Systematic review and meta-analysis

Authors: Ethan Balk, MD, MPH, Peter A L Bonis, MD
Section Editor: Joann G Elmore, MD, MPH
Deputy Editor: Carrie Armsby, MD, MPH

All topics are updated as new evidence becomes available and our peer review process is complete.

Literature review current through: Apr 2022. | **This topic last updated:** Oct 01, 2021.

INTRODUCTION

This topic review will provide an overview of how systematic reviews and meta-analyses are conducted and how to interpret them. In addition, it will provide a summary of methodologic terms commonly encountered in systematic reviews and meta-analyses.

A broader discussion of evidence-based medicine and a glossary of methodologic and biostatistical terms are presented separately. (See "Evidence-based medicine" and "Glossary of common biostatistical and epidemiological terms".)

KEY DEFINITIONS

The terms systematic review and meta-analysis are often used together, but they are not interchangeable. Not all systematic reviews include meta-analyses, though many do.

These terms are defined here since they are used throughout this topic. A glossary of other relevant terms is provided at the end of this topic. (See 'Glossary of terms' below.)

Systematic review — A systematic review is a comprehensive summary of all available evidence that meets predefined eligibility criteria to address a specific clinical question or range of questions. It is based upon a rigorous process that incorporates [1-3]:

- Systematic identification of studies that have evaluated the specific research question(s)
- Critical appraisal of the studies
- Meta-analyses (not always performed) (see 'Meta-analysis' below)
- Presentation of key findings
- Explicit discussion of the limitations of the evidence and the review

Systematic reviews contrast with traditional "narrative" reviews and textbook chapters. Such reviews generally do not exhaustively review the literature, lack transparency in the selection and interpretation of supporting evidence, generally do not provide a quantitative synthesis of the data, and are more likely to be biased [4].

Meta-analysis — Meta-analysis, which is commonly included in systematic reviews, is the statistical method of quantitatively combining or pooling results from different studies. It can be used to provide overall pooled effect estimates [5]. For example, if a drug was evaluated in multiple placebo-controlled trials that all reported mortality, meta-analysis can be used to estimate a pooled relative risk for the drug's overall effect on mortality in all of the trials together. Meta-analysis can also be used to pool other types of data such as studies on diagnostic accuracy (ie, pooled estimates on sensitivity and specificity) and epidemiologic studies (ie, pooled incidence or prevalence rates; pooled odds ratio for strength of association). Meta-regression and network meta-analysis (NMA) are enhancements to traditional meta-analysis. (See 'Meta-regression' below and 'Network meta-analysis' below.)

ADVANTAGES OF SYSTEMATIC REVIEW AND META-ANALYSIS

Clinical decisions in medicine ideally should be based upon guidance from a comprehensive assessment of the body of available knowledge. A single clinical trial, even a large one, is seldom sufficient to provide a confident answer to a clinical question. Indeed, one analysis suggested that most research claims are ultimately proven to be incorrect or inaccurate when additional studies have been performed [6]. At the same time, it is well established that large randomized controlled trials do

not always confirm the results of prior meta-analyses [7-9]. The "truth" needs to be understood by examining all sources of data as critically and objectively as possible.

There are several potential benefits to performing systematic analysis, which may also include meta-analysis:

- Unique aspects to a single randomized trial, involving the participating patient population, protocol, setting in which the trial is performed, or expertise of the involved clinicians, may limit its generalizable to other settings or individual patients. The conclusions of systematic reviews are likely to be more generalizable than single studies.
- Combining studies in meta-analyses increases the sample size and generally produces more precise estimates of the effect size (ie, estimates that have smaller confidence intervals) than a single randomized trial. Meta-analysis may also allow exploration of heterogeneity across studies to allow conclusions beyond what can be gleaned from individual studies.
- Clinicians rarely have the time or resources to critically evaluate the body of evidence relevant to a particular clinical question, and a systematic review can facilitate this investigation.
- In contrast with narrative review articles, most systematic reviews focus on a narrow, clearly defined topic and include all eligible studies, not just those chosen by the author. Systematic reviews start with a clinical or research question and form conclusions based on the evidence. This is in contrast with many narrative reviews that start with a conclusion and include evidence to support that conclusion.

Systematic review and meta-analysis are methods to synthesize the available evidence using an explicit, transparent approach that considers the strengths and weaknesses of the individual studies, populations and interventions, and specific outcomes that were assessed. Individual practitioners, policymakers, and guideline developers can use well-conducted systematic reviews to determine best patient management decisions. Organizations that develop guidelines can use the results of systematic reviews and meta-analyses to provide evidence-based recommendations for care.

STEPS TO CONDUCTING A SYSTEMATIC REVIEW AND META-ANALYSIS

Overview — Several steps are essential for conducting a systematic review or meta-analysis. These include:

- Formulating research questions (see 'Formulating research questions' below)
- Developing a protocol (see 'Developing a protocol' below)
- Searching for the evidence (see 'The literature search' below)
- Assessing the quality of studies (see 'Risk of bias assessment' below)
- Summarizing and displaying results (eg, using forest pots and a summary of findings table, as shown in the figure (figure 1)) (see 'Forest plot' below)
- Exploring reasons for heterogeneity across studies (see 'Heterogeneity' below and 'Subgroup analyses' below and 'Sensitivity analysis' below)

The basic steps, along with limitations that should be considered, are discussed here. While this topic review focuses on meta-analysis of randomized controlled trials, many of the methods and issues apply equally to meta-analyses of other comparative studies, noncomparative (single group) and other observational studies, and studies of diagnostic tests. An overview of approaches to systematic review and meta-analysis is provided in a table (table 1).

The updated 2020 Preferred Reporting Items of Systematic reviews and Meta-Analyses (PRISMA) statement emphasizes that systematic reviews should provide the protocol, data, and assessments of risk of bias (RoB) from individual studies with sufficient transparency to allow the reader to verify the results [10]. It underscores the basic questions that the clinician and investigator should ask when interpreting a systematic review. The PRISMA website provides checklists for the items that should be included in a systematic review. Several "extensions" to PRISMA have been developed for specific types of systematic reviews or meta-analyses (eg, harms, network meta-analyses, meta-analyses of diagnostic tests, individual patient data analyses) [11]. In addition, readers of systematic reviews should assess the relevance to their own practice in regard to the studied populations, settings, interventions, and outcomes assessed.

In 2011, the Institute of Medicine has published recommended standards for developing systematic reviews, which remain pertinent [12]. While these standards principally apply to publicly funded systematic reviews of comparative effectiveness research that focus specifically on treatments, most of the standards pertain to all systematic reviews. The United States Agency for Healthcare Research and Quality also has an ongoing series of articles that form a Methods Guide for Comparative Effectiveness Reviews for its Evidence-based Practice Center program and related reviews. This guide principally applies to large

overarching systematic reviews but provides insights and recommendations for addressing different types of topics and studies.

Formulating research questions — Research questions (often referred to as "key questions") are analogous to the research hypotheses of primary research studies. They should be focused and defined clearly since they determine the scope of research the systematic review will address [13].

Broad questions that cover a range of topics may not be directly answerable and are not appropriate for systematic reviews or meta-analyses. As an example, the question "What is the best treatment for chronic hepatitis B?" would need to be broken down into several smaller well-focused questions that could be addressed in individual and complementary systematic reviews. Examples of appropriate key questions may include, "How does entecavir compare with placebo for achieving hepatitis B e antigen (HBeAg) seroconversion in patients with chronic HBeAg-positive hepatitis B?" and "What is the relationship between hepatitis B genotypes and response rates to entecavir?" These and other related questions would be addressed individually and then, ideally, considered together to answer the more general question.

Key questions for studies of the effectiveness of interventions are commonly formulated according to the "PICO" method, which fully defines the **P**opulation, **I**ntervention, **C**omparator, and **O**utcomes of interest [13]. The acronym "PICOD" is sometimes used to indicate that investigators must also specify which study designs are appropriate to include (eg, all comparative studies versus only randomized trials). Other eligibility criteria may include the timing or setting of care. Variations of these criteria should be used for systematic reviews of other study designs, such as of cohort studies (without a comparator), studies of exposures, or studies of diagnostic tests.

Developing a protocol — A written protocol serves to minimize bias and to ensure that the review is implemented according to reproducible steps. A systematic review should describe the research questions and the review methodology, including the search strategy and approach to analyzing the data. Ideally, the protocol should be a collaborative effort that includes both clinical and methodology experts [14].

Publication of protocols can be useful to prevent unnecessary duplication of efforts and to enhance transparency of the systematic review. A voluntary registry, PROSPERO, was established in 2011. The database contains protocol details for systematic reviews that have health-related outcomes.

The literature search

Performing the search — The literature search should be systematic and comprehensive to minimize error and bias [13]. Most systematic reviews start with a search of an electronic database of the literature. PubMed [15] is almost universally used; other commonly searched databases include Embase [16] and the Cochrane Central Register of Controlled Trials (CENTRAL) [17]. Inclusion of additional databases should be considered for specialized topics such as complementary or alternative medicine, quality of care, or nursing. Electronic searches should be supplemented by searches of the bibliographies of retrieved articles and relevant review articles and by studies known to domain experts.

The research community has also recognized a need to incorporate the "grey literature" to diminish the risks of publication bias (selective publication of studies, possibly based on their results) and reporting bias (selective reporting of study results, possibly based on statistical significance) [12,18-20]. There is no standard definition of grey literature, but it generally refers to information obtained from sources other than published, peer-reviewed articles, such as conference proceedings, clinical trial registries, adverse events databases, government agency databases (eg, US Food and Drug Administration) and documents, unpublished industry data, dissertations, and online sites. Methods to incorporate other types of relevant information, particularly "real-world data" obtained from analyzing databases of patients undergoing routine care, are still being developed [21,22].

Publication and reporting bias — Reporting bias refers to bias that results from incomplete publishing or reporting of available research. This is a common concern and a potentially important limitation of systematic review since the missing data may affect the validity of systematic reviews [23]. There are two main categories of reporting bias:

- Publication bias – Compared with positive studies, negative studies may take longer to be published or may not be published at all [24]. This is referred to as "publication bias."
- Outcome reporting bias – "Outcome reporting bias" refers to the concern that a study may only include outcomes that are favorable and significant in the published report, while nonsignificant or unfavorable outcomes are selectively not reported.

Several methods have been developed to evaluate whether publication bias is present. However, they all involve major assumptions about possible missing studies [25]. Any evaluation of publication bias should not be considered definitive, but rather only exploratory in nature.

A commonly used method for assessing publication bias is the funnel plot, which is a scatter plot displaying the relationship between the weight of the study (eg, study size or standard error) and the observed effect size (figure 2) [26]. An asymmetric appearance, especially due to the absence of smaller negative studies, can suggest unpublished data. However, this assessment is not definitive since asymmetry could be due to factors other than unpublished negative studies (such as population heterogeneity or study quality) [23,27-29].

Other methods to evaluate reporting bias include the "trim and fill" method [30], "modeling selection process" [31,32], and testing for an excess of significant findings. These methods are beyond the scope of this topic [33].

Risk of bias assessment — The quality of an individual study has been defined as the "confidence that the trial design, conduct, and analysis has minimized or avoided biases" [34]. The risk of bias (RoB) assessment (sometimes referred to as "quality assessment") represents the extent to which trial design and methodology prevented systematic error and can help explain differences in the results of systematic reviews.

The primary value of the RoB assessment of individual studies in the meta-analysis is to determine the degree of confidence that the pooled effect estimate reflects the "truth" as best as it can be measured. One would be more likely to have high confidence in conclusions based upon "high-quality" (ie, low RoB) studies rather than "low-quality" (ie, high RoB) studies. Differences in RoB of individual studies can also be explored to help explain heterogeneity (eg, does the effect in low RoB studies differ from that in high RoB?).

The process of assessing study quality is not straightforward. Several different RoB scoring systems are available. Commonly used tools, among many others, include:

- Original Cochrane RoB tool for randomized controlled trials (with 7 questions [35])
- More complex revision of this tool, RoB 2 (with 5 overarching questions and 22 subquestions [36])

- The ROBINS-I tool (Risk Of Bias In Non-randomized Studies of Interventions, with 7 overarching questions and 31 subquestions [37])

Different methodologists use different tools depending on available time and resources, needs and purpose of the given review, and philosophical differences among researchers about the relative importance of different "quality" factors. Importantly, the assessment of a study's RoB can be limited by the need to rely on information presented in the manuscript [38].

For randomized trials, the RoB assessment typically considers the following factors:

- **Randomization method** – Some "randomization" methods are not truly random, which can be a source of bias. For example, a computer algorithm is generally preferred over a system based on day of the week or other nonrandom method.

- **Allocation concealment** – Allocation is the assignment of study participants to a treatment group. It occurs between randomization and implementation of the intervention. Allocation should be adequately concealed from the study personnel. A study may be biased if allocation is not concealed. For example, if the study used unsealed envelopes corresponding to the randomization order to assign patients to each treatment arm, the study personnel could read the contents and thereby channel certain patients into the desired treatment (eg, if they believed the investigational treatment was effective, they may channel sicker patients into that arm). This would result in imbalance between the two arms of the study (ie, the intervention arm would have sicker patients while the control arm would have healthier people), resulting in the intervention appearing to be less effective than it truly is.

- **Blinding** – Ideally, all relevant groups should be blinded to treatment assignment. This includes study participants, clinicians, data collectors, outcome assessors, and data analysts. Blinding is not always feasible. Some forms of surgery or behavioral modifications, for example, do not lend themselves to blinding of patients and providers. However, outcome assessors and data analyst can usually be blinded regardless of the type of treatment. "Double blinding" generally refers to blinding of the study participants and at least one of the study investigators, although it may not be clear who was blinded when only "double blinding" is reported. For adequate blinding, treatments with a noticeable side effect (eg, niacin) ideally should have an "active control" that mimics the side effect.

- **Differences between study groups** – Differences in the treatment groups at baseline can lead to biased results. The goal of randomization is to balance important prognostic variables relevant to the outcome(s) of interest among the different treatment groups. However, randomization is not always successful. Differences in treatment groups typically occur in trials with relatively small numbers of subjects. Researchers can attempt to adjust for baseline differences in the statistical analysis, but it is far more preferable to have balanced groups at baseline.

- **Attrition and incomplete reporting** – High rates of withdrawal of participants from a study may indicate a fundamental problem with the study design. Uneven withdrawal from different study groups can lead to bias, particularly if the reasons for withdrawal differ between, and are related to, the interventions (such as ascribing adverse events to the intervention or lack of effectiveness to the placebo). Reports should describe the reasons for patient withdrawal to allow assessment of their effect on bias and study applicability.

- **Early termination for benefit** – Stopping a trial early for benefit will, on average, overestimate treatment effects [39]. However, the degree of overestimation varies. Small trials that are stopped early with few events can result in large overestimates. In larger trials with more events (ie, >200 to 300 events), early stopping is less likely to result in serious overestimation [40]. Early termination of a trial for harm can also introduce bias (ie, overestimation of the harm); however, it is generally considered ethically obligatory to stop the trial in such circumstances. Early termination for other reasons (eg, slow accrual) is not considered a source of bias per se, though it can sometimes indicate that there are other problems with the trial (eg, the eligibility criteria may be too strict and not reflective of the patient population seen in actual clinical practice).

Other factors that may be considered when assessing the methodologic quality of a study include the accuracy of reporting (eg, details of study methodology, patient characteristics, and study results) and the appropriateness of statistical analyses. For example, an intention to treat (ITT) analysis is appropriate for assessing efficacy of a treatment since it preserves the comparability of treatment groups achieved by randomization. In some cases, it may be appropriate to perform a per protocol analysis alongside the ITT analysis, but when performed alone, per protocol analyses can lead to biased results.

The RoB assessment involves judgement. For this reason, it should generally be performed independently by two separate reviewers and there should be a process for resolving disagreements.

Meta-analysis

Statistical methods for combining data — Meta-analysis combines results across studies to provide overall estimates and confidence intervals of treatment effects. For dichotomous outcomes (ie, outcomes with two possible states, such as death versus survival), results are summarized using an odds ratio (OR), relative risk (RR; also called risk ratio), or hazard ratio (HR). Essentially, any study metric can be meta-analyzed, including continuous variables (mean, mean difference, percent change) or proportions. However, meta-analysis is not feasible if the studies measured completely different outcomes (eg, one trial measured pain scores while the other measured functional ability).

There are numerous specific methodologic details of meta-analysis that are beyond the scope of this topic. The primary consideration is whether the summary effect estimate should be calculated under the assumption of a "random effects" or a "fixed effect" model [41]. For most of the medical literature, the random effects model is the more appropriate approach. These two approaches are discussed in detail below. (See 'Random effects model' below and 'Fixed effect model' below.)

When to combine studies — The decision to combine studies should be based upon both qualitative and quantitative evaluations. Important qualitative features include the degree of similarity of populations, interventions, outcomes, study objectives, and study designs that incorporate both clinical and biologic plausibility. The systematic reviewers should provide a sufficient explanation of the rationale for combining studies to allow the readers to judge for themselves whether they agree that it was appropriate to combine the individual studies.

Quantitative methods to examine heterogeneity may also be considered in making the decision or determining if it is appropriate to combine data. These typically involve the I^2 index or Q statistic, which are described below. (See 'Heterogeneity' below and 'I2 index' below and 'Q statistic' below.)

These statistics, however, generally have low power and are thus prone to false negative results (eg, not detecting heterogeneity when it is present). Evidence of statistical heterogeneity does not preclude appropriate meta-analysis.

Precision — Precision refers to the extent to which the observed results would be reproduced exactly given the same interventions and study design. Precision is generally assessed by examining the confidence intervals (CIs). The narrower the CIs are, the more precise the estimate is. If the estimate is too imprecise (ie, CIs are too wide), our certainty in the finding is

reduced. But how wide is too wide? As a general rule, imprecision is problematic if the clinical decision based on the result (eg, to use or not use an intervention) would be different at the upper versus lower boundary of the 95% CIs. For organizations issuing guidelines, the strength of the evidence should be downgraded for imprecision in this scenario. Specific criteria for imprecision have been developed in the context of grading for guidelines [42].

Precision is different from validity, which refers to the extent to which the results reflect the "truth." The figure illustrates a conceptual example of the difference between precision and validity (figure 3).

Problematic imprecision is often encountered when the sample size is small (particularly if there are few events). An important advantage of meta-analysis is that combining studies produces more precise estimates of the effect size (ie, estimates that have narrower CIs) due to the increased sample size. (See 'Advantages of systematic review and meta-analysis' above.)

Sensitivity analysis — A meta-analysis should test how stable the overall estimates are when different subgroups of studies are analyzed and should explore heterogeneity among the studies. Meta-regression and subgroup analyses can be used to examine the influence on the overall results. When feasible, meta-analysis of individual patient data allows the most rigorous exploration of heterogeneity. (See 'Individual patient data' below.)

Explorations such as reanalyzing the data with single studies or groups of studies (eg, high RoB studies) omitted can be used to determine the degree to which overall results are being driven by these studies. Conclusions should seldom be driven by a single study since the meta-analysis would add little additional information or confidence compared with the single study alone.

Sensitivity analyses can also be used to explore such issues as publication or reporting bias. As an example, finding that meta-analysis of the largest studies yields smaller effect sizes than meta-analysis of all trials can suggest that smaller "negative" trials may be missing [43,44].

Subgroup analyses — Another way to explore heterogeneity is subgroup analysis, which involves performing separate analyses based upon clinically relevant variables. Subgroup analysis is subject to the same limitations inherent to meta-regression, including risks associated with data dredging and ecological fallacy. To minimize the risk of drawing false conclusions, subgroup analyses in meta-analyses should be:

- Specified a priori, including hypotheses for the direction of the differences (ie, they should be based upon prior evidence or knowledge)
- Limited to only a few (ie, to avoid data dredging)
- Analyzed by testing for interaction (eg, using meta-regression) rather than simply comparing the separate effect estimates

An example of subgroup analysis is shown in the figure (figure 4), which is from a meta-analysis examining the effect of continuous positive airway pressure on reducing depressive symptoms in patients with obstructive sleep apnea [45]. The investigator performed subgroup analyses to explore whether the effect differed in studies involving patients with baseline depression compared with studies of patients without baseline depression. in this case, the test for subgroup effect (ie, interaction) was statistically significant ($p<0.001$).

The approach to evaluating subgroup analyses in meta-analyses and clinical trials is discussed in greater detail separately. (See "Evidence-based medicine", section on 'Subgroup analyses'.)

Meta-regression — Regression analysis of primary studies may be used to adjust for potential confounders or explain differences in results among subjects. This meta-analytic technique is commonly known as meta-regression. In this approach, the dependent variable in the regression is the estimate of treatment effect from each individual study and the independent variables (eg, covariates such as drug dose, treatment duration, or study size) are the aggregated characteristics derived from the individual studies. Instead of individual patients serving as the units of analysis, each individual study is considered to be one observation [46-48]. Meta-regression tests the statistical interaction between the subgroup variable (eg, dose) and the treatment effect (eg, relative risk of death). It can include categorical variables (including two or more categories, such as study country or study design) and continuous variables (such as dose or follow-up duration) either singly (univariable analysis) or together (multivariable analysis).

An example of a meta-regression of early trials of zidovudine monotherapy for HIV infection is shown in a figure (figure 5) [49]. The meta-regression successfully explains the heterogeneity across studies, showing an association between treatment duration and the effect of treatment on death that was not apparent within the individual trials.

There are several caveats related to the performance and interpretation of meta-regression:

- Meta-regression and subgroup analyses (that rely on retrospective data from previously run trials) should be considered to yield hypothesis-generating, rather than conclusive, associations, in contrast to well-designed regressions of prospective study data.
- Meta-regression is not always feasible, since covariates may not be fully reported or may not be uniformly defined.
- Data dredging (analyzing every possible variable regardless of clinical relevance) can result in spurious associations [50].
- It may be difficult to account properly for patient-level variables (such as age, sex, or laboratory values). Most studies, for example, report averages for such variables (eg, a mean age of 47.1 years) that do not reflect the range of values across the study population. Making an assumption about individual data based upon aggregated statistics (known as "ecological fallacy") can produce invalid results in meta-regression [51,52]. The only reliable way to address this is to analyze patient-level data.

Individual patient data — It is sometimes possible to obtain original patient-level databases to reanalyze individual patient data in a meta-analysis [14]. Pooling individual patient data is the most rigorous form of meta-analysis. While more costly and time-consuming and limited by difficulties collecting original data, there are several benefits. These include the ability to perform meta-regressions of patient-level predictors (eg, age) without the risk of ecological fallacy; time-to-event analyses; and to include unpublished, previously unanalyzed data. However, analyses of partial databases (all that may be available with proprietary data) or of selected databases are subject to selection bias or limited generalizability of results, similar to other retrospective analyses of incomplete samples.

Network meta-analysis — When multiple different interventions are compared across trials, a network of studies can be established where all the studied interventions are linked to each other by individual trials. Network meta-analysis (NMA) evaluates all studies and all interventions simultaneously to produce multiple pairwise estimates of relative effects of each intervention compared with every other intervention [53,54].

A schematic representation of a network diagram is shown in the figure (figure 6). In reality, some network diagrams in NMAs are far more complex (figure 7).

The pairwise comparisons in NMAs are based upon both direct and indirect comparisons. For example, consider two drugs (drug A and drug B) that were each evaluated in placebo-controlled trials and directly compared with one another in a separate clinical trial (figure 6). NMA can be used to estimate the relative efficacy of drug A versus drug B based upon the direct comparison (ie, from the trial directly comparing drug A to drug B) and indirect comparisons (ie, from the placebo-controlled trials). The direct and indirect estimates are then pooled together to yield an overall estimate (or "network estimate") of the relative effect. Typically, the direct, indirect, and network estimates are reported separately in NMAs. Some of the comparisons in a NMA may be based entirely on indirect data.

When assessing the validity of an NMA, many of the same principles that are used for assessing conventional meta-analysis apply (eg, was the literature search comprehensive, were eligibility criteria for the studies clearly stated, were the individual studies assessed for RoB, how precise are the effect estimates, etc (table 2)). However, there are two concerns that are unique to NMAs [55,56]:

- **Intransitivity** – The assumption of transitivity is fundamental to NMA because the network estimates rely upon indirect comparisons. For the transitivity assumption to hold, the individual studies must be sufficiently similar in all respects other than the treatments being compared (ie, similar participants, setting, ancillary treatments, and other relevant parameters). In the example above, if studies of drug A versus placebo are systematically different than studies of drug B versus placebo (eg, if they were conducted in an earlier era), then the indirect comparison of drug A versus drug B may be biased due to these differences (ie, the difference may be partly explained by differences in disease management over the intervening decades).

- **Incoherence** – Incoherence (also called inconsistency) refers to differences between the direct and indirect estimates. Incoherence can be a consequence of bias due to methodologic limitations of the studies, publication bias, indirectness, or intransitivity. If the direct and indirect estimates are considerably different from each other, the network estimate may not be valid. Addressing incoherence and assessing its impact on the network estimate requires judgement [55].

Bayesian methods are commonly used to conduct NMA [57]. This approach has the advantage of allowing estimation of the probability of each intervention being best, which, in turn, allows interventions to be ranked. Such ranking, however, needs to be interpreted cautiously, as it can be unstable, depending on the network topology, and can have a substantial degree of imprecision [58].

READING AND INTERPRETING A SYSTEMATIC REVIEW

Key questions to consider when reading and interpreting a systematic review are summarized in the table (table 2). The reader should appraise the systematic reviews for its quality, potential sources of bias, and extent to which the findings are applicable to their specific question. Systematic review and meta-analysis are subject to the same biases observed in all research. In addition, the value of a systematic review's conclusions may be limited by the quality and applicability of the individual studies included in the review.

Since meta-analysis is a pooling of distinct individual studies, it is important to bear in mind that the overall results, even more than individual study results, are not directly interpretable as a patient-level risk (of an outcome) and cannot make personalized predictions for patients. Results must be interpreted as an average result for a population.

GLOSSARY OF TERMS

Applicability (generalizability) — The relevance of a study (or a group of studies) to a population of interest (or an individual patient). This requires an assessment of how similar the subjects of a study are to the population of interest, the relevance of the studied interventions and outcomes, and other PICO features. (See 'PICO method (PICOD, PICOS, PICOTS, others)' below.)

Ecological fallacy (ecological inference fallacy) — An error in interpreting data where inferences are made about specific individuals based upon aggregated statistics for groups of individuals.

Fixed effect model — The central assumption of a fixed effect model is that there is a single true treatment effect and that all trials provide estimates of this one true effect. Meta-analysis thus provides a pooled estimate of the single true effect. A hypothetical model for a fixed effect model meta-analysis is shown in a figure (figure 8).

The central assumption of a fixed effect model is that estimates from each study differ solely because of random error around a common true effect. This assumes that all studies represent the same population, intervention, comparator, and outcome for which there is a single "true" effect size. Fixed effects models yield effect size estimates by assigning a weight to each individual

study estimate that reflects the inherent variability in the results measured (ie, the "within-study variance" related to the standard error of the outcome).

There are limited instances when it is appropriate to use a fixed effects model for summarizing clinical trials. These include meta-analyses in which:

- There is extreme confidence that the studies are comparable (ie, characteristics of the enrolled patients, the type of intervention, comparators and outcome measures) such that any difference across studies is just due to random variation. Such an assumption is typically difficult to justify. One example of an appropriate use of the fixed effects model is meta-analysis of repeated, identical, highly controlled trials in a uniform setting, as may be done by pharmaceutical companies during early testing.

- The studies are of rare events in which one form of a fixed effects model (the Peto odds ratio) may be less biased than other methods of pooling data [59].

Forest plot — A forest plot is a graphical presentation of individual studies, typically displayed as point estimates with their associated 95% CIs on an appropriate scale, next to a description of the individual studies (figure 9). The forest plot allows the reader to see the estimate and the precision of the individual studies, appreciate the heterogeneity of results, and compare the estimates of the individual studies to the overall summary estimate.

Ideally, a forest plot should provide sufficient data for the reader to make some assessment of the individual studies in the context of the overall summary (eg, to compare sample sizes, any variations in treatments such as dose, baseline values, demographic features, and study quality).

Funnel plot — A graphical technique, with related statistical tests, to examine the studies within a systematic review for the possibility of publication bias (figure 2). (See 'Publication and reporting bias' above.)

Grey literature — A term with varying and shifting meaning that indicates sources of evidence beyond the peer-reviewed, published literature available in major databases (eg, Medline). Examples include alternative databases, conference abstracts and proceedings, unpublished studies (eg, via clinicaltrials.gov), newspaper or internet citations, citation indexes, handsearching of journals or reference lists, and domain experts.

Heterogeneity

Clinical heterogeneity — Qualitative differences in study features, such as study eligibility criteria, interventions, or methods of measuring outcomes, that may preclude appropriate meta-analysis. These features can be explicit (such as different drug doses used) or implicit (such as differences in populations depending on setting or country). Clinical heterogeneity may or may not result in statistical heterogeneity but often may not: for example, if the effect size is similar regardless of the drug dose, of the individual drug within a class of drugs, or in different populations (eg, men and women, or Japanese and American).

Statistical heterogeneity — Quantitative differences in study results across studies examining similar questions. Statistical heterogeneity may be due to clinical heterogeneity or to chance. Statistical heterogeneity is measured with a variety of tests, most commonly I^2 and the Q statistic. Other heterogeneity measures (eg, H^2, R^2, tau^2) have also been described but are infrequently used.

I2 index — The I^2 index represents the amount of variability in the effect sizes across studies that can be explained by between-study variability. For example, an I^2 value of 75 percent means that 75 percent of the variability in the measured effect sizes across studies is caused by true heterogeneity among studies. By consensus, standard thresholds for the interpretation of I^2 are 25, 50, and 75 percent to represent low, medium, and high heterogeneity, respectively [60]. However, the investigators who introduced the I^2 statistic noted that naïve categorization of I^2 values is not appropriate in all circumstances and that "the practical impact of heterogeneity in a meta-analysis also depends on the size and direction of treatment effects" [60]. The clinical implication and interpretability of a meta-analysis with a large I^2 index will be different for studies with large statistically significant effects compared with studies with smaller inconsistent effects.

Key questions — Research questions that are clearly defined and form the basis for the systematic review or meta-analysis. (See 'Formulating research questions' above.)

Meta-regression — A meta-analytic technique that permits adjustment for potential confounders and analysis of different variables to help explain differences in results across studies. Equivalent to patient-level regression, except that the unit of analysis is a study instead of a person. (See 'Meta-regression' above.)

Network meta-analysis — A technique to simultaneously meta-analyze a network of studies that evaluated related, but different, specific comparisons. It permits quantitative inferences across studies that have made indirect comparisons of interventions. An example would be the comparison of two or more drugs to each other, when each was studied only in comparison to placebo. (See 'Network meta-analysis' above.)

PICO method (PICOD, PICOS, PICOTS, others) — An acronym that stands for **P**opulation, **I**ntervention(s), **C**omparator(s), **O**utcome(s); added letters include Study **D**esign (PICOD), **S**etting (PICOS), **T**iming and Setting (PICOTS). PICO is the basis for a systematic approach in developing a key question and research protocol. While used extensively for systematic reviews, PICO is relevant to all medical research questions. Each feature is defined explicitly and comprehensively so that it is unambiguously evident which studies are eligible for inclusion in a systematic review.

Precision — Precision refers to the extent to which the observed results would be reproduced exactly, given the same interventions and study design. The precision of an effect estimate can generally be assessed by examining the confidence intervals (ie, the narrower the confidence intervals are, the more precise the estimate is). (See "Glossary of common biostatistical and epidemiological terms", section on 'Confidence interval'.)

PRISMA statement — Preferred Reporting Items of Systematic reviews and Meta-Analyses, an update of QUOROM (Quality of Reporting of Meta-analyses) statement. A guideline for reporting of systematic reviews, used as a standard by many journals.

PROSPERO — An international database of prospectively registered systematic reviews in health care. PROSPERO creates a permanent record of systematic review protocols to reduce unnecessary duplication of efforts and increase transparency. Researchers should ideally enter their protocols prospectively and update them as necessary.

Publication bias — One of several related biases in the available evidence being considered for inclusion in a systematic review. Conceptually, studies that have been published are systematically different than studies that have failed to be published, due to lack of acceptance by journals, lack of interest by authors or research grantors, or potentially, by deliberate withholding by funders. Theoretically, "positive" (statistically significant) results are more likely to be published than "negative" results. Strictly, publication bias refers specifically to missing publications about studies.

Related biases include selective outcome reporting bias, where studies are published without certain outcomes; time-lag bias, where "negative" study results tend to be delayed in their publication compared with "positive" results; location bias, where "positive" or more interesting results tend to be published in journals that are more easily accessible; language bias, pertinent in certain fields, where non-English language publications differ in study results compared with those published (from the same countries or authors) in English; and multiple or duplicate publication bias, where certain studies may be overrepresented in the literature due to duplicate or overlapping publications (that may be difficult to tease apart).

Q statistic — The "Q" statistic (or chi square test for heterogeneity) tests the hypothesis that results across studies are homogeneous. Its calculation involves summing the squared deviations from the effect measured in each study from the overall effect and weighting the contribution from each study by the inverse of its variance. The Q statistic is usually interpreted to indicate heterogeneity if its P value is <0.10. A nonsignificant value suggests that the studies are homogeneous. However, the Q statistic has limited power to detect heterogeneity in meta-analyses with few studies, while it tends to over-detect heterogeneity in meta-analyses with many studies [61].

Random effects model — The central assumption of a random effects model is that each study estimate represents a random sample from a distribution of different populations [62]. For most of the medical literature, the random effects model is the more appropriate approach. A hypothetical model for a random effects model meta-analysis is shown in the figure (figure 10). The model assumes there are multiple true treatment effects related to inherent differences in different populations or other factors, and that each trial provides an estimate of its own true effect. The meta-analysis provides a pooled estimate across (or an average of) a range of true effects. Thus, the random effects model assumes that there is not necessarily one "true" effect size but rather that the studies included have provided a glimpse of a range of "true" effects. The random effects model incorporates both "between-study variance" (to capture the range of difference effects across studies) and "within-study variance" (to capture the range of difference effects within studies) [41]. There are several methods for calculating the random effects model estimates. The optimal approaches continue to be debated [63].

Risk of bias assessment — The risk of bias (RoB) assessment (sometimes referred to as "quality assessment") represents the extent to which trial design and methodology prevented systematic error and can help explain differences in the results of systematic reviews. The primary value of the RoB assessment of individual studies in the meta-analysis is to determine the degree of confidence that the pooled effect estimate reflects the "truth" as best as it can be measured. One would be more

likely to have high confidence in conclusions based upon "high-quality" (ie, low RoB) studies rather than "low-quality" (ie, high RoB) studies. (See 'Risk of bias assessment' above.)

Sensitivity analysis — A method of exploring heterogeneity in a meta-analysis by varying which studies are included to determine the effects of such changes. Used to explore how sensitive a meta-analysis finding is to inclusion of individual studies and to evaluate possible causes of heterogeneity; for example, whether exclusion of high RoB studies influences the size of the effect. (See 'Sensitivity analysis' above.)

SUMMARY

- A systematic review is a comprehensive summary of all available evidence that meets predefined eligibility criteria to address a specific clinical question or range of questions. Meta-analysis, which is commonly included in systematic reviews, is a statistical method that quantitatively combines the results from different studies. It is commonly used to provide an overall pooled estimate of the benefit or harm of an intervention. (See 'Key definitions' above.)

- Several steps are essential for conducting a systematic review or meta-analysis. These include:

  - Formulating research questions (see 'Formulating research questions' above)
  - Developing a protocol (see 'Developing a protocol' above)
  - Searching for the evidence (see 'The literature search' above)
  - Assessing the quality of studies (see 'Risk of bias assessment' above)
  - Summarizing and displaying results (eg, using forest pots and a summary of findings table, as shown in the figure (figure 1)) (see 'Forest plot' above)
  - Exploring reasons for heterogeneity across studies (see 'Heterogeneity' above and 'Subgroup analyses' above and 'Sensitivity analysis' above)

- When reading and interpreting a systematic review, the reader should appraise the methodologic quality, assess for potential sources of bias, and consider the extent to which the findings are applicable to their specific question. Key issues to consider are summarized in the table (table 2). The value of a systematic review's conclusions may be limited by the

quality and applicability of the individual studies included in the review. (See 'Reading and interpreting a systematic review' above.)

REFERENCES

1. Systematic reviews, Chalmers I, Altman DG (Eds), BMJ Publishing Group, London 1995.
2. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med 1997; 126:376.
3. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. JAMA 1994; 272:1367.
4. Mulrow CD. The medical review article: state of the science. Ann Intern Med 1987; 106:485.
5. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. Ann Intern Med 1997; 127:820.
6. Ioannidis JP. Why most published research findings are false: author's reply to Goodman and Greenland. PLoS Med 2007; 4:e215.
7. LeLorier J, Grégoire G, Benhaddad A, et al. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997; 337:536.
8. Cappelleri JC, Ioannidis JP, Schmid CH, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA 1996; 276:1332.
9. Villar J, Carroli G, Belizán JM. Predictive ability of meta-analyses of randomised controlled trials. Lancet 1995; 345:772.
10. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021; 372:n71.
11. www.prisma-statement.org/Extensions/Default.aspx (Accessed on April 16, 2018).
12. Institute of Medicine. Finding what workds in health care: Standards for systematic reviews. The National Academies Press, Washington, DC, 2011. Available at: http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-S

ystematic-Reviews.aspx (Accessed on October 10, 2011).

13. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997; 127:380.
14. Clarke MJ, Stewart LA. Principles of and procedures for systematic reviews. In: Systematic reviews in health care: meta-analysis in context, Egger M, Smith G, Altman D (Eds), BMJ Publishing Group, London 2001. p.23.
15. http://www.ncbi.nlm.nih.gov/pubmed (Accessed on July 14, 2011).
16. http://www.embase.com/search (Accessed on July 14, 2011).
17. http://onlinelibrary.wiley.com/o/cochrane/cochrane_clcentral_articles_fs.html (Accessed on July 14, 2011).
18. Dickersin K, Chalmers I. Recognising, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the World Health Organisation. James Lind Library 2010. Available at: www.jameslindlibrary.org (Accessed on October 10, 2011).
19. Mathieu S, Boutron I, Moher D, et al. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA 2009; 302:977.
20. Kirkham JJ, Altman DG, Williamson PR. Bias due to changes in specified outcomes during the systematic review process. PLoS One 2010; 5:e9810.
21. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016; 375:2293.
22. Briere JB, Bowrin K, Taieb V, et al. Meta-analyses using real-world data to generate clinical and epidemiological evidence: a systematic literature review of existing recommendations. Curr Med Res Opin 2018; 34:2125.
23. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. J Clin Epidemiol 2000; 53:207.
24. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 1998; 279:281.
25. Vevea JL, Woods CM. Publication bias in research synthesis: sensitivity analysis using a priori weight functions. Psychol Methods 2005; 10:428.

26. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315:629.
27. Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. J Clin Epidemiol 2000; 53:477.
28. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. Stat Med 2003; 22:2113.
29. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol 2005; 58:894.
30. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000; 56:455.
31. Copas J. What works?: Selectivity models and meta-analysis. Journal of the Royal Statistical Society Series A 1999; 162:95.
32. Rosenthal R. The 'file drawer problem' and tolerance for null results. Psychol Bull 1979; 86:638.
33. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. Clin Trials 2007; 4:245.
34. Moher D, Jadad AR, Nichol G, et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995; 16:62.
35. Higgins JP, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011; 343:d5928.
36. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019; 366:l4898.
37. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355:i4919.
38. Verhagen AP, de Vet HC, de Bie RA, et al. The art of quality assessment of RCTs included in systematic reviews. J Clin Epidemiol 2001; 54:651.
39. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA 2010; 303:1180.
40. Walter SD, Guyatt GH, Bassler D, et al. Randomised trials with provision for early stopping for benefit (or harm): The impact on the estimated treatment effect. Stat Med 2019; 38:2524.

41. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7:177.
42. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 2011; 64:1283.
43. Dechartres A, Altman DG, Trinquart L, et al. Association between analytic strategy and estimates of treatment outcomes in meta-analyses. JAMA 2014; 312:623.
44. Berlin JA, Golub RM. Meta-analysis as evidence: building a better pyramid. JAMA 2014; 312:603.
45. Povitz M, Bolo CE, Heitman SJ, et al. Effect of treatment of obstructive sleep apnea on depressive symptoms: systematic review and meta-analysis. PLoS Med 2014; 11:e1001762.
46. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. Stat Med 1995; 14:395.
47. Schmid CH. Exploring heterogeneity in randomized trials via metaanalysis. Drug Inf J 1999; 33:211.
48. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002; 21:589.
49. Ioannidis JP, Cappelleri JC, Sacks HS, Lau J. The relationship between study design, results, and reporting of randomized clinical trials of HIV infection. Control Clin Trials 1997; 18:431.
50. Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. Lancet 2005; 365:1657.
51. Rothman KJ, Greenland S. Modern epidemiology, 2nd ed, Lippincott-Raven, Philadelphia 1998.
52. Geissbühler M, Hincapié CA, Aghlmandi S, et al. Most published meta-regression analyses based on aggregate data suffer from methodological pitfalls: a meta-epidemiological study. BMC Med Res Methodol 2021; 21:123.
53. Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value Health 2011; 14:417.
54. Mills EJ, Ioannidis JP, Thorlund K, et al. How to use an article reporting a multiple treatment comparison meta-analysis. JAMA 2012; 308:1246.

55. Brignardello-Petersen R, Mustafa RA, Siemieniuk RAC, et al. GRADE approach to rate the certainty from a network meta-analysis: addressing incoherence. J Clin Epidemiol 2019; 108:77.
56. Brignardello-Petersen R, Bonner A, Alexander PE, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. J Clin Epidemiol 2018; 93:36.
57. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods 2012; 3:80.
58. Trinquart L, Attiche N, Bafeta A, et al. Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. Ann Intern Med 2016; 164:666.
59. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med 2007; 26:53.
60. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327:557.
61. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods 2006; 11:193.
62. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. Lancet 1998; 351:123.
63. Jackson D, Law M, Stijnen T, et al. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. Stat Med 2018; 37:1059.

Topic 16293 Version 24.0

GRAPHICS

Corticosteroids compared with no corticosteroids for treatment of acute respiratory distress syndrome w

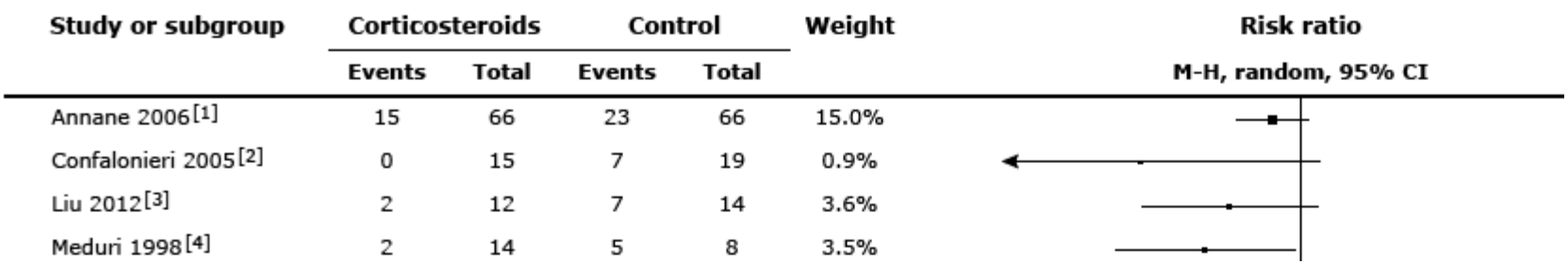
A

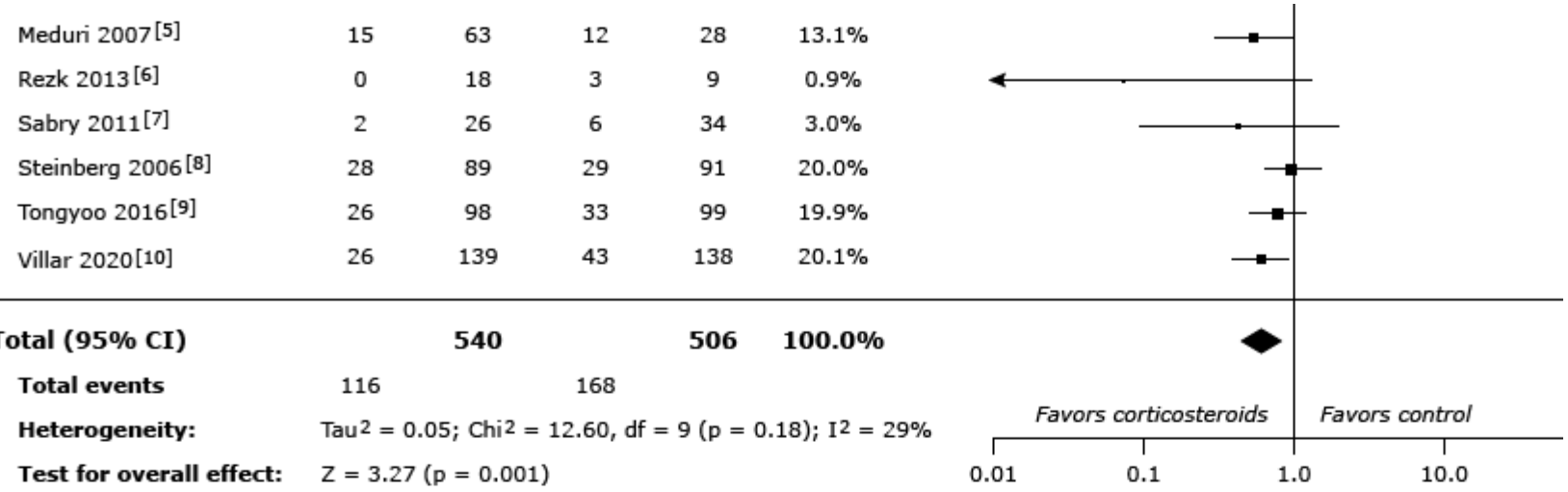
Summary of evidence for corticosteroids in the treatment of ARDS[1-10]

Outcomes	Number of participants (studies)	Certainty of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute	
				Risk without corticosteroids	Risk diff corticc
Mortality* Follow-up – Range 8 to 60 days	851 (7 RCTs)	⨁⨁⨁◐ **Moderate to high¶**	**RR 0.66** (0.53 to 0.84)	333 per 1000	113 few (53 to
Days free from mechanical ventilationΔ Follow-up – 28 days	837 (6 RCTs)	⨁⨁⨁ **Moderate** due to imprecision◇	**MD 4.0 days** (0.5 to 7.5)	8.0 days	4 m (0.5 to
Days not in the ICU§ Follow-up – 28 days	213 (2 RCTs)	⨁⨁ **Low** due to imprecision◇	**MD 7.0 days** (–2.6 to 16.6)	4.8 days	7.0 m (2.6 fewer
Nosocomial infection	1094 (10 RCTs)	⨁⨁⨁ **Moderate** due to imprecision◇	**RR 0.83** (0.67 to 1.02)	232 per 1000	39 fewe (77 fewe
Gastrointestinal bleeding	617 (6 RCTs)	⨁⨁ **Low** due to imprecision◇	**RR 1.33** (0.65 to 2.71)	42 per 1000	14 mor (15 fewe
Hyperglycemia¥	288 (2 RCTs)	⨁⨁⨁ **Moderate** due to imprecision‡	**RR 1.17** (1.01 to 1.36)	660 per 1000	112 mo (7 to
Neuromuscular weakness	316 (3 RCTs)	⨁ **Very low** due to imprecision†	**RR 1.15** (0.53 to 2.49)	130 per 1000	20 mor (61 fewer

B

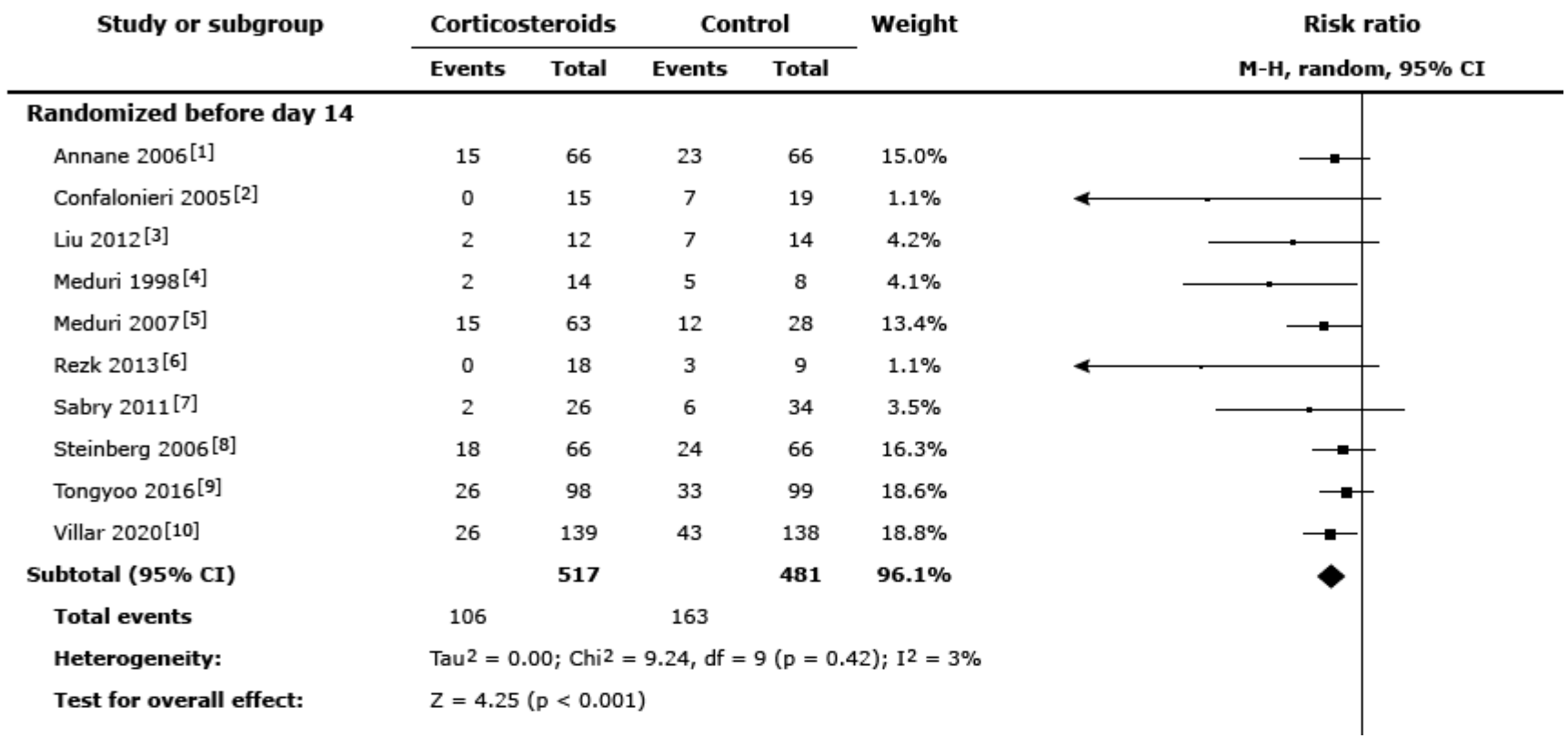
Overall mortality at 8 to 60 days



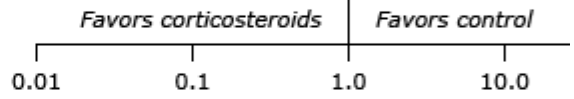


C

Mortality at 8 to 60 days, patients randomized before day 14 versus on or after day 14



Randomized day 14 or later					
Meduri 1998[4]	0	2	0	0	
Steinberg 2006[8]	8	23	2	25	3.9%
Subtotal (95% CI)		**25**		**25**	**3.9%**
Total events	8		2		
Heterogeneity:	Not applicable				
Test for overall effect:	Z = 2.00 (p = 0.05)				
Total (95% CI)		**542**		**506**	**100.0%**
Total events	114		165		
Heterogeneity:	$Tau^2 = 0.08$; $Chi^2 = 15.86$, df = 10 (p = 0.10); $I^2 = 37\%$				
Test for overall effect:	Z = 2.99 (p = 0.003)				
Test for subgroup differences:	$Chi^2 = 6.80$, df = 1 (p = 0.009); $I^2 = 85.3\%$				

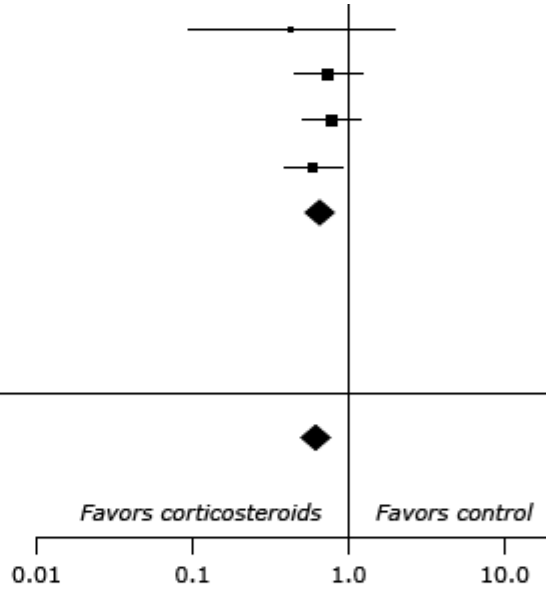


D

Mortality at 8 to 60 days, trials stopped early for benefit versus not stopped early for benefit

Study or subgroup	Corticosteroids Events	Corticosteroids Total	Control Events	Control Total	Weight	Risk ratio M-H, random, 95% CI
Trials stopped early for benefit						
Confalonieri 2005[2]	0	15	7	19	0.6%	
Meduri 1998[4]	2	14	5	8	2.4%	
Meduri 2007[5]	15	63	12	28	12.0%	
Subtotal (95% CI)		**92**		**55**	**15.0%**	
Total events	17		24			
Heterogeneity:	$Tau^2 = 0.27$; $Chi^2 = 3.13$, df = 2 (p = 0.21); $I^2 = 36\%$					
Test for overall effect:	Z = 2.17 (p = 0.03)					
Trials not stopped early for benefit						
Annane 2006[1]	15	66	23	66	14.7%	
Liu 2012[3]	2	12	7	14	2.5%	
Rezk 2013[6]	0	18	3	9	0.6%	

Sabry 2011[7]	2	26	6	34	2.0%
Steinberg 2006[8]	18	66	24	66	17.4%
Tongyoo 2016[9]	26	98	33	99	23.6%
Villar 2020[10]	26	139	43	138	24.2%
Subtotal (95% CI)		**425**		**426**	**85.0%**
Total events	89		139		
Heterogeneity:	$Tau^2 = 0.00$; $Chi^2 = 4.68$, df = 6 (p = 0.59); $I^2 = 0\%$				
Test for overall effect:	Z = 3.50 (p = 0.0005)				
Total (95% CI)		**517**		**481**	**100.0%**
Total events	106		163		
Heterogeneity:	$Tau^2 = 0.00$; $Chi^2 = 9.24$, df = 9 (p = 0.42); $I^2 = 3\%$				
Test for overall effect:	Z = 4.25 (p < 0.0001)				
Test for subgroup differences:	$Chi^2 = 1.63$, df = 1 (p = 0.20); $I^2 = 38.7\%$				

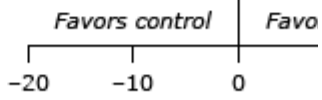


E

Ventilator-free days within first 28 days, trials stopped early for benefit versus not stopped early for benefit

Study or subgroup	Corticosteroids Mean	SD	Total	Control Mean	SD	Total	Weight	Mean differen IV, random, 95%
Trials stopped early for benefit								
Confalonieri 2005[2]	22	6.2	15	10.1	10.2	19	9.4%	
Meduri 1998[4]	14	7	14	4	6.2	8	9.3%	
Meduri 2007[5]	16.5	10.1	63	8.7	10.2	28	10.6%	
Subtotal (95% CI)			**92**			**55**	**29.3%**	
Heterogeneity:	$Tau^2 = 0.00$; $Chi^2 = 1.28$, df = 2 (p = 0.53); $I^2 = 0\%$							
Test for overall effect:	Z = 6.31 (p < 0.00001)							
Trials not stopped early for benefit								
Annane 2006[1]	4.9	8.4	85	3.1	6.9	92	13.1%	
Liu 2012[3]	13.9	11.3	12	12.8	11.3	14	6.3%	
Rezk 2013[6]	17.4	2.9	18	6.8	2.2	9	13.4%	
Steinberg 2006[8]	10.3	9.7	67	8.6	8.5	66	12.3%	

Tongyoo 2016[9]	14.4	10.6	98	12.3	10.9	99	12.4%
Villar 2020[10]	12.3	9.9	139	7.5	9	138	13.2%
Subtotal (95% CI)			**419**			**418**	**70.7%**
Heterogeneity:	$Tau^2 = 16.02$; $Chi^2 = 47.52$, df = 5 ($p < 0.00001$); $I^2 = 89\%$						
Test for overall effect:	$Z = 2.23$ ($p = 0.03$)						
Total (95% CI)			**511**			**473**	**100.0%**
Heterogeneity:	$Tau^2 = 15.41$; $Chi^2 = 56.59$, df = 8 ($p < 0.00001$); $I^2 = 86\%$						
Test for overall effect:	$Z = 3.84$ ($p = 0.0001$)						
Test for subgroup differences:	$Chi^2 = 5.66$, df = 1 ($p = 0.02$); $I^2 = 82.3\%$						

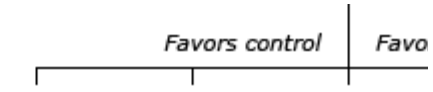


F

ICU-free days within first 28 days, trials stopped early for benefit versus not stopped early for benefit

Study or subgroup	Corticosteroids			Control			Weight	Mean differen
	Mean	SD	Total	Mean	SD	Total		IV, random, 95%
Trials stopped early for benefit								
Confalonieri 2005[2]	9.4	7.5	14	3	5.3	8	17.7%	
Meduri 1998[4]	21.6	6.3	15	10.3	9.2	19	17.9%	
Meduri 2007[5]	13.1	9.1	63	6.1	8	28	20.3%	
Subtotal (95% CI)			**92**			**55**	**55.9%**	
Heterogeneity:	$Tau^2 = 0.44$; $Chi^2 = 2.15$, df = 2 ($p = 0.34$); $I^2 = 7\%$							
Test for overall effect:	$Z = 5.67$ ($p < 0.00001$)							
Trials not stopped early for benefit								
Rezk 2013[6]	15.2	2.9	18	3.3	3	9	22.0%	
Steinberg 2006[8]	8.4	8.4	89	6.3	7.8	97	22.1%	
Subtotal (95% CI)			**107**			**106**	**44.1%**	
Heterogeneity:	$Tau^2 = 46.58$; $Chi^2 = 33.26$, df = 1 ($p < 0.00001$); $I^2 = 97\%$							
Test for overall effect:	$Z = 1.43$ ($p = 0.15$)							
Total (95% CI)			**199**			**161**	**100.0%**	

Heterogeneity:	$Tau^2 = 23.03$; $Chi^2 = 35.84$, df = 4 ($p < 0.00001$); $I^2 = 89\%$
Test for overall effect:	$Z = 3.30$ ($p = 0.0010$)
Test for subgroup differences:	$Chi^2 = 0.04$, df = 1 ($p = 0.85$); $I^2 = 0\%$



GRADE Working Group grades of evidence:

- **High certainty –** We are very confident that the true effect lies close to that of the estimate of the effect.
- **Moderate certainty –** We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, substantially different.
- **Low certainty –** Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the ef
- **Very low certainty –** We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the est

(Panels B to F) Forest plots showing effect sizes in the individual trials, pooled effect estimates, and subgroup analysis for the different outcome

- **(B)** Overall mortality at 8 to 60 days.
- **(C)** Mortality at 8 to 60 days, patients randomized before day 14 versus on or after day 14.
- **(D)** Mortality at 8 to 60 days, trials stopped early for benefit versus not stopped early for benefit.
- **(E)** Ventilator-free days within first 28 days, trials stopped early for benefit versus not stopped early for benefit.
- **(F)** ICU-free days within first 28 days, trials stopped early for benefit versus not stopped early for benefit.

ARDS: acute respiratory distress syndrome; GRADE: Grading of Recommendations Assessment, Development and Evaluation; CI: confidence inte risk ratio; MD: mean difference; ICU: intensive care unit; M-H: Mantel-Haenszel; df: degrees of freedom; SD: standard deviation; IV: inverse variar pneumonia.

* The effect estimates for this outcome are based on the pooled effect for patients who received steroids before day 14 in trials that were not sto detected a subgroup effect for patients randomized before day 14 versus on or after day 14 (panel C), so we restricted our analysis to patients ra detect a significant subgroup effect on mortality in trials stopped early for benefit versus not stopped early for benefit (panel D); however, we dic this when assessing ventilator-free days (panel E). We decided to trust the more conservative estimate based on the 7 trials that were not stoppe

¶ We rated down for the following concerns:

1. Indirectness in the outcome: The duration of follow-up in 2 trials was relatively short (<15 days)[6,7].
2. Indirectness in the patient population: We are interested in the effect of steroids in patients with ARDS managed with low tidal volume ver described use of low tidal volume ventilation[8-10], 1 trial described using larger tidal volumes[1], and 3 trials did not provide details rega enrolled only patients with severe sepsis[1,9] and 1 trial included only patients with CAP[27]; it's possible that the effect of steroids in thes shock rather than directly on ARDS.
3. Risk of bias: 2 trials provided only limited details regarding randomization and allocation concealment[6,7], and we had limited informatio it was not published in English[3]. A previous meta-analysis[11] did not detect a subgroup effect based on risk of bias but did report a stat comparing smaller trials (<60 patients) versus larger trials (≥60 patients), with a larger effect size in smaller trials.

While each concern alone was not serious enough to warrant rating down, taken together, they may limit the certainty of our findings. We there moderate to high for this outcome.

Δ We decided to trust the more conservative estimate based on the 6 trials that were not stopped early for benefit (panel E).

◇ The 95% CI for the absolute effect includes possible harm with steroids.

§ We decided to trust the more conservative estimate based on the 2 trials that were not stopped early for benefit (panel F).

¥ The 95% CI for the absolute effect approaches no effect.

‡ The 95% CI for the absolute effect varies broadly from a considerable reduction to a considerable increase in neuromuscular weakness with ste

† Defined as blood glucose >150 mg/dL or requiring insulin therapy.

References:

1. *Annane D, Sébille V, Bellissant E, Ger-Inf-05 Study Group. Effect of low doses of corticosteroids in septic shock patients with or without early acute respiratory distress syndro*
2. *Confalonieri M, Urbino R, Potena A, et al. Hydrocortisone infusion for severe community-acquired pneumonia: a preliminary randomized study. Am J Respir Crit Care Med 2(*
3. *Liu L, Li J, Huang YZ, et al. The effect of stress dose glucocorticoid on patients with acute respiratory distress syndrome combined with critical illness-related corticosteroid in*
4. *Meduri GU, Headley AS, Golden E, et al. Effect of prolonged methylprednisolone therapy in unresolving acute respiratory distress syndrome: a randomized controlled trial. JA*
5. *Meduri GU, Golden E, Freire AX, et al. Methylprednisolone infusion in early severe ARDS: results of a randomized controlled trial. Chest 2007; 131:954.*
6. *Rezk NA, Ibrahim AM. Effects of methylprednisolone in early ARDS. Egyptian Journal of Chest Diseases and Tuberculosis 2013; 62:167.*
7. *Sabry NA, Omar EE. Corticosteroids and ICU course of community acqured pneumonia in Egyptian settings. Pharmacology & Pharmacy 2011; 2:73.*
8. *Steinberg KP, Hudson LD, Goodman RB, et al. Efficacy and safety of corticosteroids for persistent acute respiratory distress syndrome. N Engl J Med 2006; 354:1671.*
9. *Tongyoo S, Permpikul C, Mongkolpun W, et al. Hydrocortisone treatment in early sepsis-associated acute respiratory distress syndrome: results of a randomized controlled t*
10. *Villar J, Ferrando C, Martínez D, et al. Dexamethasone treatment for the acute respiratory distress syndrome: a multicentre, randomised controlled trial. Lancet Respir Med 2*
11. *Meduri GU, Bridges L, Shih MC, et al. Prolonged glucocorticoid treatment is associated with improved ARDS outcomes: analysis of individual patients' data from four rando updated literature. Intensive Care Med 2016; 42:829.*

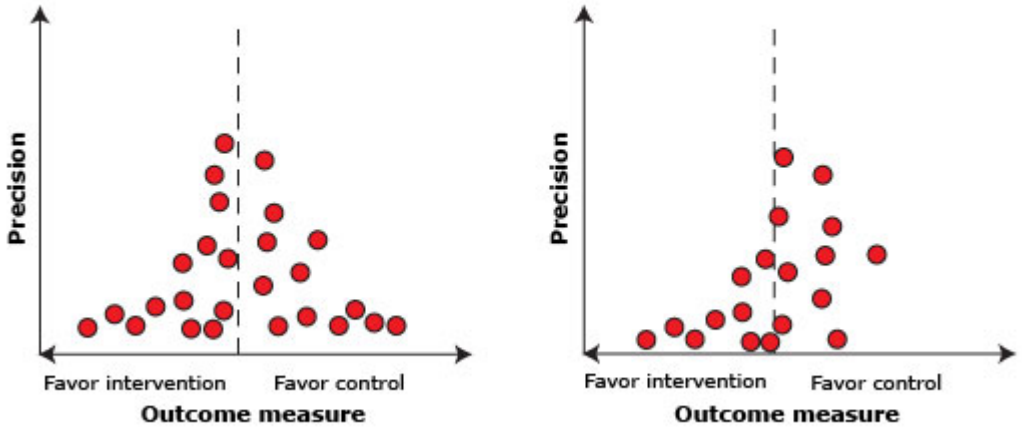
Graphic 121523 Version 2.0

Methods of summarizing studies in systematic reviews

Approach	Main model	Comments
Qualitative systematic review	Qualitative description of evidence	May be most appropriate if large degree of heterogeneity among studies.
Meta-analysis overall estimate	Fixed effects model	Assumes a single truth across populations and homogeneity among studies.
	Random effects model	Incorporates between-study heterogeneity. Wider confidence interval than fixed effects model when heterogeneity is present.
Meta-analysis of multiple interventions simultaneously	Network meta-analysis (mixed or multiple treatment comparison)	Estimates relative effects of multiple interventions against each other, even if no individual study analyzed a comparison between 2 or more specific interventions.
Meta-analysis subgroup analyses	Fixed or random effects models	Estimates treatment effects for each group. May explain heterogeneity. Subject to ecological fallacy. May not be possible due to limited reported data. Arbitrary selection of subgroups may result in spurious findings.
Meta-regression	Regression across studies	Tests interaction between subgroup and treatment effects. Can test continuous or categorical variables singly or in a multivariable analysis. May explain heterogeneity. Subject to ecological fallacy. May not be possible due to limited reported data. Arbitrary selection of subgroups may reult in spurious findings.
Individual patient data meta-analysis	Multivariate regression across individuals	Allows most complete analysis of data and evaluation of heterogeneity. Costly and resource-intensive.

Graphic 80657 Version 3.0

Funnel plots



The above graphs represent funnel plots. In each figure, the x-axis represents the magnitude of the effect and the y-axis the "precision." The dotted line on the x-axis represents "no effect"; to the right of this line, the effect favors the control group while to the left the effect favors the intervention. Circles represent individual studies. Generally, larger studies have higher precision.

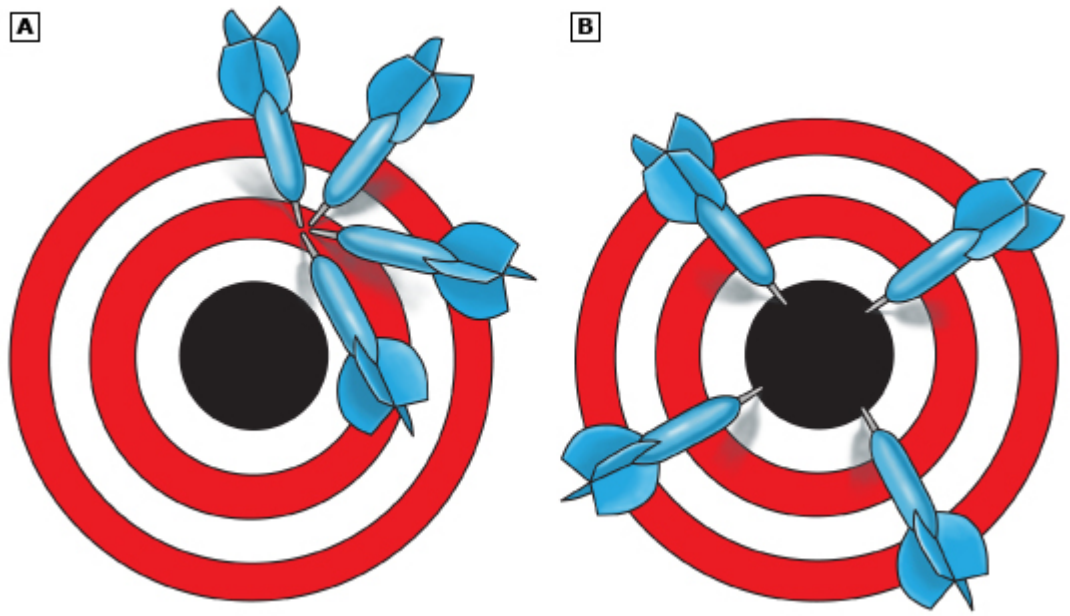
The funnel plot on the left shows several smaller studies symmetrically clustered around the "no effect" line. In contrast, the funnel plot on the right appears to be missing smaller studies that favored the control arm, suggesting that there may have been publication bias in favor of studies where the intervention succeeded.

Unfortunately, the graphical depiction of studies using this approach is subject to variable interpretation among raters, particularly when there are relatively few studies. Thus, funnel plots are not always a reliable method to clarify whether there is publication bias.

Figure courtesy of Gordon Guyatt, MD.

Graphic 65012 Version 3.0

Precision and validity

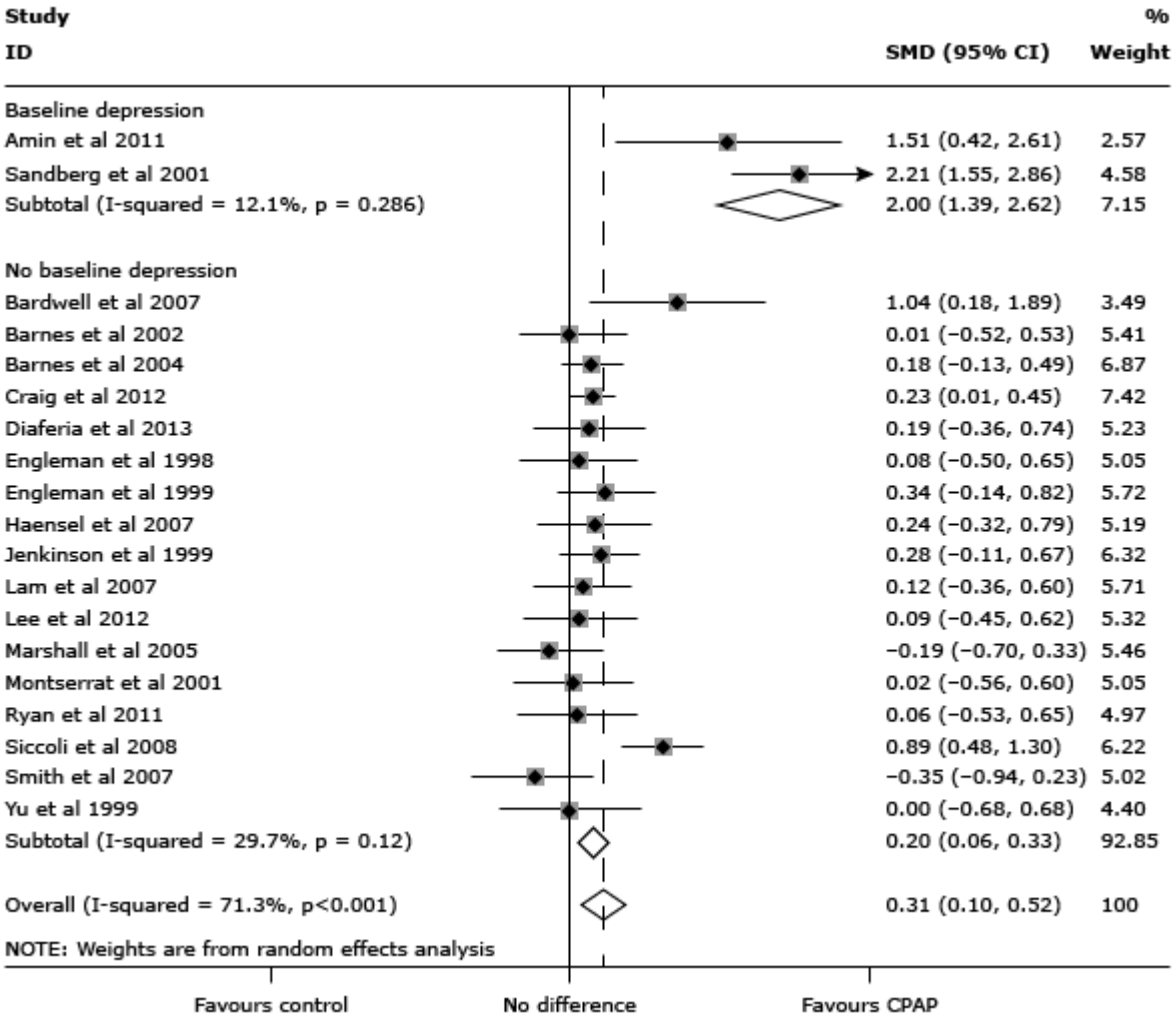


(A) Dartboard illustrating high precision but low validity. The darts can be interpreted as representing studies attempting to measure the true effect (the bullseye). The studies closely approximate an estimate, but it is not the "true" effect.

(B) Dartboard illustrating low precision but high validity. As in the previous example, the darts can be interpreted as representing studies attempting to measure the "true" effect (the bullseye). In this case, the studies closely approximate the "true" effect (the bullseye), but the studies vary in their individual estimates. Interpretation of group of studies like this give a less confident assessment that they have revealed the "truth" compared with the group of studies that gave highly precise estimates (arrows tightly clustered within the bullseye).

Graphic 52219 Version 4.0

CPAP studies forest plot

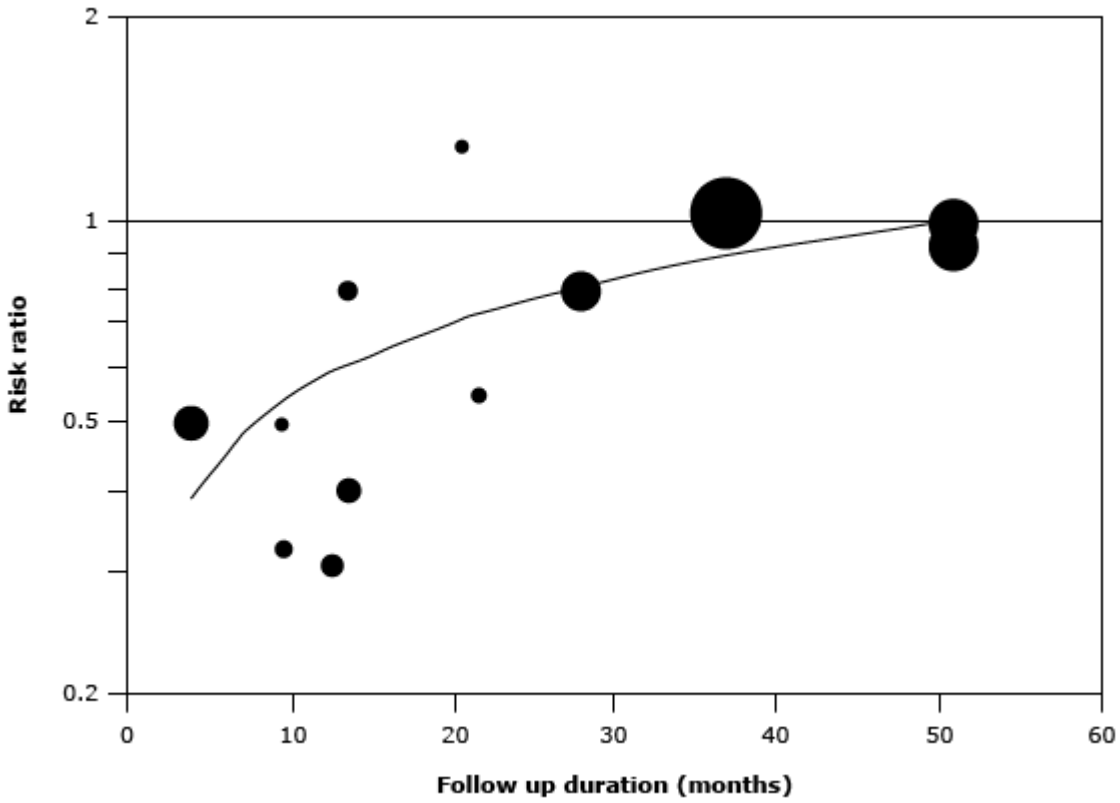


Data were calculated by a random effects model. Studies were stratified by baseline depression score. Boxes are SMDs, and lines are 95% CIs. The vertical solid line represents no difference between CPAP and control. Values to the right of the solid line favor CPAP benefit. Pooled SMDs and 95% CIs are represented by the diamond shapes.

SMD: standardized mean difference; CPAP: continuous positive airway pressure.

Graphic 104212 Version 1.0

Meta-regression of the association between follow-up duration and risk ratio of death in trials of zidovudine monotherapy



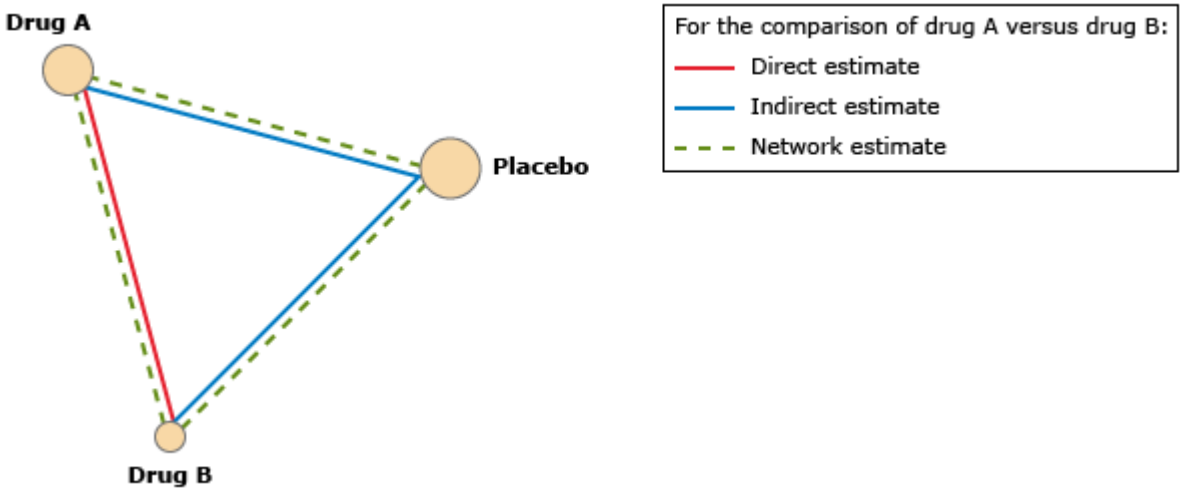
The relationship between study design, results, and reporting of randomized clinical trials of HIV infection. Decrease in the magnitude of the treatment effect with increasing mean duration of follow-up in studies of zidovudine monotherapy. Studies are represented by dots the size of which is proportional to the inverse of the variance of the observed logarithm of the risk ratio. The depicted regression line is weighted by the inverse of the variance.

Reproduced from: Ioannidis JP, Cappelleri JC, Sacks HS, Lau J. The relationship between study design, results, and reporting of randomized clinical trials of HIV infection. Control Clin Trials 1997; 18:431. Illustration

---

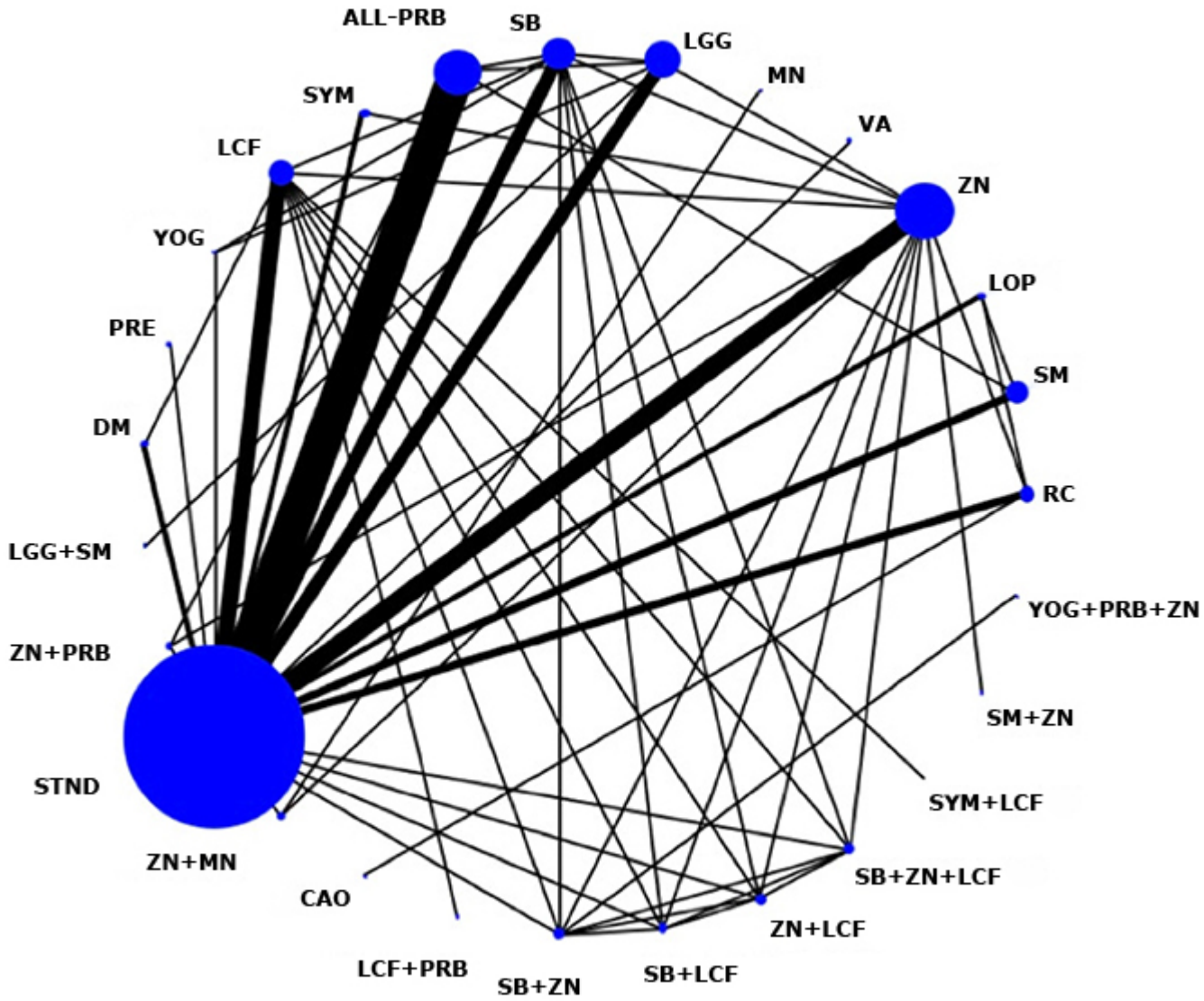
Graphic 69060 Version 1.0

Schematic diagram showing key concepts in network meta-analysis



Graphic 132754 Version 1.0

Example of a network diagram from a network meta-analysis: Interventions for acute diarrhea in children



The network diagram visually conveys the size and complexity of the network in NMA. This example shows the network diagram for an NMA that evaluated 62 clinical trials (20,256 participants)

evaluating different treatments for acute diarrhea in children. The "nodes" (blue dots) represent different treatments; the size of each node corresponds to the number of participants who received that treatment. The "edges" (black lines connecting different pairs of dots) represent trial(s) directly comparing the 2 treatments; the thickness of the line corresponds to the number of trials.

NMA: network meta-analysis; ALL-PRB: all probiotics, except LGG and SB; CAO: kaolin-pectin; DM: diluted milk; LCF: lactose-free formula; LCF+PRB: lactose-free formula plus probiotics; LGG: *Lacticaseibacillus rhamnosus* GG; LGG+SM: LGG plus smectite; LOP: loperamide; MN: micronutrients; PRE: prebiotics; RC: racecadotril; SB: *Saccharomyces boulardii*; SB+LCF: *S. boulardii* plus lactose-free formula; SB+ZN: *S. boulardii* plus zinc; SB+ZN+LCF: *S. boulardii* plus zinc plus lactose-free formula; SM: smectite; SM+ZN: smectite plus zinc; STND: standard treatment or placebo; SYM: symbiotics; SYM+LCF: symbiotics plus lactose-free formula; VA: vitamin A; YOG: yogurt; YOG+PRB+ZN: yogurt plus probiotics plus zinc; ZN: zinc; ZN+LCF: zinc plus lactose-free formula; ZN+MN: zinc plus micronutrients; ZN+PRB: zinc plus probiotics.

Graphic 132756 Version 1.0

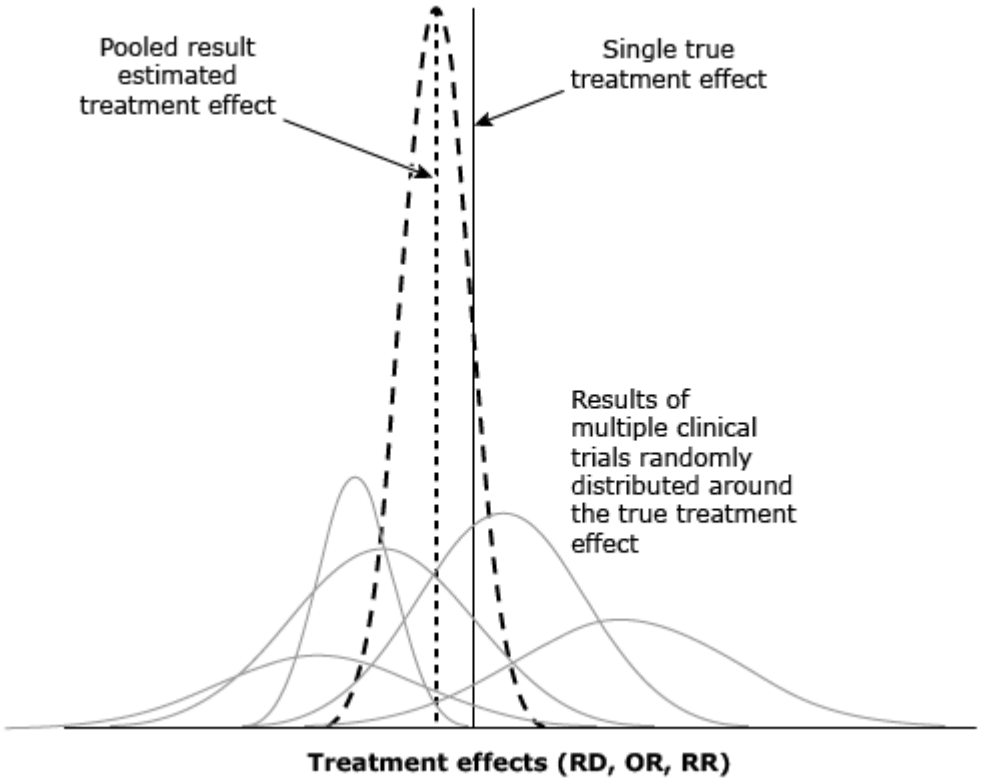
Questions to consider when interpreting a systematic review and meta-analysis

Were the systematic review and meta-analysis performed according to an explicitly defined protocol?
Were the key questions well formulated, and would their answers be clinically useful?
Did the authors clearly define the eligibility criteria for studies to be included? Eligibility criteria should clearly define **all** of the following: ▪ Study design (eg, randomized controlled trial versus observational study) ▪ Population of interest ▪ Treatment(s) and comparison(s) of interest ▪ Outcome(s) of interest
Were the eligibility criteria appropriate to capture all relevant studies? ▪ Were the population, treatment(s), comparison(s), and outcome(s) relevant to clinical practice? ▪ Were the study designs of included studies appropriate for addressing the key questions?
Was the search for relevant studies comprehensive/exhaustive? ▪ Was the search strategy reported in sufficient detail that it could be reproduced? ▪ Were important sources of "grey" literature included (eg, unpublished data)? ▪ Were the selection and assessment of studies reproducible (ie, done independently by 2 or more separate reviewers)? ▪ Were adequate explanations given for exclusion of studies?
Were the characteristics of the individual studies listed with sufficient detail to allow an assessment of the appropriateness of their inclusion?
Were the individual studies assessed for their methodologic quality (ie, risk of bias assessment)?
Was publication bias considered?
Were the statistical methods for combining results (meta-analysis) described?
Was the reporting of results clear? ▪ Were the pooled effect estimates presented with corresponding confidence intervals (rather than p values)? ▪ Were primary results for each individual study also included?

Was between-study heterogeneity assessed? ▪ Did the review attempt to explain between-study heterogeneity (ie, by performing subgroup and/or sensitivity analyses)? ▪ If subgroup analyses were performed, were they limited to only a few and were they specified a priori?
Did the authors clearly explain their conclusions, and did they address the overall certainty of effect estimates? ▪ Did they rate the certainty (or quality) of evidence for each outcome (eg, in a summary of findings table)? ▪ Was the methodologic quality (risk of bias) of the individual studies considered in formulating the overall certainty ratings?
Were limitations of the meta-analysis discussed?

Graphic 56270 Version 2.0

Fixed effects model



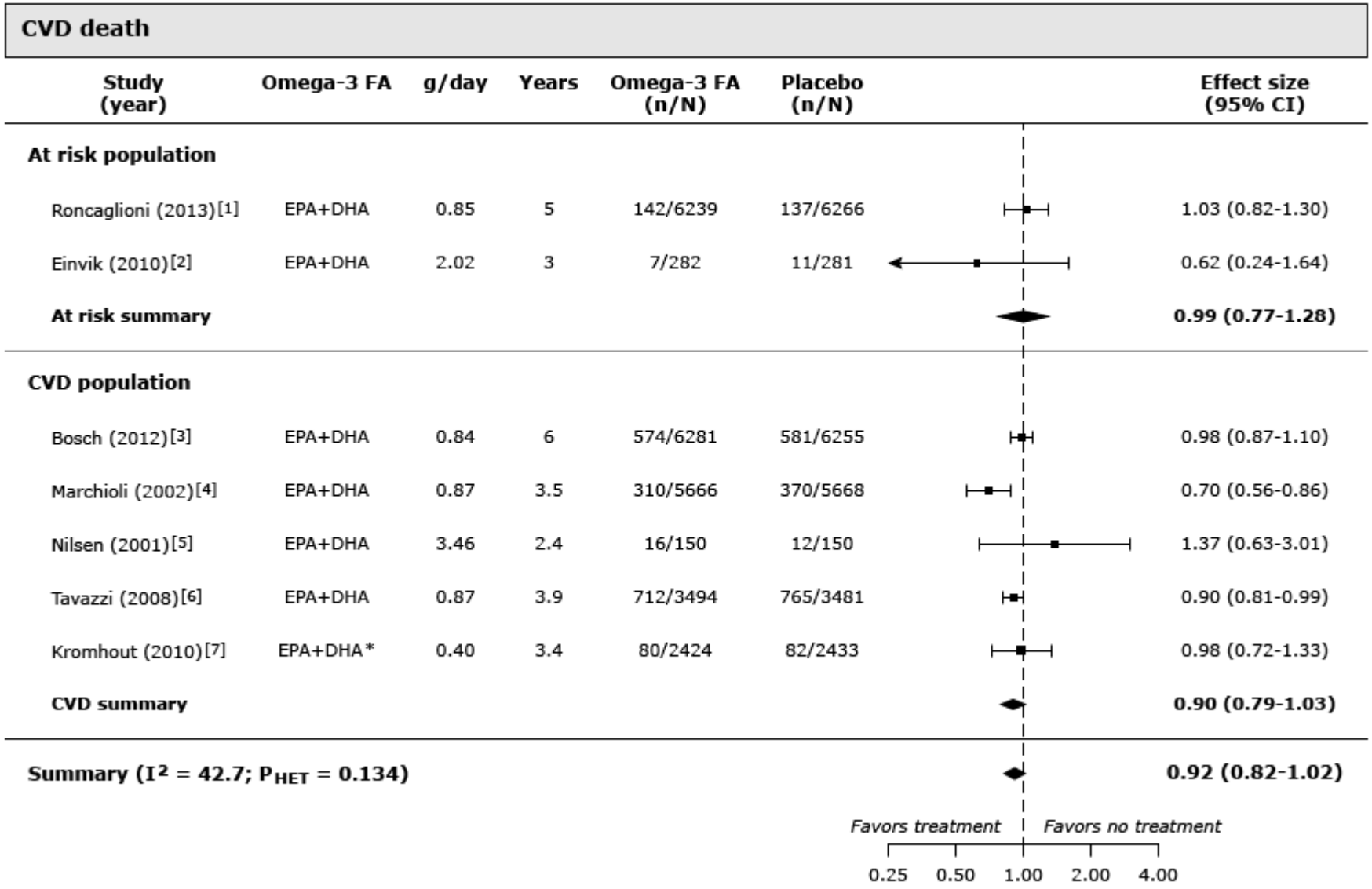
This model assumes there is a single true treatment effect and that all trials provide estimates of this one true effect.

RD: risk difference; OR: odds ratio; RR: relative risk.

Courtesy of: Joseph Lau, MD.

Graphic 80325 Version 3.0

Forest plot of randomized trials of fish oils in at-risk populations and in patients with cardiovascular disease assessing the outcome of cardiovascular death



Random effects model meta-analysis of omega-3 FA versus placebo (or no omega-3 FA), with subgroup analyses by population. In all meta-analyses, only studies that reported sufficient data are included.

CVD: cardiovascular disease; FA: fatty acids; n/N: number with outcome/number analyzed; CI: confidence interval; EPA: eicosapentaenoic acid; DHA: docosahexaenoic acid; P_{HET}: p-value of the test for statistical heterogeneity.

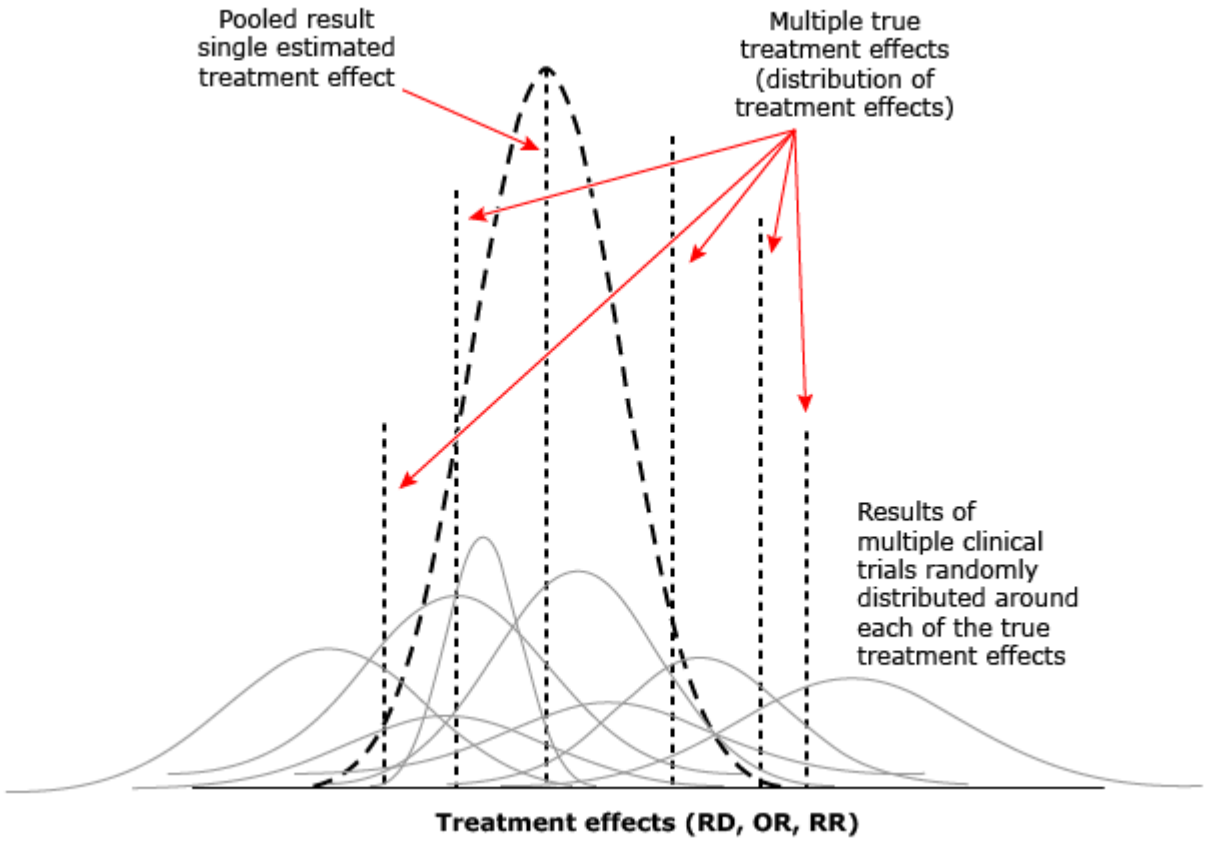
References:

1. *Roncaglioni MC, Tombesi M, Avanzini F, et al. n-3 fatty acids in patients with multiple cardiovascular risk factors. N Engl J Med 2013; 368:1800.*
2. *Einvik G, Klemsdal TO, Sandvik L, et al. A randomized clinical trial on n-3 polyunsaturated fatty acids supplementation and all-cause mortality in elderly men at high cardiovascular risk. Eur J Cardiovasc Prev Rehabil 2010; 17:588.*
3. *Bosch J, Gerstein HC, Dagenais GR, et al. n-3 fatty acids and cardiovascular outcomes in patients with dysglycemia. N Engl J Med 2012; 367:309.*
4. *Marchioli R, Barzi F, Bomba E, et al. Early protection against sudden death by n-3 polyunsaturated fatty acids after myocardial infarction: Time-course analysis of the results of the Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico (GISSI)-Prevenzione. Circulation 2002; 105:1897.*
5. *Nilsen DW, Albrektsen G, Landmark K, et al. Effects of a high-dose concentrate of n-3 fatty acids or corn oil introduced early after an acute myocardial infarction on serum triacylglycerol and HDL cholesterol. Am J Clin Nutr 2001; 74:50.*
6. *Tavazzi L, Maggioni AP, Marchioli R, et al. Effect of n-3 polyunsaturated fatty acids in patients with chronic heart failure (the GISSI-HF trial): A randomised, double-blind, placebo-controlled trial. Lancet 2008; 372:1223.*
7. *Kromhout D, Giltay EJ, Geleijnse JM, et al. n-3 fatty acids and cardiovascular events after myocardial infarction. N Engl J Med 2010; 363:2015.*

Reproduced from: Balk EM, Adam GP, Langberg V, et al. Omega-3 fatty acids and cardiovascular disease: An updated systematic review. Evidence Report/Technology Assessment No. 223. (Prepared by the Brown Evidence-based Practice Center under Contract No. 290-2012-00012-I.) AHRQ Publication No. 16-E002-EF. Rockville, MD: Agency for Healthcare Research and Quality; August 2016. https://effectivehealthcare.ahrq.gov/products/fatty-acids-cardiovascular-disease/research/.

Graphic 59845 Version 5.0

Random effects model



This model assumes that there are multiple treatment effects and that each trial provides an estimate of its own true effect.

RD: risk difference; OR: odds ratio; RR: relative risk.

Courtesy of Joseph Lau, MD.

Graphic 51482 Version 4.0

Contributor Disclosures

Ethan Balk, MD, MPH Consultant/Advisory Boards: Society for Gynecologic Surgeons [Systematic review of methodology consultant]; American Association of Gynecologic Laparoscopists [Systematic review of methodology consultant]; Kidney Diseases – Improving Global Outcomes (KDIGO) [Evidence review team, systematic review and clinical practice guideline development methodologist]. All of the relevant financial relationships listed have been mitigated. **Peter A L Bonis, MD** No relevant financial relationship(s) with ineligible companies to disclose. **Joann G Elmore, MD, MPH** No relevant financial relationship(s) with ineligible companies to disclose. **Carrie Armsby, MD, MPH** No relevant financial relationship(s) with ineligible companies to disclose.

Contributor disclosures are reviewed for conflicts of interest by the editorial group. When found, these are addressed by vetting through a multi-level review process, and through requirements for references to be provided to support the content. Appropriately referenced content is required of all authors and must conform to UpToDate standards of evidence.

Conflict of interest policy

→