Complex reviews

Methods and considerations for summarising and synthesising results in systematic reviews with complexity

prepared by

McKenzie JE, Brennan SE

10 October 2014

Table of Contents

Authors and contributors

The following staff of the Australasian Cochrane Centre were involved in the methodological review, and contributed to the preparation of the report.

Joanne McKenzie	Conceived and designed the study, screened a subset of studies for inclusion, led the analysis and the write up of the report.
Sue Brennan	Contributed to the design of the study, designed the overall search methods, screened studies for inclusion, read and coded manuscripts for the analysis, co-led the analysis and write up of the report.
Miyoung Kim	Designed and conducted the citation searches. Retrieved publications.
Steve McDonald	Designed and conducted the MEDLINE search.

Declarations of interest

All authors declare they have no financial, personal or professional interests that could be construed to have influenced the conduct or results of this review.

Executive summary

Background

Systematic reviews synthesise evidence of the effects of healthcare interventions. Complexity in systematic reviews, which may arise from several facets (e.g. diversity of settings, conditions, and outcomes), can provide challenges in applying standard meta-analytical methods. However, without the use of meta-analysis, or another synthesis method, reviews may provide little more than an assembly of available research meeting the inclusion/exclusion criteria of the review, with narrative text that may risk privileging some studies or findings above others without appropriate justification. We therefore aimed to identify summary, synthesis and presentation methods that could be readily applied in reviews with more complexity, and which may aid in better utilisation of available research and fairer representation of the results.

Aims and objectives

- To create an inventory of synthesis and presentation methods that may be readily applied in complex systematic reviews of healthcare interventions.
- To identify the pros and cons of the methods in the inventory, the questions the methods address, the circumstances in which the methods should be considered, key references to the methods, and references to examples of their use.

Methods

We undertook a review of the literature (using three sources) to identify synthesis and presentation methods that may be used in reviews with more complexity. This literature was summarised to provide an overview of the methods and guidance regarding their use.

Findings and recommendations for practice

We identified a range of synthesis methods that can be readily applied in reviews with more complexity. These synthesis methods require different levels of information from the individual studies, they address different questions, and the conclusions and recommendations that can be drawn from the various methods differs.

Key recommendations for practice include the following.

- Systematic reviewers should consider synthesis methods beyond meta-analysis (and its extensions) in circumstances where it is not possible to use meta-analysis. While methods other than meta-analysis provide more limited information, they are still likely to be preferable to a narrative summary of study by study results.
- In systematic reviews with more complexity, reviewers should consider pre-specifying an alternative synthesis method in the review protocol, to cover the circumstance where it is not possible to undertake a meta-analysis.
- Systematic reviewers should clearly acknowledge the limitations of the chosen synthesis method, and the conclusions should reflect these limitations.

Introduction

Systematic reviews synthesize evidence of the effects of healthcare interventions. They aim to make vast bodies of research accessible, in a reproducible manner that fairly represents the sum of knowledge. In systematic reviews that involve a large degree of complexity, achieving this aim can be more challenging.

Complexity in systematic reviews can arise from several facets: complexity of the interventions; diversity of settings, conditions, and outcomes; and different types of study designs. Systematic reviews evaluating interventions of importance to policy (e.g. financing interventions), public health (e.g. mass media interventions for reducing road accidents), and health service delivery (e.g. interventions to improve the safety of care) often have complexity arising in all facets. For example, a systematic review evaluating the effects of continuous quality improvement on professional practice outcomes may include a range of study designs (randomised trials, controlled before after studies, interrupted time series), may include studies in any setting (e.g. studies set in primary care and hospitals), and may encompass any condition, with a resulting diversity of outcomes (e.g. appropriate depression treatment, referred to smoking cessation program, prescribed potassium sparing diuretics). While systematic reviews of clinical interventions also involve complexity, the complexity is often limited to the intervention (e.g. multifaceted behaviour change interventions to increase adherence to dietary advice, psychological and educational interventions for preventing depression in children and adolescents) and diversity of outcomes, since the reviews are often restricted by clinical condition and study design.

Producing systematic reviews that best utilise available research, fairly represent results, and are coherently structured, requires several key elements: organisation and pre-specification of conceptually similar interventions and outcomes (outcome domains), coupled with appropriate synthesis and presentation methods. In reviews with more complexity, these elements are more difficult to achieve because of the diversity of interventions, outcomes, and challenges in applying standard meta-analytical methods. However, without these elements, reviews may provide little more than an assembly of available research meeting the inclusion/exclusion criteria of the review, with narrative text that may risk privileging some studies or findings above others without appropriate justification. In addition, reviews with little or no synthesis leave readers to make sense of the research themselves, which may result in the use of seemingly simple, yet problematic synthesis methods such as vote counting (e.g. counting the number of studies with statistically significant results) [1-3]

In this report, we address the key element of 'synthesis and presentation methods'. Our aim was to identify summary, synthesis and presentation (graphical and tabular) methods that may be readily applied in reviews with more complexity. Our specific objectives were:

- To create an inventory of synthesis and presentation methods that may be readily applied in complex systematic reviews of healthcare interventions.
- To identify the pros and cons of the methods in the inventory, the questions the methods address, the circumstances in which the methods should be considered, key references to the methods, and references to examples of their use.

We present preliminary results of our review, the full findings of which will be reported in a peer reviewed paper and will form the basis for guidance in a new chapter of the next edition of the Cochrane Handbook for Systematic Reviews of Interventions. The findings have been presented at a

recent methods symposium for public health systematic reviews held at the 2014 Cochrane Colloquium, Hyderabad, India (see https://colloquium.cochrane.org/pre-colloquium-symposia-and-workshops-20-22-september-2014).

Methods

Methodological approach and focus

We undertook a review of the literature to identify synthesis and presentation methods that may be used in reviews with more complexity. Because the overall aim of the review was to collate and gain new insights about potential approaches, the overall methodological approach to the review was similar to a scoping review. We used broad selection criteria, took an inclusive approach to ensure all papers of potential relevance to the review were considered, and used iterative search methods to enable us to follow promising leads in more depth (e.g. exploring novel methods or insights from the application of methods in other disciplines).

Our focus was on methods that could be readily applied by review authors in circumstances where there is complex, heterogeneous data (e.g. multiplicity of outcomes, diversity of interventions). We were particularly interested in circumstance where data might be sparse (e.g. few studies with limited replication of outcomes or comparisons). In such circumstances meta-analysis and more sophisticated statistical techniques can be difficult to apply. Our own review of the methods used in a sample of complex reviews and a recent review of NICE public health appraisals [4] suggest that review authors perceive a dichotomy in which the only alternative to meta-analysis is narrative summary. In fact a range of methods can be used in circumstances where meta-analysis is not possible or requires significant specialist knowledge (e.g. Bayesian meta-analysis). We wanted to identify and describe quantitative synthesis and presentation methods that could be used in these circumstances.

Meta-analysis and its extensions (meta-regression, multivariate meta-analysis, network meta-analysis) should be the first choice for quantitative synthesis. Much has been written about these methods, and while we provide information about key decision points when using these methods (question addressed, pros, cons), along with some key papers, we have not focused our attention on these methods.

Similarly, our focus is on providing guidance on the summary and synthesis of quantitative data about the effect of interventions. There is a growing literature about narrative synthesis methods (as opposed to narrative summary) for capturing information about intervention complexity and context (see Pettigrew for an overview [5]). These synthesis methods include qualitative and mixed method approaches. We have not reviewed the literature on narrative synthesis methods.

Criteria for considering publications for this review

Inclusion of publications was based primarily on relevance to addressing our aims. Publications were included if they:

- provided a description of a potentially relevant method,
- contributed to our understanding of the pros and cons of a method, or
- provided key information about the provenance of a method.

We also included examples of the application of a method (including examples of problematic approaches). These were primarily sourced from our review of the methods used in a sample of complex reviews and from citation searches.

Types of publications: We included any published work relevant to our aims. This included tutorials and descriptive reviews of available methods, books and guidance for systematic review authors, empirical studies evaluating methods, and systematic reviews in which one or more of the methods had been used.

Exclusions: We excluded publications on methods that were not suitable for synthesising or summarising quantitative data from evaluations of the effects of an intervention (e.g. methods for synthesising qualitative research). Publications that provided cursory description of a method without providing new insights (for example in the background of a paper that was not about synthesis methods) were excluded. We also excluded publications in languages other than English.

Search methods for the identification of publications

We searched for publications in three databases: Ovid MEDLINE (1946 to May 2014); the Meth4ReSyn library of research synthesis methods publications (www.citeulike.org/search/user/Meth4ReSyn) (all years, database coverage 1980 to 2010), and; the United States Agency for Healthcare Research and Quality (AHRQ) Effective Healthcare Program's Scientific Resource Center (SRC) Methods Library (www.citeulike.org/user/SRCMethodsLibrary) (database commencement to May 2014; database includes publications dating back to 1951). The SRC Methods Library includes publications on all aspects of systematic review and comparative effectiveness review methodologies. We searched for specific methods (e.g. vote counting, combining p-values), as well as more general terms that aimed to identify methods publications that dealt with complexity in reviews (e.g. complex, public health, policy, health services). The MEDLINE search strategy is available in Appendix 1a. The Meth4ReSyn and SRC Methods Libraries have limited search functionality. We, therefore, downloaded all publications from these libraries and conducted searches in Endnote (Appendix 1b).

As a supplement to the above searches, we undertook searches of authors who are known to have published in this area, and examined popular guidance for undertaking systematic reviews [6-10]. We screened reference lists from included publications. We also included relevant articles from our bibliographies, which we denote as 'ad hoc' finds.

To trace the development of methods, and identify examples of use, we conducted citation searches in Scopus or Web of Science for seminal papers (those listed in the results tables in the column "references to methods").

Selection of the publications, data extraction and management

One author (S.E.B) screened titles and abstracts of publications identified through the search. The second author (J.E.M.) screened abstracts for which there was uncertainty about the publication's inclusion.

The full text of all potentially relevant publications was retrieved and imported into NVivo 10 for screening and coding. Our inclusion decisions were iterative. Publications were initially screened for relevance and importance; those that clearly met both criteria were analysed first. Findings from the initial analysis were used to assess the importance of the remaining studies (i.e. those that initially appeared to make a minor or more peripheral contribution) and to identify the need for supplementary searches (to follow promising leads).

We created a hierarchical coding framework to collate all relevant information related to each of the aims of the review. At the top level of the hierarchy we used the following categories (examples of sub-categories are indicated in brackets):

- Type of method (e.g. graphs, plots and visual; meta-analysis and extensions; text; tabular)
- Purpose of the method (e.g. summary, synthesis, exploring heterogeneity)
- Attributes of the method (e.g. description, pros, cons, recommendations for best practice)
- Examples of use (e.g. worked examples presented in tutorial papers including head-to-head comparison of methods, actual examples from systematic reviews)
- General guidance for the summary and synthesis of results in complex reviews (e.g. reasons/arguments for or against synthesising results; implications of different approaches for decision makers)

The framework was developed iteratively, beginning with methods of which we were aware and had searched for specifically. Additional codes were created as we identified new methods.

We also used NVivo 10 to collect information about the included publications as follows:

- discipline (e.g. healthcare or other discipline);
- purpose of the paper (e.g. tutorial; evaluation or critique of method; overview of methods);
- importance of the source (e.g. key sources such as seminal papers or detailed critiques or reviews were categorised as 'high'; cursory descriptions were categorised as 'low', with the potential to revise this categorisation if no comprehensive sources were identified);
- currency (e.g. subjective judgement of whether the methods described are current).

Analysis

Data collated in NVivo10 was used to develop comprehensive summaries of published information about each of the methods, précis of which are tabulated in this report.

Results of the review

Results of the search

Our database searches identified 1002 unique references, 199 of which were included for full text review. A further 49 references were included from citation searching, reference lists and ad hoc sources. All 248 publications were included in NVivo for coding.

Summary and synthesis methods

An overview of the summary, synthesis and presentation methods that we identified and determined could be, or had been, used in systematic reviews with more complexity is depicted in Figure 1. The question each of these methods addresses is also noted. The figure depicts a continuum of approaches with the synthesis approaches (meta-analysis and its extensions) at the right being more preferable than those at the left. In the following sections, each of the methods is expanded, with the pros and cons of the methods described along with references to examples of their use. As our aim was to provide an overview of the methods, we do not provide technical detail of how to apply the methods, but instead provide key references to the methods.

Figure 1: An overview of available for methods for summary and synthesis in reviews with complexity

	Summary	Synthesis				
Methods/ Questions addressed	**Text** **Tabular** Narrative summary of evidence presented in either text of tabular form.	**Combining p-values** "Is there evidence that there is an effect in at least one study?"	**Vote counting** "Is there any evidence of an effect?" **Vote counting extension** "What is the best estimate of intervention effect?"	**Summary of effect estimates (non parametric statistics)** "What is the range and distribution of effects?"	**Meta-analysis** "What is the average intervention effect?" (random effects) **Predictive intervals** "What is the potential effect of an intervention in an individual study?"	**Exploring heterogeneity** **Subgroup analysis/meta-regression** "What factors modify the magnitude of the intervention effects?"
Plots	**Forest plots (plotting individual studies without a pooled estimate)**		**Harvest plots** **Effect direction plots**	**Box and whisker plots** **Stacked bar plots**	**Forest plots**	**Forest plots** **Box and whisker plots** **Bubble plots**

Summary: Text and tabular summary of intervention effects

In a review of synthesis methods used in NICE public health appraisals, Archana and colleagues found that nearly 80% of appraisals did not attempt a quantitative synthesis relying only on narrative summary [4]. Narrative and tabular summaries of findings are essential in any systematic review. Yet used in the absence of quantitative synthesis methods, these approaches may provide little more than an assembly of studies, leaving decision makers to make sense of the findings themselves. Table 1 describes the limitations of summarising rather than synthesising findings and the problems that may arise as a consequence.

Table 1. Summary: text and tabular

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
Summary	Text	***"Narrative summary of the evidence from individual studies"*** Results of studies summarised in the text of a publication without the use of a quantitative synthesis method (i.e. a narrative summary). Results are typically presented study by study. "Narrative summary typically includes the selection, chronicling, and ordering of evidence to produce an account of the evidence. Its form may vary from the simple recounting and description of findings through to more interpretive and explicitly reflective accounts that include commentary and higher levels of abstraction" [11]. Conceptual frameworks or logic models may be useful tools for structuring narrative summaries [1 12 13].	Provides an assembly of the available research meeting the inclusion/exclusion criteria [14].	Results summarised, not synthesized. Reviews with poorly structured results or a large number of comparisons / outcomes may selectively report (privileging some findings above others) or overemphasise some results. Interpretation of results across studies is difficult/not possible[4]. "… narrative synthesis depends substantially on using text to 'tell the story'[15]. "If the number of included studies is large, this can result in a lengthy and somewhat indigestible results section …" [14]. Readers may use inappropriate ad hoc rules to make sense of the results (e.g. counting the number of statistically significant results) [16].	[11 15]	*Examples of selective reporting:* Ko 2010 reports effects and confidence intervals for statistically significant outcomes only. Authors state they have undertaken a 'narrative synthesis' [17] ter Wee 2012 reports results for all outcomes, but only the effect estimate and its CI/p-value are reported in the text when there is a statistically significant effect [18]. Schouten 2008 only reports statistically significant effects in full in the text and in the tables [19].

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
	Tabular	***"Tabular summary of the evidence from individual studies"*** Provides a structured method for presenting data. E.g. comparison, outcome (professional performance (adherence to recommended practice), patient outcome), study design, potential effect modifiers.	More likely to report all results of all outcomes (i.e. may be less likely to selectively include results). Results available for others to synthesize.	Results summarised, not synthesized. Overwhelming amount of information which is difficult for a reader to interpret (often multiple outcomes per study).		*Tables that enhance interpretation of results by enabling comparison across studies:* Giguère 2012 tabulates results by comparison and outcome category (tables 4-10) [20] Marteau 2010 tabulates results by outcome category (table 1) [21] *Tables that enable interpretation of results for each study, but not comparison across studies:* ter Wee *Ann* 2012 reports all available results, study by study, for three outcome domains (Table 3 results from RCTs) [18] Boonyasai 2007 reports results, study by study, for outcomes in each outcome category (significant and non-significant) (eTable2) [22]

Synthesis: Combining p-values

Combining p-values is a crude synthesis approach with a long history [23]. This approach has little in the way of redeeming features since it provides no information on the magnitude of intervention effect, and the hypothesis tested (whether there is an intervention effect in at least one study) is of limited practical interest (Table 2). Common reasons for using this method include impoverished reporting of results (no or little information reported beyond p-values), and where non-parametric analyses have been performed in the individual studies [24].

Table 2. Combining p-values

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
Combining p-values	n/a	***"Is there evidence that there is an effect in at least one study?"*** Combines p-values across studies (e.g. using the method of Fisher 1932 or Stouffer) [23]. Tests the hypothesis that there is no effect in every study (i.e. the null hypothesis in every study is true). Alternative hypothesis is that there is an effect in at least one study (i.e. the alternative hypothesis in at least one study is true) [24]. Several methods for combining p-values are available. Extensions that adjust for correlated p-values (as arise when there are multiple p-values from the same study) are available [25].	May be used in circumstances where (i) studies report results from non-parametric analyses; (ii) no or little information is reported beyond p-values; (iii) outcomes are different across studies (e.g. different serious side effects across studies); (iv) the statistical tests differ across the studies (as long as the same hypothesis is tested); (v) there is diversity in study populations such that a pooled effect size is not interpretable, but still meaningful to test if there is an effect in at least one study [23 24].	Does not provide information on the size of effect [24]. Difficult to interpret the result of the test since there may only be an effect in one study [24]. Cannot say how many studies there was an effect in. Sample sizes of the studies will influence the power of the test. Therefore, when combining p-values from small studies, failure to reject the null hypotheses needs to be interpreted carefully.	[23-25]	Cucherat 2000 meta-analysed p-values to answer the question of whether there was evidence that homeopathic treatments were effective in patients with any condition [26]. More specifically, the authors state *"*Using this approach [combining p-values], the null hypothesis tested is that the effect of interest (in this case, the efficacy of homeopathic treatment) is not present in any of the trials considered. If the null hypothesis is rejected, we can conclude that in at least one trial there is a non-null effect. This point is crucial and, unfortunately, the results from analysis using this approach are often misinterpreted. We cannot say in how many trials homeopathy is efficacious and we cannot estimate the size of the effect. The major advantage of this approach is that P values from any statistical test for the hypothesis of interest can be combined. If the results are interpreted with sufficient precaution, this approach provides a way to combine results from very dissimilar trials with differing outcomes and statistical tests.*"* (pg 28 [26]).

Synthesis: Vote counting

Vote counting has been described as "a last resort in situations when standard meta-analytical methods cannot be applied" [27]. However, it is not uncommon that reviews with no or very few meta-analyses use vote counting informally to summarise results across studies (e.g. 3 of 5 studies found ...). In the absence of any quantitative synthesis, readers may vote count themselves. These informal and *ad hoc* approaches are problematic because they lead to erroneous conclusions. If review authors anticipate that meta-analysis may not be possible, and vote-counting might be used, then a better approach is to pre-specify appropriate vote-counting methods. Similar to combining p-values, vote counting methods would be considered in circumstances where there was impoverished reporting of results within the studies. Specifically, in the circumstance where no estimates of intervention effect are reported, but the direction of the effect is available (i.e. whether the intervention is beneficial or harmful). An extension to conventional vote counting is available that can yield an estimate of pooled intervention effect [28].

Table 3. Vote counting and its extensions

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
Vote counting	Text/ Tabular	***"Is there any evidence of an effect?"*** [27] Compares the number of studies showing harm vs the number of studies showing benefit (regardless of statistical significance or size of results) [1 27]. As with other meta-analytical approaches, methods need to be pre-specified to select a single outcome from studies that report multiple within an outcome domain (e.g. multiple measures of pain) so as not to give studies differential weighting on the basis of the number of outcomes collected. Sign test (non-parametric test) used to assess the statistical significance of evidence for the existence of an effect in either direction. Under the null hypothesis of no true intervention effect (e.g. is the average effect size zero?), we would expect the proportion of studies with effect estimates favouring the intervention (or alternative) to be 50%. Under the alternative hypothesis, of an intervention effect, we would expect the proportion of studies with favourable	Provides a method for synthesizing effects in circumstances when (i) standard meta-analytical methods are difficult to apply (e.g. inconsistent outcomes, no appropriate estimates of precision available (not able to adjust for unit of analysis errors); different effect measures (mean difference, odds ratio)); or (ii) there is diversity in study populations such that pooled effect size is not interpretable, but it is still meaningful to test if there is on average an effect [23 27 29]. Requires a minimal amount of statistical data. [30] Simple to apply [24].	Provides no information on the magnitude of the effects of the intervention (only information on the direction). Problematic if subjective decisions or statistical significance are used to define the number of positive and the number of negative studies [27]. No account of the differential weighting across studies (i.e. ignores the precisions of the studies) [27]. Vote counting of statistically signigicant results also suffers from problems with unit of analysis errors and underpowered studies (those with point estimates which are clinically	[1-3 23 27-30]	*Examples of methods that are not recommended:* Ioannidis 2008 cites an example where the authors used their own "home made" rules to indicate effectiveness [3]. ter Wee 2012 use vote counting, but it is not clear what constitutes a "positive", "conflicting", or "negative" categorisation because it is not defined in the paper (table 4) [18]. Schouten 2008 use vote counting within outcome categories for each study (e.g. 1 of 3 process of care outcomes were 'significant') (Table 2). There is no synthesis across studies.[19] *Examples of vote counting based on direction of effect (recommended):* Flodgren 2011[35] counts the number of outcomes favouring the intervention regardless of statistical significance or size of results. However, they do not adjust for the issue that for a particular outcome domain, studies contribute a different number of outcomes, and therefore will inappropriately receive differential weighting just on the basis of the number of outcomes collected (Table 6).

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
		effect estimates to be greater than 50% (e.g. is the average effect non-zero?) [23 28 29]. The proportion of studies with favourable results along with a 95% confidence interval for this proportion can be calculated [28].		important, but not statistically significant) [3 31 32]. Vote counting based on statistical significance does not identify the underlying truth as the number of studies increases (not consistent) [30 33 34].		
Vote counting extension	Pooled effect	***"What is the best estimate of the intervention effect?"*** ***"Is there evidence of an effect?"*** A method to estimate the standardised mean difference (SMD) from vote counts. Estimates of SMD can be calculated from scenarios where studies either report (i) the direction of the effect or (ii) the direction of effect and statistical significance. Pooled estimates of SMD from these approaches can be combined with meta-analytic SMDs (calculated from effect estimates and standard errors). Adaption to the method needs to be used when the results are all in one direction [28].	Can combine studies reporting different degrees of information. When using meta-analysis to pool only available effect estimates, studies without this information are excluded. Unlike vote counting based on stat. sig., this approach will identify the underlying truth (i.e. the estimator is consistent) as the number of studies increases. The assumptions required for using this approach are not too strict. Weights the studies according to the information they provide (studies' precisions).	Complex method to apply [28 36]. Based on fixed effect model (assumes the effect sizes are homogeneous), so does not provide information about the potential heterogeneity of effects across studies [28 36]. Assumes that sample sizes are large enough so that the effect estimates are normally distributed [28 36]. Requires the same effect measure to have been used across the studies (e.g. difference in means, difference in SMD) [28 36].	[28 36]	None identified from citation search.
	Harvest plot	Visual plot of vote counting results (hypothesis testing approach) to display the "distribution of evidence" [14]. Harvest plot groups studies based on whether they demonstrate a positive, negative, or no effect. Plot can be 'visually' weighted and annotated to highlight study characteristics, e.g. risk of bias domains (e.g. allocation concealment), proximal vs distal outcomes, study design. [14 37]	May help make sense of a lot of data quickly. More fairly represents the results compared with a narrative synthesis. Helpful for identifying areas where there is a lack of research (e.g. either poor quality or no studies). Similarly, helps identify compelling evidence and	As above with conventional vote counting. In addition, colouring and height of bars may disproportionately draw the readers' eye to studies which should not receive so much 'visual' weight. Plots help summarise the results, but problematic if interpreted as providing a	[14 37]	[38-40]

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
			'deviant' cases. [14] Helps focus narrative discussion.	definitive statement about the results of the review. [14]		
	Effect direction plot	Visual display of the direction of the effects for each outcome within a study, or the direction of effects for outcome domains (categories) across studies. Arrows are used to indicate the reported direction of effect (improvement, deterioration, no change/conflicting results). An indication of study size and statistical significance incorporated (using size and colour). The version of the plot that depicts the direction of effects for outcome domains (when there are multiple outcomes per domain) uses a process to synthesise results based on the proportion of results in a consistent direction and statistical significance.	Conveys a large amount of complex information and allows comparison across studies, outcome domains, as well as other characteristics (e.g. study quality, design). Provides a clear visual structure for presenting the results. Aims to improve the accessibility and transparency of the written summary. Rules for determining the look of the arrows are pre-specified (e.g. size and colour of the arrows).	Complex plot to interpret (has many footnotes). Does not convey information about effect estimates for the studies that may have these available. May lead viewers to vote count based on statistical significance, given the arrows denoting statistical significance are a dominant colour. Not clear that the synthesis approach to combine the direction of effects across outcomes within a domain has performance validity.	[41]	[42]

Synthesis: Summary of effect estimates (non-parametric statistics)

The estimates of intervention effect may be described using non-parametric statistics such as the median, interquartile range, and range. These statistics provide information on the typical estimates of intervention effect and their distribution. This synthesis approach is less preferable than meta-analysis for the reasons described in Table 4, but may be considered in circumstances where estimates of intervention are available, but variances of estimates are not (and cannot be calculated from other statistics, or reasonably imputed). Note that this method has some advantage compared with conventional vote counting and combining p-values in that it provides information on the magnitude of intervention effects.

Table 4. Summary of effect estimates

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
Summary of effect estimates	Non-parametric statistics	***"What is the range and distribution of effects?"*** [2] The median effect size across studies is calculated. Other statistics such as the interquartile range (IQR) and range are also calculated. In studies that have multiple outcomes measuring the same underlying construct (e.g. health practitioner adherence to recommended practice); an approach may be used to select one outcome per outcome domain (category). The selection process should be set up *a priori* to avoid selective inclusion of results. Possible processes include (i) selecting the primary outcome defined in the study; (ii) panel rates importance of outcomes independent of results; (iii) the outcome with the median effect is included in the absence of specification of the primary outcome (agnostic approach). The chosen effect measure is used to 'characterise' the outcome of the study [1]. When the outcome with the median effect is chosen, and median of these median effects is calculated, this approach is referred to as the 'median-of-medians' approach.	Provides a method for synthesizing results when difficult to undertake a meta-analysis (e.g. missing variances of effects). Provides information on the magnitude of effects. Use of medians to summarise across studies may be beneficial (compared with calculating an average) to minimise the impact of publication bias (small studies with large effects more likely to be published), since the median is robust against outliers [10].	Standard calculation of non-parametric statistics does not weight effects; small studies are as influential as large studies. Weighted non-parametric statistics can be calculated, but require a priori decisions about how to weight the effects (e.g. study sample size [10]). Such approaches may be problematic. Range and IQR of effects may be explained by biases (e.g. studies at a high risk of bias).	[1 2 10 31 43]	Shojania 2011 [44] and Arditi 2012 [45] both present median effect size with IQR. They use pre-specified methods to select one outcome per outcome domain and use the median of medians approach.

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
	Box and whisker plots	Visual display of the distribution of effects. Displays median, Inter-quartile range (IQR), and range.	Plot is in common usage, facilitating rapid and correct interpretation [46].	Standard use of the plot does not account for the weights of the studies. A box plot of weighted estimates is available [47]. Does not account for uncertainty in estimates [48].	[47 48]	Ivers 2012 presents box plots to visually display the distribution of effects for each of the comparisons included in the review (figure 10) [49]
	Stacked bar plots	Plot depicts distribution and range of effects with the stacks distinguishing results which are statistically significant, statistically non-significant, and have a unit of analysis error.	Provides information regarding statistical significance of the effects compared with the box and whisker plot.	Standard use of the plot does not account for the weights of the studies. Does not account for uncertainty in estimates [48].	[31]	Grimshaw 2004 presents results for dichotomous process measures in stacked bar plots [31]

Synthesis: Meta-analysis and its extensions

Meta-analysis is a statistical analysis of the results of several independent studies leading to a quantitative summary [50]. Meta-analysis has many benefits compared with other synthesis approaches (Table 5), and importantly, provides decision makers with information about the likely magnitude of the effects of an intervention (and certainty in the estimates of these effects). This information allows decision makers to make informed choices about whether to fund, recommend, or use interventions. For this reason and those noted in Table 5, meta-analysis is the method of choice and systematic reviewers should aim to use this approach whenever possible. A commonly used and simple meta-analysis approach only requires an estimate of intervention effect and its variance from each included study [27]. While the variance of the estimated effect may not be directly reported in a journal article, it can often be calculated from other reported statistics.

Commonly reported reasons for avoiding meta-analysis include concerns of too much diversity across interventions, settings, participants, outcomes, metrics, and study designs [3]. Extensions of meta-analytical methods such as prediction intervals, subgroup analysis, and meta-regression can be used to address some of these concerns (Table 5). Subgroup analysis and meta-regression allow quantification of the degree to which factors may modify the magnitude of the intervention effect, thus, potentially explaining the observed statistical heterogeneity. Results of subgroup and meta-regression analyses may be of particular interest to policy makers since the results may provide an indication of whether certain elements of an intervention may be more effective, or whether the setting/context in which the intervention was delivered may impact on its effectiveness.

Table 5. Meta-analysis, meta-analytic extensions, exploring heterogeneity

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
Meta-analysis	Meta-analysis	***"What is the average intervention effect?"*** (random effects model) ***"What is the best estimate of the intervention effect?"*** (fixed effect model) Meta-analysis is a statistical analysis of the results of several independent studies leading to a quantitative summary [50]. The pooled effect is a weighted average of the effects from the component studies. The weights differ depending on the particular meta-analysis method used. Two models are the fixed effect and random effects models. These models have different inference goals. Fixed effect model (conditional model):	Allows interpretation of the effectiveness of the intervention that is difficult, if not impossible, to achieve without meta-analysis (particularly as the number of studies increases) [4]. Provides an estimate of the intervention effect (average or best estimate), and in addition, appropriately adjusts for the amount of information each study provides. Increases power and	Meta-analysis can be misleading when specific study designs, within-study biases, variation across studies, and reporting biases are not considered [4 27] A pooled intervention effect may have limited meaning when a large degree of diversity exists across the interventions such that an average effect is not meaningful [4]. For example "… interventions	[27 50-53]	Akl 2011 illustrates the use of outcome categorisation and methods to combine multiple effects from the same outcome domain, to synthesise effects (and make good use of available evidence).[54] Smedslund 2011 illustrates the use of outcome categories to enable synthesis of effects [55]

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
		provides an estimate of intervention effect for the studies included in the meta-analysis, or studies that are sufficiently similar to those in the meta-analysis. Random effects model (unconditional model): provides an average estimate of intervention effect for the universe of studies from which the sample of studies was obtained [51].	improves precision, allowing a question to be answered that cannot be answered by the individual studies [27]. Provides a basis for "exploring heterogeneity and attempting to account for it" [4].	to increase the uptake of safety practices varied between studies (e.g., interventions ranged from educational initiatives, through vouchers to reduce the price of equipment, through to the free provision and fitting of equipment), and therefore, by fitting the data into a meta-analysis framework of ''intervention'' vs. ''usual care,'' the interpretation of the resulting pooled effect was unclear - exactly what does the pooled effect relate to?" [4]		
	Predictive intervals	"What is the potential effect of an intervention in an individual study?" Provides a predicted range for the true effects in an individual study. Complements information provided by random effects meta-analysis.	Particularly helpful when it is not possible to identify factors predictive of heterogeneity. May be clearly used to demonstrate the inappropriateness of relying on the estimate of average effect.	Problematic when calculated from studies at a high risk of bias since it will encompass heterogeneity introduced by these biases, in addition to heterogeneity caused by factors of interest (e.g. intensity of intervention) [56].	[56 57]	See Riley 2011 for examples [56]
	Multivariate meta-analysis	An extension of standard meta-analysis that allows for correlation between effect estimates (as arises when multiple outcomes are measured per study) [53].	Allows inclusion of multiple effect estimates from a single study in a meta-analysis, such as estimates from multiple intervention arms from the same trial or outcomes measured at different time points [5].	Complex method to apply. "Requires information about within study correlations" [5]	[53 58]	Del Re 2013a and 2013b provide examples of multivariate meta-analyses [59 60]
	Forest plots	Displays effect estimates and confidence intervals for each study. Can display the meta-analytic effect or not. The area of the block representing the point estimate for each study is proportional to the	Provides a familiar and efficient method of presenting effects and their confidence intervals. This familiarity may facilitate	Requires variances of the effect estimates. Lines depicting confidence intervals may be	[46-48 61 62]	See Schriger 2010 [46] (figure 1) for a visual depiction and explanation of all elements of a forest plot.

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
		weight the study receives in the meta-analysis. Studies can be ordered by date or characteristics that might explain heterogeneity (e.g. risk of bias, population, intervention duration) to convey maximal information [46 61]. "Summary forest plots" can be used to depict the results of multiple meta-analyses (rather than the results of multiple studies), which can be used for displaying different outcomes on the same plot , for example, or may have utility in overviews of systematic reviews [48].	faster and more accurate interpretation [46 48]. The variation in symbol size appropriately draws attention to the studies that contribute the most information [47].	interpreted incorrectly as if all points along the line are equally likely estimates of the intervention effect. Variants of the standard forest plot, with tapering ends, have been developed to give greater prominence to estimates toward the middle, which are more likely [48 61]. Small studies may be more prominent because of the length of the confidence intervals [61], but this is countered by using a block for the point estimate that is proportionate to the study's weight [47].		
Exploring heterogeneity	Sub-group analysis	***"What factors modify the magnitude of the intervention effects?"*** Explores factors that may modify the size of the intervention effects (discrete factors). A hypothesis test to investigate if there are differences between two or more subgroups. Subgroups are typically participant (e.g. sex), trial (e.g. risk of bias), intervention (e.g. complexity of intervention), or study (e.g. individually vs cluster randomised) characteristics.	Provides hypotheses regarding what (set) of factors might be necessary for the intervention to be effective. Necessary information for decision makers. May provide leads for future research.	Observational analysis, so potentially suffers from confounding bias [27]. Factors found to modify intervention effects may not be causative, but instead associated with the factor and outcome [27]. False negative and false positive significance tests are likely to rapidly increase as the number of subgroup analyses undertaken increases [27]. Careful interpretation of the results are required.	[27 63]	Achana 2014 provides a worked example of how meta-analysis and subgroup analyses were used to explore heterogeneity in a public health review [4]. The authors also illustrate how meta-analytic extensions can be used to address limitations of traditional meta-analytic techniques.
	Meta-regression	Explores factors that may modify the size of the intervention effects (discrete and continuous factors). Can be used to investigate components ('active ingredients') of multifaceted	Provides hypotheses regarding what (set) of factors might be necessary for the intervention to be effective. Necessary information for decision	Observational analysis (weaker interpretation compared to effects observed in randomised trials); may suffer from confounding bias and	[1 2 31 64-69]	O'Brien 2007 provides an example of meta-regression with multiple covariates in a review of educational outreach for improving professional practice. The model includes baseline measures of the outcome (compliance)] [70]. Thompson 2002 provides an example of the use of

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
		interventions that may modify effects. Extensions are available that allow for correlation across effect estimates (as may arise when (i) multiple outcomes are measured on participants (e.g. depression using different measurement scales),(ii) outcomes measured on participants at multiple time points, (iii) multiple treatment groups, where effects of treatment are estimated against a common comparator (e.g. placebo)) [64 65] Meta-regression has been used to model the median effects from each study, weighting by, for example, the number of practitioners [2].	makers. Allows investigation of what components of a set of interventions may be more effective. These components may then be examined in future trials (e.g. using factorial designs).	aggregation bias (ecological fallacy) [66]. Small number of studies with multiple factors leads to overfitting and spurious claims of association. Permutation test approaches have been developed for assessing the true statistical significance of an observed meta-regression result [67 68]. In practice meta-regression investigating which components modify the effects of an intervention is difficult when there are many components. Problems occur from the need to fit interaction terms (can't assume additive effects); too few studies; combinations of components highly correlated (e.g. some components always occur together) [31]. Practical limitation also occur when there is no measure of variance of the effect, or no data on the factor [66]. Reviewers are often interested in examining the relationship between effect sizes for an adherence outcome and a baseline measure of adherence, which poses technical problems		a bubble plots to display results of a meta-regression [66]. Gardner 2010 provides an example of meta-regression using behaviour change theory to categorise intervention components, and investigates if these components modify the effects of the intervention Michie 2009 provides another example of meta-regression where interventions are classified using behaviour change theory, in this case looking at interventions to promote healthy eating and physical activity [71].

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
				(regression to the mean).		
	Other statistical approaches	Kruskal-Wallis one way ANOVA (or Mann-Whitney U test for two groups) Non-parametric method for testing if the ranks of the effects differ by potential modification factors (categorical) [31].	Provides a more objective measure (compared with visual inspection) of whether the factor modifies the magnitude of effects. Can be undertaken when it is not possible to undertake a meta-regression (e.g. missing variances of effects).	Provides a p-value for the association but no certainty in the difference in medians between groups. Suffers from some of the same issues described under meta-regression (observational, confounding, small number of studies, aggregation bias).	[1 31]	Arditi 2012 examined whether features of a clinician reminder (e.g. inclusion of advice, delivered at point of care) influenced the distribution of intervention effects. Mann-Whitney tests were used to determine whether there were statistically significant differences in the median effect size (IQR) with/without each of the features. Results are presented graphically. [45]. Shojania 2009 provides a similar example in which different study features (e.g. study design, setting, intervention duration) were examined. Kruskall-Wallis and Mann-Whitney tests were used to determine whether there were statistically significant differences in the median effect sizes with/without each of the features. Results are presented graphically [44].
	Forest plots	Displays effect estimates and confidence intervals for each study. The confidence intervals help indicate the extent of heterogeneity [48]. Can display the meta-analytic effect or not. The area of the block representing the point estimate for each study is proportional to the weight the study receives in the meta-analysis. Studies can be grouped based on characteristics that might explain heterogeneity (e.g. risk of bias, population, intervention duration) [46 61], revealing patterns not otherwise apparent [46].	Provides a familiar and efficient method of presenting effects and their confidence intervals. This familiarity may facilitate faster and more accurate interpretation [46 48]. Visual inspection of forest plots can aid in identifying the presence or absence of heterogeneity [47], especially when studies are ordered by effect size [46]	Ordering of studies alphabetically by author diminishes the potential of forest plots to reveal heterogeneity and its sources [46] Visual inspection of plots to assess heterogeneity is best done in combination with other methods for assessing heterogeneity, where these are applicable.	[47 48 62] Also: [46 61 72]	See Schriger 2010 [46] (figure 1) for a visual depiction of use of subgroups in a forest plot to aid visual inspection of heterogeneity.
	Box and whisker plots	Visual displays of the distribution of effects by the potential modifying factors.	Plot is in common usage, facilitating rapid and correct interpretation [46].	Standard use of the plot does not account for the weights of the studies. A box plot of weighted estimates is available [47].	[47]	Ivers 2012 uses box and whisker plots to visually display the distribution of effects of audit and feedback for health professionals in the presence of different modifying factors, such as attributes of the feedback (e.g. delivered by colleagues, frequency of feedback) [49]. Giguere 2012 is a similar example, looking at the effects of health professional educational materials [20].

Method	Modes	Questions answered/ Description of method	Pros	Cons	References to methods	References to examples of use
	Bubble plots	Graphical presentation of the association between the effect size of an intervention and potential modifying factors (e.g. components of the intervention, intervention duration, setting). Bubbles plots are a form of scatter plot with the effect size on the vertical axis and potential modifying factor on the horizontal axis. The precision of each effect estimate is represented by the size of the bubble (in EPOC reviews, the size of the bubble sometimes represents the number of healthcare professionals [70 73]). Studies can be denoted by numbers. Variations of the bubble plot are available [48].	Facilitates understanding of the relationship between the effect of the intervention and potential modifying factors. Can convey multiple levels of information (e.g. relationship, precision of effects). Multiple factors can be plotted (e.g. multiple lines representing categorical factors) [48].	Individual studies cannot be easily identified if numbers are not used to label studies [48]	[2 48 73]	O'Brien 2007 uses a bubble plot to visually depict the relationship between the effect of educational outreach interventions and the clinical behaviour targeted by the intervention (prescribing or other behaviours) [70]. Other examples are given in Anzures-Cabrera 2010 [48] and Grimshaw 2003 [2].

Conclusion

Our review of the literature has identified that a range of synthesis approaches exist which can readily be applied in systematic reviews with additional complexity. While meta-analysis and its extensions offers many benefits, and should be the first synthesis choice whenever possible, there are circumstances where these methods cannot be applied (e.g. incomplete reporting of results in the studies). In these circumstances, the use of other synthesis methods is likely to be preferable to only providing a narrative summary of study by study results, with the risk that some studies or findings will be promoted above others without appropriate justification. When using other synthesis methods, systematic reviewers need to be aware of the questions these methods address, acknowledge the limitations of the methods, and have appropriately cautious conclusions. Regardless of the synthesis approach adopted, we recommend that systematic reviewers pre-specify the proposed synthesis methods in their review protocol.

References

1. Bravata DM, McDonald KM, Shojania KG, et al. Challenges in systematic reviews: Synthesis of topics related to the delivery, organization, and financing of health care. Annals of Internal Medicine 2005;**142**(12 II):1056-65

2. Grimshaw J, McAuley LM, Bero LA, et al. Systematic reviews of the effectiveness of quality improvement strategies and programmes. Quality & safety in health care 2003;**12**(4):298-303

3. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. BMJ 2008;**336**(7658):1413-15 doi: citeulike-article-id:13189227 doi: 10.1136/bmj.a117[published Online First: Epub Date]|.

4. Achana F, Hubbard S, Sutton A, et al. An exploration of synthesis methods in public health evaluations of interventions concludes that the use of modern statistical methods would be beneficial. Journal of Clinical Epidemiology 2014;**67**(4):376-90 doi: http://dx.doi.org/10.1016/j.jclinepi.2013.09.018[published Online First: Epub Date]|.

5. Petticrew M, Rehfuess E, Noyes J, et al. Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. Journal of Clinical Epidemiology 2013;**66**(11):1230-43 doi: 10.1016/j.jclinepi.2013.06.005[published Online First: Epub Date]|.

6. Higgins J, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester (UK): John Wiley & Sons, 2008.

7. NICE. *Methods for the development of NICE public health guidance.* 3rd ed. ed, 2012.

8. EPPI-Centre T. EPPI-Centre Methods for Conducting Systematic Reviews: Updated 2010. London: University of London, 2007.

9. Centre for Reviews and Dissemination. Systematic Reviews: CRD's guidance for undertaking reviews in health care. York: University of York, 2008.

10. Shojania KG, McDonald KM, Wachter RM, et al. Volume 1—Series Overview and Methodology (AHRQ Publication No. 04-0051-1). Closing the Quality Gap: A Critical Analysis of Quality Improvement Strategies Rockville: Agency for Healthcare Research and Quality, 2004.

11. Dixon-Woods M, Agarwal S, Jones D, et al. Synthesising qualitative and quantitative evidence: a review of possible methods. Journal of Health Services Research & Policy 2005;**10**(1):45-53 doi: citeulike-article-id:13190565[published Online First: Epub Date]|.

12. Anderson LM, Oliver SR, Michie S, et al. Investigating complexity in systematic reviews of interventions by using a spectrum of methods. Journal of clinical epidemiology 2013;**66**(11):1223-29 doi: citeulike-article-id:13192897 doi: 10.1016/j.jclinepi.2013.06.014[published Online First: Epub Date]|.

13. Anderson LM, Petticrew M, Rehfuess E, et al. Using logic models to capture complexity in systematic reviews. Research Synthesis Methods 2011;**2**(1):33-42 doi: citeulike-article-id:13194387 doi:10.1002/jrsm.32[published Online First: Epub Date]|.

14. Ogilvie D, Fayter D, Petticrew M, et al. The harvest plot: a method for synthesising evidence about the differential effects of interventions. BMC Medical Research Methodology 2008;**8**:8 doi: citeulike-article-id:13189378 doi: 10.1186/1471-2288-8-8[published Online First: Epub Date]|.

15. Popay J, Roberts H, Sowden A, et al. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme, 2006:1-92.

16. Valentine JC, Pigott TD, Rothstein HR. How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis. Journal of Educational and Behavioral Statistics 2010;**35**(2):215-47 doi: 10.3102/1076998609346961[published Online First: Epub Date]|.

17. Ko H, Turner T, Jones C, et al. Patient-held medical records for patients with chronic disease: a systematic review. Quality & safety in health care 2010;**19**(5):e41 doi: 10.1136/qshc.2009.037531[published Online First: Epub Date]|.

18. ter Wee MM, Lems WF, Usan H, et al. The effect of biological agents on work participation in rheumatoid arthritis patients: a systematic review. Annals of the rheumatic diseases 2012;**71**(2):161-71 doi: 10.1136/ard.2011.154583[published Online First: Epub Date]|.

19. Schouten LMT, Hulscher MEJL, Everdingen JJEv, et al. Evidence for the impact of quality improvement collaboratives: systematic review. BMJ 2008;**336**(7659):1491-94 doi: 10.1136/bmj.39570.749884.BE[published Online First: Epub Date]|.

20. Giguere A, Legare F, Grimshaw J, et al. Printed educational materials: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012;**10**:CD004398 doi: 10.1002/14651858.CD004398.pub3[published Online First: Epub Date]|.

21. Marteau TM, French DP, Griffin SJ, et al. Effects of communicating DNA-based disease risk estimates on risk-reducing behaviours. Cochrane Database Syst Rev 2010(10):CD007275 doi: 10.1002/14651858.CD007275.pub2[published Online First: Epub Date]|.

22. Boonyasai RT, Windish DM, Chakraborti C, et al. Effectiveness of teaching quality improvement to clinicians: a systematic review. JAMA : the journal of the American Medical Association 2007;**298**(9):1023-37 doi: 10.1001/jama.298.9.1023[published Online First: Epub Date]|.

23. Borenstein M, Hedges LV, Higgins JPT, et al. Meta-Analysis Methods Based on Direction and p-Values. Introduction to Meta-Analysis: John Wiley & Sons, Ltd, 2009:325-30.

24. Abrams K, Jones DR. Meta-analysis and the synthesis of evidence. Mathematical Medicine and Biology 1995;**12**(3-4):297-313 doi: citeulike-article-id:13191982 doi: 10.1093/imammb/12.3-4.297[published Online First: Epub Date]|.

25. Alves G, Yu YK. Accuracy evaluation of the unified p-value from combining correlated p-values. PloS one 2014;**9**(3):e9122 doi: citeulike-article-id:13187510 doi: 10.1371/journal.pone.0091225[published Online First: Epub Date]|.

26. Cucherat M, Haugh MC, Gooch M, et al. Evidence of clinical efficacy of homeopathy. A meta-analysis of clinical trials. HMRAG. Homeopathic Medicines Research Advisory Group. European journal of clinical pharmacology 2000;**56**(1):27-33

27. Deeks JJ, Higgins JPT, Altman DG. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]: The Cochrane Collaboration, 2011.

28. Bushman BJ, Wang MC. Vote-counting procedures in meta-analysis. The Hand. of Res. Synthesis and Meta-Analysis, 2nd Ed., 2009:207-20.

29. Borenstein M, Hedges LV, Higgins JPT, et al. Vote Counting – A New Name for an Old Problem. Introduction to Meta-Analysis: John Wiley & Sons, Ltd, 2009:251-55.

30. Hedges LV, Olkin I. Vote-counting methods in research synthesis. Psychological Bulletin 1980;**88**(2):359-69

31. Grimshaw JM, Thomas RE, MacLennan G, et al. Effectiveness and efficiency of guideline dissemination and implementation strategies. Health Technology Assessment 2004;**8**(6):iii-iv, 1-72 doi: citeulike-article-id:13190619[published Online First: Epub Date]|.

32. Verbeek J, Ruotsalainen J, Hoving JL. Synthesizing study results in a systematic review. Scandinavian Journal of Work, Environment and Health 2012;**38**(3):282-90

33. Combs JG, Ketchen, Crook TR, et al. Assessing Cumulative Evidence within 'Macro' Research: Why Meta-Analysis Should be Preferred Over Vote Counting. Journal of Management Studies 2011;**48**(1):178-97 doi: citeulike-article-id:13194601 doi: 10.1111/j.1467-6486.2009.00899.x[published Online First: Epub Date]|.

34. Friedman L. Why vote-count reviews don't count. Biological Psychiatry 2001;**49**(2):161-62 doi: citeulike-article-id:10949778 doi: 10.1016/s0006-3223(00)01075-1[published Online First: Epub Date]|.

35. Flodgren G, Eccles MP, Shepperd S, et al. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. Cochrane Database Syst Rev 2011(7):CD009255 doi: 10.1002/14651858.cd009255[published Online First: Epub Date]|.

36. Bushman BJ, Wang MC. A Procedure for Combining Sample Standardized Mean Differences and Vote Counts to Estimate the Population Standardized Mean Difference in Fixed Effects Models. Psychological methods 1996;**1**(1):66-80

37. Crowther M, Avenell A, MacLennan G, et al. A further use for the Harvest plot: a novel method for the presentation of data synthesis. Research Synthesis Methods 2011;**2**(2):79-83 doi: citeulike-article-id:13194581 doi: 10.1002/jrsm.37[published Online First: Epub Date]|.

38. Thomas S, Fayter D, Misso K, et al. Population tobacco control interventions and their effects on social inequalities in smoking: Systematic review. Tob. Control 2008;**17**(4):230-37 doi: 10.1136/tc.2007.023911[published Online First: Epub Date]|.

39. Magnée T, Burdorf A, Brug J, et al. Equity-specific effects of 26 dutch obesity-related lifestyle interventions. American Journal of Preventive Medicine 2013;**44**(6):e57-e66 doi: 10.1016/j.amepre.2012.11.041[published Online First: Epub Date]|.

40. Reichow B, Servili C, Yasamy MT, et al. Non-Specialist Psychosocial Interventions for Children and Adolescents with Intellectual Disability or Lower-Functioning Autism Spectrum Disorders: A Systematic Review. PLoS Med. 2013;**10**(12):1-27 doi: 10.1371/journal.pmed.1001572[published Online First: Epub Date]|.

41. Thomson HJ, Thomas S. The effect direction plot: visual display of non-standardised effects across multiple outcome domains. Research Synthesis Methods 2013;**4**(1):95-101 doi: 10.1002/jrsm.1060[published Online First: Epub Date]|.

42. Thomson H, Thomas S, Sellstrom E, et al. Housing improvements for health and associated socio-economic outcomes. Cochrane Database Syst Rev 2013;**2**:CD008657 doi: 10.1002/14651858.CD008657.pub2[published Online First: Epub Date]|.

43. Rosen L, Ben Noach M, Rosenberg E. Missing the forest (plot) for the trees? A critique of the systematic review in tobacco control. BMC Medical Research Methodology 2010;**10**(1):34

44. Shojania KG, Jennings A, Mayhew A, et al. The effects of on-screen, point of care computer reminders on processes and outcomes of care. Cochrane Database Syst Rev 2009(3):CD001096 doi: 10.1002/14651858.CD001096.pub2[published Online First: Epub Date]|.

45. Arditi C, Rege-Walther M, Wyatt JC, et al. Computer-generated reminders delivered on paper to healthcare professionals; effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2012;**12**:CD001175 doi: 10.1002/14651858.CD001175.pub3[published Online First: Epub Date]|.

46. Schriger DL, Altman DG, Vetter JA, et al. Forest plots in reports of systematic reviews: a cross-sectional study reviewing current practice. International Journal of Epidemiology 2010;**39**(2):421-29 doi: 10.1093/ije/dyp370[published Online First: Epub Date]|.

47. Bax L, Ikeda N, Fukui N, et al. More than numbers: the power of graphs in meta-analysis. American Journal of Epidemiology 2009;**169**(2):249-55 doi: citeulike-article-id:13187983 doi: 10.1093/aje/kwn340[published Online First: Epub Date]|.

48. Anzures-Cabrera J, Higgins JPT. Graphical displays for meta-analysis: An overview with suggestions for practice. Research Synthesis Methods 2010;**1**(1):66-80 doi: citeulike-article-id:13195378 doi: 10.1002/jrsm.6[published Online First: Epub Date]|.

49. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database of Systematic Reviews 2012(6):CD000259

50. Huque M. Experiences with meta-analysis in NDA submissions. Proc Biopharmaceutical Section Am Statist Assoc 1988;**2**:28-33

51. Hedges L, Vevea J. Fixed- and random-effects models in meta-analysis. Psychological methods 1998;**3**(4):486-504

52. Borenstein M, Hedges LV, Higgins JPT, et al. A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods 2010;**1**(2):97-111 doi: 10.1002/jrsm.12[published Online First: Epub Date]|.

53. Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. Statistics in medicine 2011;**30**(20):2481-98 doi: citeulike-article-id:13195116 doi: 10.1002/sim.4172[published Online First: Epub Date]|.

54. Akl EA, Oxman AD, Herrin J, et al. Framing of health information messages. Cochrane Database Syst Rev 2011(12):CD006777 doi: 10.1002/14651858.CD006777.pub2[published Online First: Epub Date]|.

55. Smedslund G, Berg RC, Hammerstrom KT, et al. Motivational interviewing for substance abuse. Cochrane Database Syst Rev 2011(5):CD008063 doi: 10.1002/14651858.CD008063.pub2[published Online First: Epub Date]|.

56. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011;**342**:d459 doi: citeulike-article-id:13195586 doi: 10.1136/bmj.d549[published Online First: Epub Date]|.

57. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society. Series A: Statistics in Society 2009;**172**(1):137-59 doi: citeulike-article-id:13188272 doi: 10.1111/j.1467-985X.2008.00552.x[published Online First: Epub Date]|.

58. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. Journal of the Royal Statistical Society. Series A: Statistics in Society 2009;**172**(4):789-811 doi: citeulike-article-id:13188710 doi: 10.1111/j.1467-985X.2008.00593.x[published Online First: Epub Date]|.

59. Del Re AC, Maisel N, Blodgett J, et al. The declining efficacy of naltrexone pharmacotherapy for alcohol use disorders over time: A multivariate meta-analysis. Alcohol. Clin. Exp. Res. 2013;**37**(6):1064-68 doi: 10.1111/acer.12067 10.1111/j.1360-0443.2012.04054.x [Epub ahead of print]

60. Del Re AC, Maisel N, Blodgett JC, et al. Placebo group improvement in trials of pharmacotherapies for alcohol use disorders: A multivariate meta-analysis examining change over time. J. Clin. Psychopharmacol. 2013;**33**(5):649-57 doi: 10.1097/JCP.0b013e3182983e73[published Online First: Epub Date]|.

61. Schild AHE, Voracek M. Less is less: a systematic review of graph use in meta-analyses. Research Synthesis Methods 2013;**4**(3):209-19 doi: citeulike-article-id:13193569 doi: 10.1002/jrsm.1076[published Online First: Epub Date]|.

62. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001;**322**(7300):1479-80 doi: citeulike-article-id:13191686 doi: 10.1136/bmj.322.7300.1479[published Online First: Epub Date]|.

63. Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to meta-analysis*. Chichester, West Sussex, UK: John Wiley & Sons, 2009.

64. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods 2010;**1**(1):39-65 doi: 10.1002/jrsm.5[published Online First: Epub Date]|.

65. Jackson D, Riley RD. A refined method for multivariate meta-analysis and meta-regression. Statistics in medicine 2014;**33**(4):541-54 doi: 10.1002/sim.5957[published Online First: Epub Date]|.

66. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Statistics in medicine 2002;**21**(11):1559-73 doi: citeulike-article-id:13191375 doi: 10.1002/sim.1187[published Online First: Epub Date]|.

67. Harbord RM, Higgins JPT. Meta-regression in Stata. Stata J. 2008;**8**(4):493-519

68. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Statistics in medicine 2004;**23**(11):1663-82 doi: 10.1002/sim.1752[published Online First: Epub Date]|.

69. Sutton AJ, Higgins JPT. Recent developments in meta-analysis. Statistics in medicine 2008;**27**(5):625-50

70. O'Brien MA, Rogers S, Jamtvedt G, et al. Educational outreach visits: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2007(4):CD000409 doi: 10.1002/14651858.CD000409.pub2[published Online First: Epub Date]|.

71. Michie S, Abraham C, Whittington C, et al. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health psychology : official journal of the Division of Health Psychology, American Psychological Association 2009;**28**(6):690-701 doi: 10.1037/a0016136[published Online First: Epub Date]|.

72. Schild AHE, Voracek M. Finding your way out of the forest without a trail of bread crumbs: development and evaluation of two novel displays of forest plots. Research Synthesis Methods 2014:n/a-n/a doi: 10.1002/jrsm.1125[published Online First: Epub Date]|.

73. Effective Practice and Organisation of Care (EPOC). Synthesising results when it does not make sense to do a meta-analysis. EPOC Resources for review authors. Oslo: Norwegian Knowledge Centre for the Health Services. 2014. http://epocoslo.cochrane.org/epoc-specific-resources-review-authors.

Appendix 1a

MEDLINE search strategy

Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily, Ovid MEDLINE(R) and Ovid OLDMEDLINE(R) <1946 to Present>		
#	**Search Statement**	**Results**
1	(synthesis or syntheses or summari?e$ or review$).tw.	1876263
2	(harvest adj3 plot$).tw.	34
3	(vote$ adj3 count$).tw.	99
4	(direction adj3 plot$).tw.	28
5	(combin$ adj3 p value$).tw.	366
6	("non?parametric" or "non parametric").tw.	15817
7	(median adj3 medians).tw.	16
8	whisker.tw.	2064
9	stacked.tw.	6741
10	2 or 3 or 4 or 5 or 7 or 8 or 9	9344
11	1 and 2	14
12	1 and 3	49
13	1 and 4	1
14	1 and 5	27
15	1 and 6	992
16	1 and 7	1
17	1 and 8	69
18	1 and 9	539
19	1 and 10	699
20	((synthesis or syntheses or summari?e$ or review$) adj10 ("non?parametric" or "non parametric")).tw.	110
21	from 20 keep 1	1
22	(whisker adj3 plot$).tw.	72
23	1 and 22	6
24	(stacked and (plot$ or graph or graphs)).tw.	84
25	1 and 24	4

Appendix 1b

Terms used to search the Meth4ReSyn and SRC methods libraries in Endnote

	Endnote search terms	Descriptions	No retrieved
1	synthes OR summar OR review [ti OR ab]	papers related to synthesis or summary or reviews	4967
2	meta-anal OR meta anal [ti OR ab]	papers related to meta-analysis	3021
3	1 OR 2	papers related to summary, synthesis, reviews or meta-analysis	6219
4	complex [ti] AND 3	set 3 limited to papers that include reference to complexity in the title	53
5	public health OR policy OR health services [ti] AND 3	set 3 limited to papers that include reference to public health, policy or health services in the title	190
6	graph OR plot or visual [ti]	papers related to graphs, plots or visual depiction	133
7	tabul or table or matri [ti]	papers related to tabulation	61
8	non-parametric [ti OR ab]	papers related to non-parametric methods	27
9	vote [ti OR ab]	papers related to vote counting methods	57
10	ad hoc [ti OR ab]	papers related to ad hoc synthesis methods	20
11	p-value [ti OR ab] AND 3	set 3 limited to papers that include reference to combining p-values	86
12	4 to 11 [all refs unique]	papers related to summary, synthesis, reviews or meta-analysis and complexity, public health, policy, or health services (sets 4 and 5); papers reporting specific methods (sets 6 to 11) duplicates removed	612