, socio-economic status, education) (O'Neill et al 2014)	In a review of newer-generation antidepressants for depressive disorders in children and adolescents, a subgroup analysis was undertaken to examine if the effects of the antidepressants were modified by *age*. The rationale was based on the findings of another review that suggested that children and adolescents may respond differently to antidepressants. The age groups were defined as 'children' (aged approximately 6 to 12 years), 'adolescents' (aged approximately 13 to 18 years), and 'children and adolescents' (when the study included both children and adolescents, and results could not be obtained separately by these subpopulations) (Hetrick et al 2012).
Setting	Setting of care (primary care, hospital, community) Rurality (urban, rural, remote) Socio-economic setting (low and middle-income countries, high-income countries) Hospital ward (e.g. intensive care unit, general medical ward, outpatient)	In a review of hip protectors for preventing hip fractures in older people, separate comparisons were specified based on *setting* (institutional care or community-dwelling) for the critical outcome of hip fracture (Santesso et al 2014).

particularly for population characteristics measured at the participant level (see Chapters 10 and 26, Fisher et al 2017). All subgroup analyses should ideally be planned a priori and stated as a secondary objective in the protocol, and not driven by the availability of data.

In practice, it may be difficult to assign included studies to defined subpopulations because of missing information about the population characteristic, variability in how the population characteristic is measured across studies (e.g. variation in the method used to define the severity of heart failure), or because the study does not wholly fall within (or report the results separately by) the defined subpopulation. The latter issue mainly applies for participant characteristics but can also arise for settings or geographic locations where these vary within studies. Review authors should consider planning for these scenarios (see example reviews Hetrick et al 2012, Safi et al 2017; Table 3.2.b, column 3).

3.2.2 Defining interventions and how they will be grouped

In some reviews, predefining the intervention (MECIR Box 3.2.c) may be straightforward. For example, in a review of the effect of a given anticoagulant on deep vein thrombosis, the intervention can be defined precisely. A more complicated definition might be required for a multi-component intervention composed of dietary advice, training and support groups to reduce rates of obesity in a given population.

The inherent complexity present when defining an intervention often comes to light when considering how it is thought to achieve its intended effect and whether the effect is likely to differ when variants of the intervention are used. In the first example, the anticoagulant warfarin is thought to reduce blood clots by blocking an enzyme that depends on vitamin K to generate clotting factors. In the second, the behavioural intervention is thought to increase individuals' self-efficacy in their ability to prepare healthy food. In both examples, we cannot assume that all forms of the intervention will work in the same way. When defining drug interventions, such as anticoagulants, factors such as the drug preparation, route of administration, dose, duration, and frequency should be considered. For multi-component interventions (such as interventions to reduce rates of obesity), the common or core features of the interventions must be defined, so that the review authors can clearly differentiate them from other interventions not included in the review.

In general, it is useful to consider **exactly what is delivered, who delivers it, how it is delivered, where it is delivered, when and how much is delivered, and whether the intervention can be adapted or tailored**, and to consider this for each type of intervention included in the review (see the TIDieR checklist (Hoffmann et al 2014)). As argued in Chapter 17, separating interventions into 'simple' and 'complex' is a false dichotomy; all interventions can be complex in some ways. The critical issue for review authors is to identify the most important factors to be considered in a specific review. Box 3.2.b outlines some factors to consider when developing broad criteria for the 'Types of interventions' (and comparisons).

Once interventions eligible for the review have been broadly defined, decisions should be made about how variants of the intervention will be handled in the synthesis. Differences in intervention characteristics across studies occur in all reviews. If these reflect minor differences in the form of the intervention used in practice (such as small differences in the duration or content of brief alcohol counselling interventions), then

Table 3.2.b A process for planning intervention groups for synthesis

Step	Considerations	Examples
1. Identify intervention characteristics that may modify the effect of the intervention.	Consider whether differences in interventions characteristics might modify the size of the intervention effect importantly. Content-specific research literature and expertise should inform this step. The TIDieR checklist – a tool for describing interventions – outlines the characteristics across which an intervention might differ (Hoffmann et al 2014). These include 'what' materials and procedures are used, 'who' provides the intervention, 'when and how much' intervention is delivered. The iCAT-SR tool provides equivalent guidance for complex interventions (Lewin et al 2017).	***Exercise interventions*** differ across multiple characteristics, which vary in importance depending on the review. In a review of exercise for osteoporosis, whether the exercise is weight-bearing or non-weight-bearing may be a key characteristic, since the mechanism by which exercise is thought to work is by placing stress or mechanical load on bones (Howe et al 2011). Different mechanisms apply in reviews of exercise for knee osteoarthritis (muscle strengthening), falls prevention (gait and balance), cognitive function (cardiovascular fitness). The differing mechanisms might suggest different ways of grouping interventions (e.g. by intensity, mode of delivery) according to potential modifiers of the intervention effects.
2a. Label and define intervention groups to be considered in the synthesis.	For each intervention group, provide a short label (e.g. supportive psychotherapy) and describe the core characteristics (criteria) that will be used to assign each intervention from an included study to a group. Groups are often defined by intervention content (especially the active components), such as materials, procedures or techniques (e.g. a specific drug, an information leaflet, a behaviour change technique). Other characteristics may also be used, although some are more commonly used to define subgroups (see Chapter 10, Section 10.11.5): the purpose or theoretical underpinning, mode of delivery, provider, dose or intensity, duration or timing of the intervention (Hoffmann et al 2014). In specifying groups: • focus on 'clinically' meaningful groups that will inform selection and implementation of an intervention in practice;	In a review of psychological therapies for coronary heart disease, a single group was specified for meta-analysis that included all types of therapy. Subgroups were defined to examine whether intervention effects were modified by intervention components (e.g. cognitive techniques, stress management) or mode of delivery (e.g. individual, group) (Richards et al 2017). In a review of psychological therapies for panic disorder (Pompoli et al 2016), eight types of therapy were specified: 1) psychoeducation; 2) supportive psychotherapy (with or without a psychoeducational component); 3) physiological therapies; 4) behaviour therapy; 5) cognitive therapy; 6) cognitive behaviour therapy (CBT); 7) 7. third-wave CBT; and

	• consider whether a system exists for defining interventions (see Step 3); • for hard-to-describe groups, provide brief examples of interventions in each group; and • pilot the criteria to ensure that groups are sufficiently distinct to enable categorization, but not so narrow that interventions are split into many groups, making synthesis impossible (see also Step 4). Logic models may help structure the synthesis (see Chapter 2, Section 2.4.1 and Chapter 17, Section 17.2.1).	8) psychodynamic therapies. Groups were defined by the theoretical basis of each therapy (e.g. CBT aims to modify maladaptive thoughts through cognitive restructuring) and the component techniques used.
2b. Define levels for groups based on dose or intensity.	For groups based on 'how much' of an intervention is used (e.g. dose or intensity), criteria are needed to quantify each group. This may be straightforward for easy-to-quantify characteristics, but more complex for characteristics that are hard to quantify (e.g. duration or intensity of rehabilitation or psychological therapy). The levels should be based on how the intervention is used in practice (e.g. cut-offs for low and high doses of a supplement based on recommended nutrient intake), or on a rationale for how the intervention might work.	In reviews of exercise, intensity may be defined by training time (session length, frequency, program duration), amount of work (e.g. repetitions), and effort/energy expenditure (exertion, heart rate) (Regnaux et al 2015). In a review of organized inpatient care for stroke, acute stroke units were categorized as 'intensive', 'semi-intensive' or 'non-intensive' based on whether the unit had continuous monitoring, high nurse staffing, and life support facilities (Stroke Unit Trialists Collaboration 2013).
3. Determine whether there is an existing system for grouping interventions. *Consider this step with step 2a.*	In some fields, intervention taxonomies and frameworks have been developed for labelling and describing interventions, and these can make it easier for those using a review to interpret and apply findings. Using an agreed system is preferable to developing new groupings. Existing systems should be assessed for relevance and usefulness. The most useful systems:	***Generic systems*** The *behaviour change technique* (BCT) *taxonomy* (Michie et al 2013) categorizes intervention elements such as goal setting, self-monitoring and social support. A protocol for a review of social media interventions used this taxonomy to describe interventions and examine different BCTs as potential effect modifiers (Welch et al 2018).

(*Continued*)

Table 3.2.b (Continued)

Step	Considerations	Examples
	• use terminology that is understood by those using or implementing the intervention; • are developed systematically and based on consensus, preferably with stakeholders including clinicians, patients, policy makers, and researchers; and • have been validated through successful use in a range of applications (ideally, including in systematic reviews). Systems for grouping interventions may be generic, widely applicable across clinical areas, or specific to a condition or intervention type. Some Cochrane Groups recommend specific taxonomies.	The *behaviour change wheel* has been used to group interventions (or components) by function (e.g. to educate, persuade, enable) (Michie et al 2011). This system was used to describe the components of dietary advice interventions (Desroches et al 2013). ***Specific systems*** Multiple reviews have used the consensus-based taxonomy developed by the Prevention of Falls Network Europe (ProFaNE) (e.g. Verheyden et al 2013, Kendrick et al 2014). The taxonomy specifies broad groups (e.g. exercise, medication, environment/assistive technology) within which are more specific groups (e.g. exercise: gait, balance and functional training; flexibility; strength and resistance) (Lamb et al 2011).
4. Plan how the specified groups will be used in synthesis and reporting.	Decide whether it is useful to pool all interventions in a single meta-analysis ('lumping'), within which specific characteristics can be explored as effect modifiers (e.g. in subgroups). Alternatively, if pooling all interventions is unlikely to address a useful question, separate synthesis of specific interventions may be more appropriate ('splitting'). Determining the right analytic approach is discussed further in Chapter 2, Section 2.3.2.	In a review of exercise for knee osteoarthritis, the different categories of exercise were combined in a single meta-analysis, addressing the question 'what is the effect of exercise on knee osteoarthritis?'. The categories were also analysed as subgroups within the meta-analysis to explore whether the effect size varied by type of exercise (Fransen et al 2015). Other subgroup analyses examined mode of delivery and dose.
5. Decide how to group interventions with multiple components or co-interventions.	Some interventions, especially those considered 'complex', include multiple components that could also be implemented independently (Guise et al 2014, Lewin et al 2017). These components might be eligible for inclusion in the review alone, or eligible only if used alongside an eligible intervention. Options for considering multi-component interventions may include the following. • Identifying intervention components for meta-regression or a components-based network meta-analysis (see Chapter 11 and Welton et al 2009, Caldwell and Welton 2016, Higgins et al 2019).	**Grouping by main component:** In a review of psychological therapies for panic disorder, two of the eight eligible therapies (psychoeducation and supportive psychotherapy) could be used alone or as part of a multi-component therapy. When accompanied by another eligible therapy, the intervention was categorized as the other therapy (i.e. psychoeducation + cognitive behavioural therapy was categorized as cognitive behavioural therapy) (Pompoli et al 2016). **Separate group:** In a review of psychosocial interventions for smoking cessation in pregnancy, two approaches were used. All intervention types were included in a single meta-analysis

	• Grouping based on the 'main' intervention component (Caldwell and Welton 2016). • Specifying a separate group ('multi-component interventions'). 'Lumping' multi-component interventions together may provide information about their effects in general; however, this approach may lead to unexplained heterogeneity and/or inability to identify which components are effective (Caldwell and Welton 2016). • Reporting results study by study. An option if components are expected to be so diverse that synthesis will not be interpretable. • Excluding multi-component interventions. An option if the effect of the intervention of interest cannot be discerned. This approach may reduce the relevance of the review. The first two approaches may be challenging but are likely to be most useful (Caldwell and Welton 2016). See Section 3.2.3.1. for the special case of when a co-intervention is administered in both treatment arms.	with subgroups for multi-component, single and tailored interventions. Separate meta-analyses were also performed for each intervention type, with categorization of multi-component interventions based on the 'main' component (Chamberlain et al 2017).
6. Build in contingencies by specifying both specific and broader intervention groups.	Consider grouping interventions at more than one level, so that studies of a broader group of interventions can be synthesized if too few studies are identified for synthesis in more specific groups. This will provide flexibility where review authors anticipate few studies contributing to specific groups (e.g. in reviews with diverse interventions, additional diversity in other PICO elements, or few studies overall, see also Chapter 2, Section 2.5.3.	In a review of psychosocial interventions for smoking cessation, the authors planned to group any psychosocial intervention in a single comparison (addressing the higher level question of whether, on average, psychosocial interventions are effective). Given that sufficient data were available, they also presented separate meta-analyses to examine the effects of specific types of psychosocial interventions (e.g. counselling, health education, incentives, social support) (Chamberlain et al 2017).

MECIR Box 3.2.c Relevant expectations for conduct of intervention reviews

C7: Predefining unambiguous criteria for interventions and comparators (**Mandatory**)

Define in advance the eligible interventions and the interventions against which these can be compared in the included studies.

Predefined, unambiguous eligibility criteria are a fundamental prerequisite for a systematic review. Specification of comparator interventions requires particular clarity: are the experimental interventions to be compared with an inactive control intervention (e.g. placebo, no treatment, standard care, or a waiting list control), or with an active control intervention (e.g. a different variant of the same intervention, a different drug, a different kind of therapy)? Any restrictions on interventions and comparators, for example, regarding delivery, dose, duration, intensity, co-interventions and features of complex interventions should also be predefined and explained.

Box 3.2.b Factors to consider when developing criteria for 'Types of interventions'

- What are the experimental and control (comparator) interventions of interest?
- Does the intervention have variations (e.g. dosage/intensity, mode of delivery, personnel who deliver it, frequency, duration or timing of delivery)?
- Are all variations to be included (for example, is there a dose below which the intervention may not be clinically appropriate, will all providers be included)?
- Will studies including only part of the intervention be included?
- Will studies including the intervention of interest combined with another intervention (co-intervention) be included?
- Have the different meanings of phrases such as 'control', 'placebo', 'no intervention' or 'usual care' been considered?

an overall synthesis can provide useful information for decision makers. Where differences in intervention characteristics are more substantial (such as delivery of brief alcohol counselling by nurses versus doctors), and are expected to have a substantial impact on the size of intervention effects, these differences should be examined in the synthesis. What constitutes an important difference requires judgement, but in general differences that alter decisions about how an intervention is implemented or whether the intervention is used or not are likely to be important. In such circumstances, review authors should consider specifying separate groups (or subgroups) to examine in their synthesis.

Clearly defined intervention groups serve two main purposes in the synthesis. First, the way in which interventions are grouped for synthesis (meta-analysis or other synthesis) is likely to influence review findings. Careful planning of intervention groups makes best use of the available data, avoids decisions that are influenced by study findings (which may introduce bias), and produces a review focused on questions relevant to decision makers. Second, the intervention groups specified in a protocol provide a standardized terminology for describing the interventions throughout the review, overcoming the varied descriptions used by study authors (e.g. where different labels are used for the same intervention, or similar labels used for different techniques) (Michie et al 2013). This standardization enables comparison and synthesis of information about intervention characteristics across studies (common characteristics and differences) and provides a consistent language for reporting that supports interpretation of review findings.

Table 3.2.b outlines a process for planning intervention groups as a basis for/precursor to synthesis, and the decision points and considerations at each step. The table is intended to guide, rather than to be prescriptive and, although it is presented as a sequence of steps, the process is likely to be iterative, and some steps may be done concurrently or in a different sequence. The process aims to minimize data-driven approaches that can arise once review authors have knowledge of the findings of the included studies. It also includes principles for developing a flexible plan that maximizes the potential to synthesize in circumstances where there are few studies, many variants of an intervention, or where the variants are difficult to anticipate. In all stages, review authors should consider how to categorize studies whose reports contain insufficient detail.

3.2.3 Defining which comparisons will be made

When articulating the PICO for each synthesis, defining the intervention groups alone is not sufficient for complete specification of the planned syntheses. The next step is to define the comparisons that will be made between the intervention groups. Setting aside for a moment more complex analyses such as network meta-analyses, which can simultaneously compare many groups (Chapter 11), standard meta-analysis (Chapter 10) aims to draw conclusions about the comparative effects of two groups at a time (i.e. which of two intervention groups is more effective?). These comparisons form the basis for the syntheses that will be undertaken if data are available. Cochrane Reviews sometimes include one comparison, but most often include multiple comparisons. Three commonly identified types of comparisons include the following (Davey et al 2011).

- Intervention versus placebo (e.g. placebo drug, sham surgical procedure, psychological placebo). Placebos are most commonly used in the evaluation of pharmacological interventions, but may be also be used in some non-pharmacological evaluations. For example:
  - newer generation antidepressants versus placebo (Hetrick et al 2012); and
  - vertebroplasty for osteoporotic vertebral compression fractures versus placebo (sham procedure) (Buchbinder et al 2018).
- Intervention versus control (e.g. no intervention, wait-list control, usual care). Both intervention arms may also receive standard therapy. For example:

  - chemotherapy or targeted therapy plus best supportive care (BSC) versus BSC for palliative treatment of esophageal and gastroesophageal-junction carcinoma (Janmaat et al 2017); and
  - personalized care planning versus usual care for people with long-term conditions (Coulter et al 2015).
- Intervention A versus intervention B. A comparison of active interventions may include comparison of the same intervention delivered at different time points, for different lengths of time or different doses, or two different interventions. For example:
  - early (commenced at less than two weeks of age) versus late (two weeks of age or more) parenteral zinc supplementation in term and preterm infants (Taylor et al 2017);
  - high intensity versus low intensity physical activity or exercise in people with hip or knee osteoarthritis (Regnaux et al 2015);
  - multimedia education versus other education for consumers about prescribed and over the counter medications (Ciciriello et al 2013).

The first two types of comparisons aim to establish the effectiveness of an intervention, while the last aims to compare the effectiveness of two interventions. However, the distinction between the placebo and control is often arbitrary, since any differences in the care provided between trials with a control arm and those with a placebo arm may be unimportant, especially where 'usual care' is provided to both. Therefore, placebo and control groups may be determined to be similar enough to be combined for synthesis.

In reviews including multiple intervention groups, many comparisons are possible. In some of these reviews, authors seek to synthesize evidence on the comparative effectiveness of all their included interventions, including where there may be only indirect comparison of some interventions across the included studies (Chapter 11, Section 11.2.1). However, in many reviews including multiple intervention groups, a limited subset of the possible comparisons will be selected. The chosen subset of comparisons should address the most important clinical and research questions. For example, if an established intervention (or dose of an intervention) is used in practice, then the synthesis would ideally compare novel or alternative interventions to this established intervention, and not, for example, to no intervention.

3.2.3.1 Dealing with co-interventions

Planning is needed for the special case where the *same* supplementary intervention is delivered to both the intervention and comparator groups. A supplementary intervention is an additional intervention delivered alongside the intervention of interest, such as massage in a review examining the effects of aromatherapy (i.e. aromatherapy plus massage versus massage alone). In many cases, the supplementary intervention will be unimportant and can be ignored. In other situations, the effect of the intervention of interest may differ according to whether participants receive the supplementary therapy. For example, the effect of aromatherapy among people who receive a massage may differ from the effect of the aromatherapy given alone. This will be the case if the intervention of interest interacts with the supplementary intervention leading to larger (synergistic) or smaller (dysynergistic/antagonistic) effects than the intervention

of interest alone (Squires et al 2013). While qualitative interactions are rare (where the effect of the intervention is in the opposite direction when combined with the supplementary intervention), it is possible that there will be more variation in the intervention effects (heterogeneity) when supplementary interventions are involved, and it is important to plan for this. Approaches for dealing with this in the statistical synthesis may include fitting a random-effects meta-analysis model that encompasses heterogeneity (Chapter 10, Section 10.10.4), or investigating whether the intervention effect is modified by the addition of the supplementary intervention through subgroup analysis (Chapter 10, Section 10.11.2).

3.2.4 Selecting, prioritizing and grouping review outcomes

3.2.4.1 Selecting review outcomes

Broad outcome domains are decided at the time of setting up the review PICO (see Chapter 2). Once the broad domains are agreed, further specification is required to define the domains to facilitate reporting and synthesis (i.e. the PICO for each synthesis) (see Chapter 2, Section 2.3). The process for specifying and grouping outcomes largely parallels that used for specifying intervention groups.

Reporting of outcomes should rarely determine study eligibility for a review. In particular, studies should not be excluded because they do not report results of an outcome they may have measured, or provide 'no usable data' (MECIR Box 3.2.d). This is essential to avoid bias arising from selective reporting of findings by the study authors (see Chapter 13). However, in some circumstances, the measurement of certain outcomes may be a study eligibility criterion. This may be the case, for example, when the review addresses the

MECIR Box 3.2.d Relevant expectations for conduct of intervention reviews

C8: Clarifying role of outcomes (**Mandatory**)

Clarify in advance whether outcomes listed under 'Criteria for considering studies for this review' are used as criteria for including studies (rather than as a list of the outcomes of interest within whichever studies are included).	Outcome measures should not always form part of the criteria for including studies in a review. However, some reviews do legitimately restrict eligibility to specific outcomes. For example, the same intervention may be studied in the same population for different purposes (e.g. hormone replacement therapy, or aspirin); or a review may address specifically the adverse effects of an intervention used for several conditions. If authors do exclude studies on the basis of outcomes, care should be taken to ascertain that relevant outcomes are not available because they have not been measured rather than simply not reported.

C14: Predefining outcome domains (**Mandatory**)

Define in advance outcomes that are critical to the review, and any additional important outcomes.

Full specification of the outcomes includes consideration of outcome domains (e.g. quality of life) and outcome measures (e.g. SF-36). Predefinition of outcome reduces the risk of selective outcome reporting. The *critical outcomes* should be as few as possible and should normally reflect at least one potential benefit and at least one potential area of harm. It is expected that the review should be able to synthesize these outcomes if eligible studies are identified, and that the conclusions of the review will be based largely on the effects of the interventions on these outcomes. Additional important outcomes may also be specified. Up to seven critical and important outcomes will form the basis of the GRADE assessment and summarized in the review's abstract and other summary formats, although the review may measure more than seven outcomes.

C15: Choosing outcomes (**Mandatory**)

Choose only outcomes that are critical or important to users of the review such as healthcare consumers, health professionals and policy makers.

Cochrane Reviews are intended to support clinical practice and policy, and should address outcomes that are critical or important to consumers. These should be specified at protocol stage. Where available, established sets of core outcomes should be used. Patient-reported outcomes should be included where possible. It is also important to judge whether evidence of resource use and costs might be an important component of decisions to adopt the intervention or alternative management strategies around the world. Large numbers of outcomes, while sometimes necessary, can make reviews unfocused, unmanageable for the user, and prone to selective outcome reporting bias.

	Biochemical, interim and process outcomes should be considered where they are important to decision makers. Any outcomes that would not be described as critical or important can be left out of the review.
C16: Predefining outcome measures (**Highly desirable**)	
Define in advance details of what will constitute acceptable outcome measures (e.g. diagnostic criteria, scales, composite outcomes).	Having decided what outcomes are of interest to the review, authors should clarify acceptable ways in which these outcomes can be measured. It may be difficult, however, to predefine adverse effects.

potential for an intervention to *prevent* a particular outcome, or when the review addresses a specific purpose of an intervention that can be used in the same population for different purposes (such as hormone replacement therapy, or aspirin).

In general, systematic reviews should aim to **include outcomes that are likely to be meaningful to the intended users and recipients of the reviewed evidence.** This may include clinicians, patients (consumers), the general public, administrators and policy makers. Outcomes may include survival (mortality), clinical events (e.g. strokes or myocardial infarction), behavioural outcomes (e.g. changes in diet, use of services), patient-reported outcomes (e.g. symptoms, quality of life), adverse events, burdens (e.g. demands on caregivers, frequency of tests, restrictions on lifestyle) and economic outcomes (e.g. cost and resource use). It is critical that outcomes used to assess adverse effects as well as outcomes used to assess beneficial effects are among those addressed by a review (see Chapter 19).

Outcomes that are trivial or meaningless to decision makers should not be included in Cochrane Reviews. Inclusion of outcomes that are of little or no importance risks overwhelming and potentially misleading readers. Interim or surrogate outcomes measures, such as laboratory results or radiologic results (e.g. loss of bone mineral content as a surrogate for fractures in hormone replacement therapy), while potentially helpful in explaining effects or determining intervention integrity (see Chapter 5, Section 5.3.4.1), can also be misleading since they may not predict clinically important outcomes accurately. Many interventions reduce the risk for a surrogate outcome but have no effect or have harmful effects on clinically relevant outcomes, and some interventions have no effect on surrogate measures but improve clinical outcomes.

Various sources can be used to develop a list of relevant outcomes, including input from consumers and advisory groups (see Chapter 2), the clinical experiences of the review authors, and evidence from the literature (including qualitative research about outcomes important to those affected (see Chapter 21)). A further driver of outcome selection is consideration of outcomes used in related reviews. Harmonization of outcomes across reviews addressing related questions facilitates broader evidence

synthesis questions being addressed through the use of Overviews of reviews (see online Chapter V).

Outcomes considered to be meaningful, and therefore addressed in a review, may not have been reported in the primary studies. For example, quality of life is an important outcome, perhaps the most important outcome, for people considering whether or not to use chemotherapy for advanced cancer, even if the available studies are found to report only survival (see Chapter 18). A further example arises with timing of the outcome measurement, where time points determined as clinically meaningful in a review are not measured in the primary studies. Including and discussing all important outcomes in a review will highlight gaps in the primary research and encourage researchers to address these gaps in future studies.

3.2.4.2 Prioritizing review outcomes

Once a full list of relevant outcomes has been compiled for the review, authors should prioritize the outcomes and select the outcomes of most relevance to the review question. The GRADE approach to assessing the certainty of evidence (see Chapter 14) suggests that review authors separate outcomes into those that are 'critical', 'important' and 'not important' for decision making.

The critical outcomes are the essential outcomes for decision making, and are those that would form the basis of a 'Summary of findings' table or other summary versions of the review, such as the Abstract or Plain Language Summary. 'Summary of findings' tables provide key information about the amount of evidence for important comparisons and outcomes, the quality of the evidence and the magnitude of effect (see Chapter 14, Section 14.1). There should be no more than seven outcomes included in a 'Summary of findings' table, and those outcomes that will be included in summaries should be specified at the protocol stage. They should generally not include surrogate or interim outcomes. They should not be chosen on the basis of any anticipated or observed magnitude of effect, or because they are likely to have been addressed in the studies to be reviewed. Box 3.2.c summarizes the principal factors to consider when selecting and prioritizing review outcomes.

Box 3.2.c Factors to consider when selecting and prioritizing review outcomes

- Consider outcomes relevant to all potential decision makers.
- Critical outcomes are those that are essential for decision making, and should usually have an emphasis on patient-important outcomes and be determined by core outcomes sets.
- Additional outcomes important to decision makers may also be included in the review. Any outcomes not considered important to decision makers should be excluded from the review.
- Up to seven critical and important outcomes should be selected for inclusion in summary versions of the review, including 'Summary of findings' tables, Abstracts and Plain Language Summaries. Remember that summaries may be read alone, and should include the most important outcomes for decision makers.
- Ensure that outcomes cover potential as well as actual adverse effects.

3.2.4.3 Defining and grouping outcomes for synthesis

Table 3.2.c outlines a process for planning for the diversity in outcome measurement that may be encountered in the studies included in a review and which can complicate, and sometimes prevent, synthesis. Research has repeatedly documented inconsistency in the outcomes measured across trials in the same clinical areas (Harrison et al 2016, Williamson et al 2017). This inconsistency occurs across all aspects of outcome measurement, including the broad domains considered, the outcomes measured, the way these outcomes are labelled and defined, and the methods and timing of measurement. For example, a review of outcome measures used in 563 studies of interventions for dementia and mild cognitive impairment found that 321 unique measurement methods were used for 1278 assessments of cognitive outcomes (Harrison et al 2016). Initiatives like COMET (Core Outcome Measures in Effectiveness Trials) aim to encourage standardization of outcome measurement across trials (Williamson et al 2017), but these initiatives are comparatively new and review authors will inevitably encounter diversity in outcomes across studies.

The process begins by describing the scope of each outcome domain in sufficient detail to enable outcomes from included studies to be categorized (Table 3.2.c Step 1). This step may be straightforward in areas for which core outcome sets (or equivalent systems) exist (Table 3.2.c Step 2). The methods and timing of outcome measurement also need to be specified, giving consideration to how differences across studies will be handled (Table 3.2.c Steps 3 and 4). Subsequent steps consider options for dealing with studies that report multiple measures within an outcome domain (Table 3.2.c Step 5), planning how outcome domains will be used in synthesis (Table 3.2.c Step 6), and building in contingencies to maximize potential to synthesize (Table 3.2.c Step 7).

3.3 Determining which study designs to include

Some study designs are more appropriate than others for answering particular questions. Authors need to consider a priori what study designs are likely to provide reliable data with which to address the objectives of their review (MECIR Box 3.3.a). Sections 3.3.1 and 3.3.2 cover randomized and non-randomized designs for assessing treatment effects; Chapter 17 (Section 17.2.5) discusses other study designs in the context of addressing intervention complexity.

3.3.1 Including randomized trials

Because Cochrane Reviews address questions about the effects of health care, they focus primarily on randomized trials and randomized trials should be included if they are feasible for the interventions of interest (MECIR Box 3.3.b). Randomization is the only way to prevent systematic differences between baseline characteristics of participants in different intervention groups in terms of both known and unknown (or unmeasured) confounders (see Chapter 8), and claims about cause and effect can be based on their findings with far more confidence than almost any other type of study. For clinical interventions, deciding who receives an intervention and who does not is influenced by many factors, including prognostic factors. Empirical evidence

Table 3.2.c A process for planning outcome groups for synthesis

Step	Considerations	Examples
1. Fully specify outcome domains.	For each outcome domain, provide a short label (e.g. cognition, consumer evaluation of care) and describe the domain in sufficient detail to enable eligible outcomes from each included study to be categorized. The definition should be based on the concept (or construct) measured, that is 'what' is measured. 'When' and 'how' the outcome is measured will be considered in subsequent steps. Outcomes can be defined hierarchically, starting with very broad groups (e.g. physiological/clinical outcomes, life impact, adverse events), then outcome domains (e.g. functioning and perceived health status are domains within 'life impact'). Within these may be narrower domains (e.g. physical function, cognitive function), and then specific outcome measures (Dodd et al 2018). The level at which outcomes are grouped for synthesis alters the question addressed, and so decisions should be guided by the review objectives. In specifying outcome domains: • definitions should reflect existing systems if available, or relevant literature and terminology understood by decision makers; • where outcomes are likely to be inconsistently labelled and described, listing examples may convey the scope of the domain; • consider the level at which domains will be defined (broad versus narrow) and the implications for reporting and synthesis: combining diverse outcomes may lead to unexplained heterogeneity whereas narrowly specified outcomes may prevent synthesis when few studies report specific measures;	In a review of computer-based interventions for sexual health promotion, three broad outcome domains were defined (cognitions, behaviours, biological) based on a conceptual model of how the intervention might work. Each domain comprised more specific domains and outcomes (e.g. condom use, seeking health services such as STI testing); listing these helped define the broad domains and guided categorization of the diverse outcomes reported in included studies (Bailey et al 2010). In a protocol for a review of social media interventions for improving health, the rationale for synthesizing broad groupings of outcomes (e.g. health behaviours, physical health) was based on prediction of a common underlying mechanism by which the intervention would work, and the review objective, which focused on overall health rather than specific outcomes (Welch et al 2018).

	• a causal path or logic model may help identify logical groupings of related outcomes for reporting and analysis, and alternative levels at which to synthesize.	
2. Determine whether there is an existing system for identifying and grouping important outcomes.	Systems for categorizing outcomes include core outcome sets including the COMET and ICHOM initiatives, and outcome taxonomies (Dodd et al 2018). These systems define agreed outcomes that should be measured for specific conditions (Williamson et al 2017).These systems can be used to standardize the varied outcome labels used across studies and enable grouping and comparison (Kirkham et al 2013). Agreed terminology may help decision makers interpret review findings. The COMET website provides a database of core outcome sets agreed or in development. Some Cochrane Groups have developed their own outcome sets. While the availability of outcome sets and taxonomies varies across clinical areas, several taxonomies exist for specifying broad outcome domains (e.g. Dodd et al 2018, ICHOM 2018).	In a review of combined diet and exercise for preventing gestational diabetes mellitus, a core outcome set agreed by the Cochrane Pregnancy and Childbirth group was used (Shepherd et al 2017). In a review of decision aids for people facing health treatment or screening decisions (Stacey et al 2017), outcome domains were based on criteria for evaluating decision aids agreed in the International Patient Decision Aids Standards (IPDAS). Doing so helped to assess the use of aids across diverse clinical decisions. The Cochrane Consumers and Communication Group has an agreed taxonomy to guide specification of outcomes of importance in evaluating communication interventions (Cochrane Consumers & Communication Group).
3. Define the outcome time points.	A key attribute of defining an outcome is specifying the time of measurement. In reviews, time frames, and not specific time points, are often specified to handle the likely diversity in timing of outcome measurement across studies (e.g. a 'medium-term' time frame might be defined as including outcomes measured between 6 and 12 months). In specifying outcome timing: • focus on 'clinically meaningful' time points (e.g. considering the course of the condition over time and duration of the intervention may determine whether	In a review of psychological therapies for panic disorder, the main outcomes were 'short-term' (≤ 6 months from treatment commencement). 'Long-term' outcomes (> 6 months from treatment commencement) were considered important, but not specified as critical because of concerns of participant attrition (Pompoli et al 2018). In contrast, in a review of antidepressants, a clinically meaningful time frame of 6 to 12 months might be specified for the critical outcome 'depression', since this is the recommended treatment duration. However, it may be anticipated that many studies will be of shorter

(*Continued*)

Table 3.2.c (Continued)

Step	Considerations	Examples
	short-term or long-term outcomes are important); • consider whether there are agreed or accepted outcome time points (e.g. standards in a clinical area such as an NIH task force suggestion for at least 6 to 12 months follow-up for chronic low back pain (Deyo et al 2014), or core outcome sets (Williamson et al 2017); • consider carefully the width of the time frame (e.g. what constitutes 'short term' for this review?). Narrow time frames may lead to few studies in the synthesis. Broad time frames may lead to multiplicity (see Step 5) and difficulties with interpretation if the timing is very diverse across studies.	duration with short-term follow-up, so an additional important outcome of 'depression (< 3 months)' might also be specified.
4. Specify the measurement tool or measurement method.	For each outcome domain, specify: • measurement methods or tools that provide an appropriate assessment of the domain or specific outcome (e.g. including clinical assessment, laboratory tests, objective measures, and patient-reported outcome measures (PROMs)); • whether different methods or tools are comparable measures of a domain, which has implications for synthesis (Step 6). Minimum criteria for inclusion of a measure may include: • adequate evidence of *reliability* (e.g. consistent scores across time and raters when the outcome is unchanged), and *validity* (e.g. comparable results to similar measures, including a gold standard if available); and	In a review of interventions to support women to stop smoking, objective (biochemically validated) and subjective (self-report) measures of smoking cessation were specified separately to examine bias due to the method used to measure the outcome (Step 6) (Chamberlain et al 2017). In a review of high-intensity versus low-intensity exercise for osteoarthritis, measures of pain were selected based on relevance of the content and properties of the measurement tool (i.e. evidence of validity and reliability) (Regnaux et al 2015).

	• for self-reported measures, items that cover the outcome/domain and are developed using theory, empirical evidence and consumer involvement. Measures may be identified from core outcome sets (e.g. Williamson et al 2017, ICHOM 2018) or systematic reviews of instruments (see COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative for a database of examples).	
5. Specify how multiplicity of outcomes will be handled.	For a particular domain, multiple outcomes within a study may be available for inclusion. This may arise from: • multiple outcomes measured within a domain (e.g. 'anxiety' and 'depression' in a 'mental health' domain); • multiple methods to measure the outcome (e.g. self-reported depression, clinician-rated depression), or tools/instruments (e.g. Hamilton Depression Rating Scale, Beck Depression Inventory), as well as their subscales; • multiple time points measured within a time frame. Effects of the intervention calculated from these different sources of multiplicity are statistically dependent, since they have been calculated using the same participants. To deal with this dependency, select only one outcome per study for a particular comparison, or use a meta-analysis method that accounts for the dependency (see Step 6). Pre-specify the method of selection from multiple outcomes or measures in the protocol, using an approach that is independent of the result (see Chapter 9, Table 9.3.c) (López-López et al 2018). Document all eligible outcomes or measures in the 'Characteristics of included studies' table, noting which was selected and why.	The following hierarchy was specified to select one outcome per domain in a review examining the effects of portion, package or tableware size (Hollands et al 2015): • the study's primary outcome; • the outcome that was most proximal to the health outcome in the context of the specific intervention; • the outcome that provided the largest-scale measure of the domain (e.g. total amount of food consumed selected ahead of amount of vegetables consumed). Selection of the outcome was made blinded to the results. All available outcome measures were documented in the 'Characteristics of included studies' table. In a review of audit and feedback for healthcare providers, the outcome domains were 'provider performance' (e.g. compliance with recommended use of a laboratory test) and 'patient health outcomes' (e.g. smoking status, blood pressure) (Ivers et al 2012). For each domain, outcomes were selected using the following hierarchy: • the study's primary outcome; • the outcome used in the sample size calculation; and • the outcome with the median effect.

(*Continued*)

Table 3.2.c (Continued)

Step	Considerations	Examples
	Multiplicity can arise from the reporting of multiple analyses of the same outcome (e.g. analyses that do and do not adjust for prognostic factors; intention-to-treat and per-protocol analyses) and multiple reports of the same study (e.g. journal articles, conference abstracts). Approaches for dealing with this type of multiplicity should also be specified in the protocol (López-López et al 2018). It may be difficult to anticipate all forms of multiplicity when developing a protocol. Any post-hoc approaches used to select outcomes or results should be noted in the 'Differences between protocol and review' section.	
6. Plan how the specified outcome domains will be used in the synthesis.	When different measurement methods or tools have been used across studies, consideration must be given to how these will be synthesized. Options include the following. • Synthesize different measures of the same outcome (or outcome domain) together. This approach is likely to maximize the potential to synthesize. A subgroup or sensitivity analysis might be undertaken to examine if the effects are modified by, or robust to, the type of measurement method or tool (Chapter 10, Sections 10.11.2 and 10.14). There may be increased heterogeneity, warranting use of a random-effects model (Chapter 10, Section 10.10.4). • Synthesize each outcome measure separately (e.g. separate meta-analyses of Beck's Depression Inventory and Hamilton Depression Rating Scale). However, when the measurement methods all provide a measure of the same domain, multiple meta-analyses	In a review of interventions to support women to stop smoking, separate outcome domains were specified for biochemically validated measures of smoking and self-report measures. The two domains were meta-analysed together, but sensitivity analyses were undertaken restricting the meta-analyses to studies with only biochemically validated outcomes, to examine if the results were robust to the method of measurement (Chamberlain et al 2017). In a review of psychological therapies for youth internalizing and externalizing disorders, most studies contributed multiple effects (e.g. in one meta-analysis of 443 studies, there were 5139 included measures). The authors used multilevel modelling to address the dependency among multiple effects contributed from each study (Weisz et al 2017).

	can lead to difficulties in interpretation and an increase in the type I error rate (Bender et al 2008, López-López et al 2018). • Include all the available effect estimates, using a meta-analysis method that models or accounts for the dependency. This option has the advantage of using all information which may lead to greater precision in estimating the intervention effects (López-López et al 2018). Options include multivariate meta-analysis (Mavridis and Salanti 2013), multilevel models (Konstantopoulos 2011) or robust variance estimation (Hedges et al 2010) (see López-López et al 2018 for further discussion).	
7. Where possible, build in contingencies by specifying both specific and broader outcome domains.	Consider building in flexibility to group outcomes at different levels or time intervals. Inflexible approaches can undermine the potential to synthesize, especially when few studies are anticipated, or there is likely to be diversity in the way outcomes are defined and measured and the timing of measurement. If insufficient studies report data for meaningful synthesis using the narrower domains, the broader domains can be used (see also Chapter 2, Section 2.5.3).	Consider a hypothetical review aiming to examine the effects of behavioural psychological interventions for the treatment of overweight and obese adults. A specific outcome is body mass index (BMI). However, also specifying a broader outcome domain 'indicator of body mass' will facilitate synthesis in the circumstance where few studies report BMI, but most report an indicator of body mass (such as weight or waist circumference). This is particularly important when few studies may be anticipated or there is expected diversity in the measurement methods or tools.

MECIR Box 3.3.a Relevant expectations for conduct of intervention reviews

C9: Predefining study designs (**Mandatory**)

Define in advance the eligibility criteria for study designs in a clear and unambiguous way, with a focus on features of a study's design rather than design labels.	Predefined, unambiguous eligibility criteria are a fundamental prerequisite for a systematic review. This is particularly important when non-randomized studies are considered. Some labels commonly used to define study designs can be ambiguous. For example a 'double blind' study may not make it clear who was blinded; a 'case-control' study may be nested within a cohort, or be undertaken in a cross-sectional manner; or a 'prospective' study may have only some features defined or undertaken prospectively.

C11: Justifying choice of study designs (**Mandatory**)

Justify the choice of eligible study designs.	It might be difficult to address some interventions or some outcomes in randomized trials. Authors should be able to justify why they have chosen either to restrict the review to randomized trials or to include non-randomized studies. The particular study designs included should be justified with regard to appropriateness to the review question and with regard to potential for bias.

MECIR Box 3.3.b Relevant expectations for conduct of intervention reviews

C10: Including randomized trials (**Mandatory**)

Include randomized trials as eligible for inclusion in the review, if it is feasible to conduct them to evaluate the interventions and outcomes of interest.	Randomized trials are the best study design for evaluating the efficacy of interventions. If it is feasible to conduct them to evaluate questions that are being addressed by the review, they must be considered eligible for the review. However, appropriate exclusion criteria may be put in place, for example regarding length of follow-up.

suggests that, on average, non-randomized studies produce effect estimates that indicate more extreme benefits of the effects of health care than randomized trials. However, the extent, and even the direction, of the bias is difficult to predict. These issues are discussed at length in Chapter 24, which provides guidance on when it might be appropriate to include non-randomized studies in a Cochrane Review.

Practical considerations also motivate the restriction of many Cochrane Reviews to randomized trials. In recent decades there has been considerable investment internationally in establishing infrastructure to index and identify randomized trials. Cochrane has contributed to these efforts, including building up and maintaining a database of randomized trials, developing search filters to aid their identification, working with MEDLINE to improve tagging and identification of randomized trials, and using machine learning and crowdsourcing to reduce author workload in identifying randomized trials (Chapter 4, Section 4.6.6.2). The same scale of organizational investment has not (yet) been matched for the identification of other types of studies. Consequently, identifying and including other types of studies may require additional efforts to identify studies and to keep the review up to date, and might increase the risk that the result of the review will be influenced by publication bias. This issue and other bias-related issues that are important to consider when defining types of studies are discussed in detail in Chapters 7 and 13.

Specific aspects of study design and conduct should be considered when defining eligibility criteria, even if the review is restricted to randomized trials. For example, whether cluster-randomized trials (Chapter 23, Section 23.1) and crossover trials (Chapter 23, Section 23.2) are eligible, as well as other criteria for eligibility such as use of a placebo comparison group, evaluation of outcomes blinded to allocation sequence, or a minimum period of follow-up. There will always be a trade-off between restrictive study design criteria (which might result in the inclusion of studies that are at low risk of bias, but very few in number) and more liberal design criteria (which might result in the inclusion of more studies, but at a higher risk of bias). Furthermore, excessively broad criteria might result in the inclusion of misleading evidence. If, for example, interest focuses on whether a therapy improves survival in patients with a chronic condition, it might be inappropriate to look at studies of very short duration, except to make explicit the point that they cannot address the question of interest.

3.3.2 Including non-randomized studies

The decision of whether non-randomized studies (and what type) will be included is decided alongside the formulation of the review PICO. The main drivers that may lead to the inclusion of non-randomized studies include: (i) when randomized trials are unable to address the effects of the intervention on harm and long-term outcomes or in specific populations or settings; or (ii) for interventions that cannot be randomized (e.g. policy change introduced in a single or small number of jurisdictions) (see Chapter 24). Cochrane, in collaboration with others, has developed guidance for review authors to support their decision about when to look for and include non-randomized studies (Schünemann et al 2013).

Non-randomized designs have the commonality of not using randomization to allocate units to comparison groups, but their different design features mean that they are variable in their susceptibility to bias. Eligibility criteria should be based on explicit

study design features, and not the study labels applied by the primary researchers (e.g. case-control, cohort), which are often used inconsistently (Reeves et al 2017; see Chapter 24).

When non-randomized studies are included, review authors should consider how the studies will be grouped and used in the synthesis. The Cochrane Non-randomized Studies Methods Group taxonomy of design features (see Chapter 24) may provide a basis for grouping together studies that are expected to have similar inferential strength and for providing a consistent language for describing the study design.

Once decisions have been made about grouping study designs, planning of how these will be used in the synthesis is required. Review authors need to decide whether it is useful to synthesize results from non-randomized studies and, if so, whether results from randomized trials and non-randomized studies should be included in the same synthesis (for the purpose of examining whether study design explains heterogeneity among the intervention effects), or whether the effects should be synthesized in separate comparisons (Valentine and Thompson 2013). Decisions should be made for each of the different types of non-randomized studies under consideration. Review authors might anticipate increased heterogeneity when non-randomized studies are synthesized, and adoption of a meta-analysis model that encompasses heterogeneity is wise (Valentine and Thompson 2013) (such as a random effects model, see Chapter 10, Section 10.10.4). For further discussion of non-randomized studies, see Chapter 24.

3.4 Eligibility based on publication status and language

Chapter 4 contains detailed guidance on how to identify studies from a range of sources including, but not limited to, those in peer-reviewed journals. In general, a strategy to include studies reported in all types of publication will reduce bias (Chapter 7). There would need to be a compelling argument for the exclusion of studies on the basis of their publication status (MECIR Box 3.4.a), including unpublished studies, partially published studies, and studies published in 'grey' literature sources. Given the additional challenge in obtaining unpublished studies, it is possible that any unpublished studies identified in a given review may be an unrepresentative subset of all the unpublished studies in existence. However, the bias this introduces is of less concern than the bias

MECIR Box 3.4.a Relevant expectations for conduct of intervention reviews

C12: Excluding studies based on publication status (**Mandatory**)

Include studies irrespective of their publication status, unless exclusion is explicitly justified.	Obtaining and including data from unpublished studies (including grey literature) can reduce the effects of publication bias. However, the unpublished studies that can be located may be an unrepresentative sample of all unpublished studies.

introduced by excluding all unpublished studies, given what is known about the impact of reporting biases (see Chapter 13 on bias due to missing studies, and Chapter 4, Section 4.3 for a more detailed discussion of searching for unpublished and grey literature).

Likewise, while searching for, and analysing, studies in any language can be extremely resource-intensive, review authors should consider carefully the implications for bias (and equity, see Chapter 16) if they restrict eligible studies to those published in one specific language (usually English). See Chapter 4 (Section 4.4.5) for further discussion of language and other restrictions while searching.

3.5 Chapter information

Authors: Joanne E McKenzie, Sue E Brennan, Rebecca E Ryan, Hilary J Thomson, Renea V Johnston, James Thomas

Acknowledgements: This chapter builds on earlier versions of the *Handbook*. In particular, Chapter 5, edited by Denise O'Connor, Sally Green and Julian Higgins.

Funding: JEM is supported by an Australian National Health and Medical Research Council (NHMRC) Career Development Fellowship (1143429). SEB and RER's positions are supported by the NHMRC Cochrane Collaboration Funding Program. HJT is funded by the UK Medical Research Council (MC_UU_12017-13 and MC_UU_12017-15) and Scottish Government Chief Scientist Office (SPHSU13 and SPHSU15). RVJ's position is supported by the NHMRC Cochrane Collaboration Funding Program and Cabrini Institute. JT is supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care North Thames at Barts Health NHS Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

3.6 References

Bailey JV, Murray E, Rait G, Mercer CH, Morris RW, Peacock R, Cassell J, Nazareth I. Interactive computer-based interventions for sexual health promotion. *Cochrane Database of Systematic Reviews* 2010; **9**: CD006483.

Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, Thorlund K. Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology* 2008; **61**: 857–865.

Buchbinder R, Johnston RV, Rischin KJ, Homik J, Jones CA, Golmohammadi K, Kallmes DF. Percutaneous vertebroplasty for osteoporotic vertebral compression fracture. *Cochrane Database of Systematic Reviews* 2018; **4**: CD006349.

Caldwell DM, Welton NJ. Approaches for synthesising complex mental health interventions in meta-analysis. *Evidence-Based Mental Health* 2016; **19**: 16–21.

Chamberlain C, O'Mara-Eves A, Porter J, Coleman T, Perlen S, Thomas J, McKenzie J. Psychosocial interventions for supporting women to stop smoking in pregnancy. *Cochrane Database of Systematic Reviews* 2017; **2**: CD001055.

Ciciriello S, Johnston RV, Osborne RH, Wicks I, deKroo T, Clerehan R, O'Neill C, Buchbinder R. Multimedia educational interventions for consumers about prescribed and over-the-counter medications. *Cochrane Database of Systematic Reviews* 2013; **4**: CD008416.

Cochrane Consumers & Communication Group. Outcomes of Interest to the Cochrane Consumers & Communication Group: taxonomy. http://cccrg.cochrane.org/.

COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative. COSMIN database of systematic reviews of outcome measurement instruments. https://database.cosmin.nl/.

Coulter A, Entwistle VA, Eccles A, Ryan S, Shepperd S, Perera R. Personalised care planning for adults with chronic or long-term health conditions. *Cochrane Database of Systematic Reviews* 2015; **3**: CD010523.

Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology* 2011; **11**: 160.

Desroches S, Lapointe A, Ratte S, Gravel K, Legare F, Turcotte S. Interventions to enhance adherence to dietary advice for preventing and managing chronic diseases in adults. *Cochrane Database of Systematic Reviews* 2013; **2**: CD008722.

Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, Carrino J, Chou R, Cook K, DeLitto A, Goertz C, Khalsa P, Loeser J, Mackey S, Panagis J, Rainville J, Tosteson T, Turk D, Von Korff M, Weiner DK. Report of the NIH Task Force on research standards for chronic low back pain. *Journal of Pain* 2014; **15**: 569–585.

Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *Journal of Clinical Epidemiology* 2018; **96**: 84–92.

Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ* 2017; **356**: j573.

Fransen M, McConnell S, Harmer AR, Van der Esch M, Simic M, Bennell KL. Exercise for osteoarthritis of the knee. *Cochrane Database of Systematic Reviews* 2015; **1**: CD004376.

Guise JM, Chang C, Viswanathan M, Glick S, Treadwell J, Umscheid CA. *Systematic reviews of complex multicomponent health care interventions. Report No. 14-EHC003-EF*. Rockville, MD: Agency for Healthcare Research and Quality; 2014.

Harrison JK, Noel-Storr AH, Demeyere N, Reynish EL, Quinn TJ. Outcomes measures in a decade of dementia and mild cognitive impairment trials. *Alzheimer's Research and Therapy* 2016; **8**: 48.

Hedges LV, Tipton E, Johnson M, C. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods* 2010; **1**: 39–65.

Hetrick SE, McKenzie JE, Cox GR, Simmons MB, Merry SN. Newer generation antidepressants for depressive disorders in children and adolescents. *Cochrane Database of Systematic Reviews* 2012; **11**: CD004851.

Higgins JPT, López-López JA, Becker BJ, Davies SR, Dawson S, Grimshaw JM, McGuinness LA, Moore THM, Rehfuess E, Thomas J, Caldwell DM. Synthesizing quantitative evidence in systematic reviews of complex health interventions. *BMJ Global Health* 2019; **4**: e000858.

Hoffmann T, Glasziou P, Barbour V, Macdonald H. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014; **1687**: 1–13.

Hollands GJ, Shemilt I, Marteau TM, Jebb SA, Lewis HB, Wei Y, Higgins JPT, Ogilvie D. Portion, package or tableware size for changing selection and consumption of food, alcohol and tobacco. *Cochrane Database of Systematic Reviews* 2015; **9**: CD011045.

Howe TE, Shea B, Dawson LJ, Downie F, Murray A, Ross C, Harbour RT, Caldwell LM, Creed G. Exercise for preventing and treating osteoporosis in postmenopausal women. *Cochrane Database of Systematic Reviews* 2011; **7**: CD000333.

ICHOM. The International Consortium for Health Outcomes Measurement 2018. http://www.ichom.org/.

IPDAS. International Patient Decision Aid Standards Collaboration (IPDAS) standards. www.ipdas.ohri.ca.

Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, O'Brien MA, Johansen M, Grimshaw J, Oxman AD. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews* 2012; **6**: CD000259.

Janmaat VT, Steyerberg EW, van der Gaast A, Mathijssen RH, Bruno MJ, Peppelenbosch MP, Kuipers EJ, Spaander MC. Palliative chemotherapy and targeted therapies for esophageal and gastroesophageal junction cancer. *Cochrane Database of Systematic Reviews* 2017; **11**: CD004063.

Kendrick D, Kumar A, Carpenter H, Zijlstra GAR, Skelton DA, Cook JR, Stevens Z, Belcher CM, Haworth D, Gawler SJ, Gage H, Masud T, Bowling A, Pearl M, Morris RW, Iliffe S, Delbaere K. Exercise for reducing fear of falling in older people living in the community. *Cochrane Database of Systematic Reviews* 2014; **11**: CD009848.

Kirkham JJ, Gargon E, Clarke M, Williamson PR. Can a core outcome set improve the quality of systematic reviews? A survey of the Co-ordinating Editors of Cochrane Review Groups. *Trials* 2013; **14**: 21.

Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods* 2011; **2**: 61–76.

Lamb SE, Becker C, Gillespie LD, Smith JL, Finnegan S, Potter R, Pfeiffer K. Reporting of complex interventions in clinical trials: development of a taxonomy to classify and describe fall-prevention interventions. *Trials* 2011; **12**: 125.

Lewin S, Hendry M, Chandler J, Oxman AD, Michie S, Shepperd S, Reeves BC, Tugwell P, Hannes K, Rehfuess EA, Welch V, Mckenzie JE, Burford B, Petkovic J, Anderson LM, Harris J, Noyes J. Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Medical Research Methodology* 2017; **17**: 76.

López-López JA, Page MJ, Lipsey MW, Higgins JPT. Dealing with multiplicity of effect sizes in systematic reviews and meta-analyses. *Research Synthesis Methods* 2018; **9**: 336–351.

Mavridis D, Salanti G. A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research* 2013; **22**: 133–158.

Michie S, van Stralen M, West R. The Behaviour Change Wheel: a new method for characterising and designing behaviour change interventions. *Implementation Science* 2011; **6**: 42.

Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, Eccles MP, Cane J, Wood CE. The behavior change technique taxonomy (v1) of 93 hierarchically clustered

techniques: building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine* 2013; **46**: 81–95.

Moraes VY, Lenza M, Tamaoki MJ, Faloppa F, Belloti JC. Platelet-rich therapies for musculoskeletal soft tissue injuries. *Cochrane Database of Systematic Reviews* 2014; **4**: CD010071.

O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, Evans T, Pardo Pardo J, Waters E, White H, Tugwell P. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. *Journal of Clinical Epidemiology* 2014; **67**: 56–64.

Pompoli A, Furukawa TA, Imai H, Tajika A, Efthimiou O, Salanti G. Psychological therapies for panic disorder with or without agoraphobia in adults: a network meta-analysis. *Cochrane Database of Systematic Reviews* 2016; **4**: CD011004.

Pompoli A, Furukawa TA, Efthimiou O, Imai H, Tajika A, Salanti G. Dismantling cognitive-behaviour therapy for panic disorder: a systematic review and component network meta-analysis. *Psychological Medicine* 2018; **48**: 1–9.

Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series-paper 5: a checklist for classifying studies evaluating the effects on health interventions – a taxonomy without labels. *Journal of Clinical Epidemiology* 2017; **89**: 30–42.

Regnaux J-P, Lefevre-Colau M-M, Trinquart L, Nguyen C, Boutron I, Brosseau L, Ravaud P. High-intensity versus low-intensity physical activity or exercise in people with hip or knee osteoarthritis. *Cochrane Database of Systematic Reviews* 2015; **10**: CD010203.

Richards SH, Anderson L, Jenkinson CE, Whalley B, Rees K, Davies P, Bennett P, Liu Z, West R, Thompson DR, Taylor RS. Psychological interventions for coronary heart disease. *Cochrane Database of Systematic Reviews* 2017; **4**: CD002902.

Safi S, Korang SK, Nielsen EE, Sethi NJ, Feinberg J, Gluud C, Jakobsen JC. Beta-blockers for heart failure. *Cochrane Database of Systematic Reviews* 2017; **12**: CD012897.

Santesso N, Carrasco-Labra A, Brignardello-Petersen R. Hip protectors for preventing hip fractures in older people. *Cochrane Database of Systematic Reviews* 2014; **3**: CD001255.

Shepherd E, Gomersall JC, Tieu J, Han S, Crowther CA, Middleton P. Combined diet and exercise interventions for preventing gestational diabetes mellitus. *Cochrane Database of Systematic Reviews* 2017; **11**: CD010443.

Squires J, Valentine J, Grimshaw J. Systematic reviews of complex interventions: framing the review question. *Journal of Clinical Epidemiology* 2013; **66**: 1215–1222.

Stacey D, Légaré F, Lewis K, Barry MJ, Bennett CL, Eden KB, Holmes-Rovner M, Llewellyn-Thomas H, Lyddiatt A, Thomson R, Trevena L. Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews* 2017; **4**: CD001431.

Stroke Unit Trialists Collaboration. Organised inpatient (stroke unit) care for stroke. *Cochrane Database of Systematic Reviews* 2013; **9**: CD000197.

Taylor AJ, Jones LJ, Osborn DA. Zinc supplementation of parenteral nutrition in newborn infants. *Cochrane Database of Systematic Reviews* 2017; **2**: CD012561.

Valentine JC, Thompson SG. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; **4**: 26–35.

Vaona A, Banzi R, Kwag KH, Rigon G, Cereda D, Pecoraro V, Tramacere I, Moja L. E-learning for health professionals. *Cochrane Database of Systematic Reviews* 2018; **1**: CD011736.

Verheyden GSAF, Weerdesteyn V, Pickering RM, Kunkel D, Lennon S, Geurts ACH, Ashburn A. Interventions for preventing falls in people after stroke. *Cochrane Database of Systematic Reviews* 2013; **5**: CD008728.

Weisz JR, Kuppens S, Ng MY, Eckshtain D, Ugueto AM, Vaughn-Coaxum R, Jensen-Doss A, Hawley KM, Krumholz Marchette LS, Chu BC, Weersing VR, Fordwood SR. What five decades of research tells us about the effects of youth psychological therapy: a multilevel meta-analysis and implications for science and practice. *American Psychologist* 2017; **72**: 79–117.

Welch V, Petkovic J, Simeon R, Presseau J, Gagnon D, Hossain A, Pardo Pardo J, Pottie K, Rader T, Sokolovski A, Yoganathan M, Tugwell P, DesMeules M. Interactive social media interventions for health behaviour change, health outcomes, and health equity in the adult population. *Cochrane Database of Systematic Reviews* 2018; **2**: CD012932.

Welton NJ, Caldwell DM, Adamopoulos E, Vedhara K. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *American Journal of Epidemiology* 2009; **169**: 1158–1165.

Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, Clarke M, Gargon E, Gorst S, Harman N, Kirkham JJ, McNair A, Prinsen CAC, Schmitt J, Terwee CB, Young B. The COMET Handbook: version 1.0. *Trials* 2017; **18**: 280.