Methods Guide
for Comparative Effectiveness Reviews

Expanded Guidance on Selected Quantitative Synthesis Topics





Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

This report is based on research conducted by the Tufts Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10055-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials that are clearly noted in the document. Further reproduction of those copyrighted materials is prohibited without the specific permission of copyright holders.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Lau J, Terrin N, Fu R. Expanded Guidance on Selected Quantitative Synthesis Topics. Methods Research Report. (Prepared by the Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-I.) AHRQ Publication No. 13-EHC024-EF. Rockville, MD. Agency for Healthcare Research and Quality; March 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Authors:

Joseph Lau, M.D.[a]

Norma Terrin, Ph.D.[a]

Rochelle Fu, Ph.D.[b]

[a]Tufts Evidence-based Practice Center
[b]Oregon Evidence-based Practice Center

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

Strong methodological approaches to systematic review improve the transparency, consistency, and scientific rigor of these reports. Through a collaborative effort of the Effective Health Care (EHC) Program, the Agency for Healthcare Research and Quality (AHRQ), the EHC Program Scientific Resource Center, and the AHRQ Evidence-based Practice Centers have developed a Methods Guide for Comparative Effectiveness Reviews. This Guide presents issues key to the development of Systematic Reviews and describes recommended approaches for addressing difficult, frequently encountered methodological issues.

The Methods Guide for Comparative Effectiveness Reviews is a living document, and will be updated as further empiric evidence develops and our understanding of better methods improves. We welcome comments on this Methods Guide paper. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Task Order Officer
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Expanded Guidance on Selected Quantitative Synthesis Topics

Abstract

This report provides expanded guidance on several topics that originally appeared in Chapter 9 ("Conducting Quantitative Synthesis When Comparing Medical Interventions") of the 2007 draft "Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews." Selected topics from this chapter were posted on the Effective Health Care Program Web site after public comments and were also published as a journal manuscript. The topics in the current report were cut from the 2007 draft methods reference guide to make the currently posted quantitative synthesis document a manageable length. The current report complements the posted document and includes the following topics: combining a small number of studies, combining composite outcome, control rate meta-regression, and interpretation and translation of results of meta-analyses.

The first three topics of this report focus on whether meta-analyses should be conducted in the settings encountered and on the selection of appropriate methods should it be decided to carry out meta-analyses. The section on combining small number of studies provides the rationale for why meta-analyses of small number (two to four) of studies could be unreliable and gives guidance on performing meta-analyses that have only few studies. The section on combining composite outcome discusses the rationale for using composite outcomes as well as the potential for misinterpretation of clinical trials when such outcomes are used and provides guidance on carrying out the proper analyses and interpretation. The section on control rate meta-regression discusses settings in which heterogeneous treatment effects may be related to varying baseline risk. The proper method of performing control rate meta-regression is discussed. Finally, the section on interpretation and translation of results of meta-analyses provides practical guidance on interpreting meta-analysis results of binary and continuous outcomes, as well as time to event and count data. This report ends with a section that provides instructions for reporting of meta-analyses.

Contents

Background 1
Combining a Small Number of Studies 2
Combining Composite Outcomes 4
Statistical Efficiency in Clinical Trials 4
Misinterpretation in Clinical Trials 4
Example 5
Composite Outcomes in Meta-Analysis 5
Summary 5
Control Rate Meta-Regressions 6
Interpretation and Translation of Results of Meta-Analyses 7
Binary Outcomes 7
Continuous Outcomes 9
Time to Event Data and Count Data 9
Instructions for Reporting the Quantitative Synthesis of Studies 12
References 15

Tables
Table 1. Number needed to treat with statins to prevent one cardiovascular event in 5 years. 11
Table 2. Summary of headings for reporting the quantitative synthesis of studies: methods section. 13
Table 3. Summary of headings for reporting the quantitative synthesis of studies: Results Section 14

Figures
Figure 1. Distributions of meta-analysis results (summary odds ratio and between-study variance) in subsets of 3 studies drawn from a set of 10. 2

Background

This report provides expanded guidance on several topics that originally appeared in Chapter 9 ("Conducting Quantitative Synthesis When Comparing Medical Interventions") of the 2007 draft "Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews."[1] Selected topics from this chapter were posted on the Effective Health Care Program Web site[2] after public comments and were also published as a journal manuscript.[3] The topics in the current report were cut from the 2007 draft methods reference guide to make the currently posted quantitative synthesis document a manageable length. The current report complements the posted document and includes the following topics: combining a small number of studies, combining composite outcome, control rate meta-regression, and interpretation and translation of results of meta-analyses.

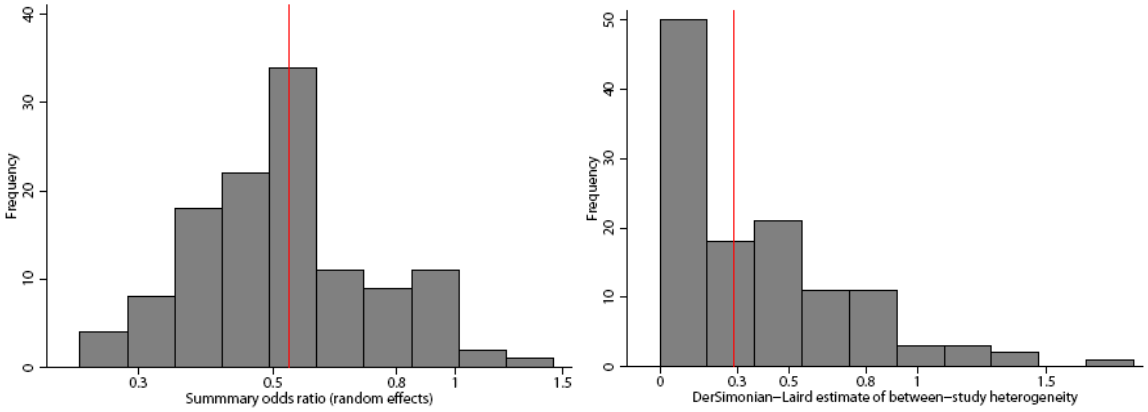
The first three topics of this report focus on whether meta-analyses should be conducted in the settings encountered and on the selection of appropriate methods should it be decided to carry out meta-analyses. The section on combining small number of studies provides the rationale why meta-analyses of small number (two to four) of studies could be unreliable and gives guidance on performing meta-analyses that have only few studies. The section on combining composite outcome discusses the rationale for using composite outcomes as well as the potential for misinterpretation of clinical trials when such outcomes are used and provides guidance on carrying out the proper analyses and interpretation. The section on control rate meta-regression discusses settings in which heterogeneous treatment effects may be related to varying baseline risk. The proper method of performing control rate meta-regression is discussed. Finally, the section on interpretation and translation of results of meta-analyses provides practical guidance on interpreting meta-analysis results of binary and continuous outcomes, as well as time to event and count data. This report ends with a section that provides instructions for reporting of meta-analyses.

Combining a Small Number of Studies

There is no general rule for deciding the minimum number of studies in a meta-analysis, and it is possible to combine results even if there are only two studies. When interpreting the results, the precision of the studies is as relevant as the number of studies. Thus the meta-analysis of three "mega-trials" will be more reliable than the meta-analysis of three small trials, all other factors being equal. Therefore, determining whether to include studies in a meta-analysis will depend on the extent of their clinical and methodological diversity.

As an example of the hazards of relying on too little information, consider a meta-analysis of palliative chemotherapy versus supportive care and/or delayed chemotherapy for the treatment of advanced or metastatic colorectal cancer.[4] The summary random effects odds ratio for death within 12 months is 0.53 (95% CI: 0.34, 0.83) in favor of palliative chemotherapy (10 studies). The studies are statistically heterogeneous, with significant Q statistic and I^2=60 percent. Suppose that only 3 of the 10 studies had actually been completed, while the others never made it past the planning stages. The results would be quite different depending on which 3 studies had been completed. Figure 1 displays the wide variation in the summary odds ratio and estimated between-study variability among all 120 possible subsets of size 3.

Figure 1. Distributions of meta-analysis results (summary odds ratio and between-study variance) in subsets of 3 studies drawn from a set of 10.



Note: Vertical lines indicate the estimates for the entire set of 10 studies.

Statistical heterogeneity is difficult to infer when the total amount of information (sum of the precisions in the individual studies) is low,[5] although the I^2 statistic can be used as a descriptor[6] to help with the determination of whether to combine. Meta-regression should generally be avoided when there are few studies, because of low power.

Although it is not feasible to determine whether there is statistical heterogeneity among a small number of studies, a random effects model is preferred when heterogeneity is suspected. The classical random effects models (e.g., DerSimonian and Laird[7]) assume that the between-study variance is known, when actually it is estimated from the data. Hence the methods tend to underestimate the error associated with parameter estimates, particularly when the number of studies is small. The larger the true between-study variance, the less accurate the DerSimonian and Laird confidence limits.[8] A Bayesian meta-analysis with a vague prior distribution on the between study variance is recommended when there are too few studies to accurately estimate the between-study variance.[9]

One should consider whether the rarity of eligible studies is an indication of publication bias or selective outcome reporting, or whether the intervention being studied is novel and the specific scientific field is relatively new and immature. It is not unusual for estimates based on a handful of early studies to shift considerably over time as more studies are published on the same topic.[10] Thus, the interpretation of results should take into account the number of years since the first publication, or meta-analysis should be deferred until more studies are available.

In summary, when few studies (i.e., two to four) are available for meta-analysis:

- Clinical and methodological similarity should be taken into consideration when determining whether to combine them.
- Statistical heterogeneity is more difficult to address, but can be handled better with Bayesian random effects models than with classical methods.
- Meta-regression should be avoided.
- Interpretation should take into account the precision of the individual studies as well as the number of studies.
- Maturity of the field of investigation also needs to be considered.

Combining Composite Outcomes

A composite outcome can be binary (0/1) or time-to-event. If it is binary, it takes the value 1 if any of several possible events occurs. For example, “cardiovascular event (yes/no)” could be defined as a composite of MI, stroke, and death from cardiovascular disease. If a composite outcome is time-to-event, it takes the value of the time until the first event. Although the use of a composite outcome as the primary outcome in a clinical trial can reduce sample size requirements, that approach may lead to serious misinterpretation of the data. In meta-analysis, sample size is less of a concern than in clinical trials, and thus the motivation for using composite outcomes is diminished.

Statistical Efficiency in Clinical Trials

Composite outcomes can improve statistical efficiency, increasing power for a given sample size. If power is projected to be insufficient to analyze each of several outcomes separately while maintaining a low overall Type I error rate, investigators may be interested in using a composite as the primary outcome. Furthermore, composites have a larger number of events than the component outcomes, and thus they can increase the power for time-to-event analyses and binary analyses of relative measures (odds ratio or relative risk). So composites can improve power both by handling the multiple testing problem and increasing the number of events. However, a composite can also reduce power by diluting component outcomes that are affected by the treatment with others that are not.[11] Composite outcomes may be considered in the context of clinical trials for which there are several relevant outcomes of similar clinical importance pertaining to the same disease process. In addition to homogeneity of clinical importance, there should be an expectation that the risk ratio of treatment benefit will be similar across the component outcomes.[11]

Misinterpretation in Clinical Trials

Composite outcomes pose a dilemma with regard to interpretation. For example, if an intervention results in a reduced risk for the composite of hospitalization and death, the intervention may have decreased hospitalizations while having no effect or a negative effect on survival. In reporting the result, it would be hard to avoid the suggestion of a reduction in mortality even if there was none. Requirements for meaningful composites include homogeneity of clinical importance as well as homogeneity of treatment benefit across the component outcomes.[12,13] Furthermore, the statement of the result should make clear the extent to which the component outcomes contributed to the finding. Empirical research found that in most clinical trials with composite endpoints, there was heterogeneity of clinical importance of the component outcomes; in about a third of the trials, results for the components were not reported; and in those trials that did report results for the components, more than half had heterogeneity of treatment effect.[14] Another review of composite outcomes found that only 60 percent of trials provided reliable estimates for both the composite and its components. The components were judged to be of similar importance in only 18 percent of trials. Indeed, death was the most important component in 83 percent of trials. Other problems included post hoc and inconsistent definitions of the composite.[15]

Example

In a trial that randomized 120 patients with in-stent stenosis of a saphenous vein graft to radiation or placebo, the composite outcome of death from cardiac causes, Q wave myocardial infarction, and revascularization of the target vessel, there were 43 events in the placebo arm compared with 22 for the intervention. Death and MI together accounted for 6 events in the placebo arm and 5 in the intervention arm. Thus, despite the composite outcome definition's inclusion of death and MI, the trial provided little information on these outcomes.[14,16]

Composite Outcomes in Meta-Analysis

Because of the large number of patients contributing data, systematic review diminishes and may eliminate the primary motivation for analyzing composites; that is, increasing statistical power. Furthermore, meta-analyses of the individual components of the composite yield more meaningful results. When a meta-analysis of a composite outcome is undertaken, trials without data for all component outcomes should be graded as having high risk of bias. Only composite outcomes that are generally agreed upon and in wide usage by the research community should be used in meta-analysis, and the meta-analyses of individual components should also be performed. Creating de novo composite outcomes without a precedent by the meta-analysts should be avoided. Statistical and clinical homogeneity of the components should be verified.

Summary

- Composite outcomes typically increase statistical efficiency.
- The additional power may not be necessary for many meta-analyses.
- Interpretation of composite outcome results is fraught.
- Only widely accepted composite outcomes should be used in meta-analysis.
- The components of the composite should be homogeneous with respect to clinical importance and magnitude of treatment benefit.
- For most composites used in clinical trials, there is heterogeneity of clinical importance across the components.
- Meta-analyses of the individual components should also be performed.

Control Rate Meta-Regressions

Patients with higher underlying risk for mortality and other outcomes may experience different benefits or harms from treatment than patients with lower underlying risk.[17] For studies with binary outcomes, the "control rate" refers to the proportion of subjects in the control group who experience the event. The control rate may be affected by disease severity, concomitant treatments, followup duration, as well as other factors that differ across studies,[18,19] and may thus be viewed as a study-level proxy for these factors. It is used to test for interaction between underlying population risk and treatment benefit, via control-rate meta-regression. However, advanced methods must be employed to obtain the correct level of statistical significance.

Even in the absence of a true linear relationship between treatment effect and control rate, the expected slope for the regression of treatment effect on control rate is non-zero. This bias is caused by measurement error in the control rate estimate and correlation between the control rate and treatment effect estimates.[20,21] Simple weighted regressions tend to identify a significant relation between control rate and treatment effect twice as often as more suitable approaches including hierarchical meta-regression models[19] and Bayesian meta-regressions.[21]

Thompson, Smith, and Sharp[21] illustrated the hazards of using a naïve meta-regression model to assess the relation between the control rate and mortality in a meta-analysis of the effectiveness of endoscopic sclerotherapy in patients with cirrhosis and esophagogastric varices.[22] The naïve approach estimated a statistically significant negative slope for the regression of odds ratio on control rate, implying that the higher the underlying risk, the more effective the treatment. In contrast, a Bayesian analysis that accounted for all sources of variability and correlation found a much weaker relation.[21]

The presence of a control rate effect varies according to the metric. The risk difference is more highly correlated with the control rate than is the relative risk or odds ratio and is constrained by the control rate particularly when the control rate is small. Schmid et al. demonstrated this empirically and showed that the relationship with the control rate is inflated using the risk difference metric.[19] In an empirical evaluation control rate effects were seen in 14 percent, 13 percent or 31 percent of 115 meta-analyses of binary outcomes when the measure of choice was the odds ratio, the risk ratio, or the risk difference, respectively.[19] The differences in the percentages between the relative (odds ratio, risk ratio) and the absolute (risk difference) metrics is related to the greater heterogeneity of the risk difference. For example, a risk ratio of 1.5 corresponds to very different risk differences at various levels of baseline risk (0.5 percent at 1 percent control rate, and 5 percent at 10 percent control rate).

A scatter plot of treatment effect against control rate is a useful ad hoc approach to visually assess whether there may be a relation between the two. A quick way to rule out the presence of a control rate effect is by a weighted regression of the effect size on the control rate. A negative finding would be most likely replicated by the more complicated methods; a positive finding would need to be verified by a more comprehensive method.

In summary, if the control group event rate is a plausible proxy for average within-study severity of illness of the study population, then:

- Consider a control rate meta-regression to explain between-study treatment effect heterogeneity.
- The use of a relative metric (risk ratio, odds ratio) is preferred in control rate meta-regression.
- Use a scatter plot to search for a systematic change in the effect size at different control rates.

- Use a simple weighted regression of the effect size on the control rate to rule out presence of a control rate effect; if the slope is significantly different than 0, advanced methods must be used to obtain the correct level of statistical significance.

Interpretation and Translation of Results of Meta-Analyses

CERs should present summary effects in a way that makes it easy for readers to interpret and apply these findings appropriately. This section discusses different ways of presenting and interpreting various effect measures.

Binary Outcomes

Three effect measures could be used for binary outcomes in meta-analyses including risk difference (RD), relative risks (RR) and odds ratios (OR). It should be noted that there is no single perfect metric that is adequate in all settings. Each has its limitations and the proper interpretation requires additional data in order to fully inform the decision maker.

RD is generally considered as being most easily understood by clinicians and patients, and is the absolute difference in probabilities of an event between two intervention groups. Interpretation of RD is straightforward. For example, a RD of 5 percent between the intervention and placebo groups indicates that the risk of an event in the intervention group is 5 percent higher than the risk in the placebo group. Investigators should note that the clinical relevance of RD (as well as for RR and OR) depends on the underlying event rates. A RD of 2 percent could be clinically significant if the change is from 3 percent to 1 percent of an event, and less significant if the intervention reduces the risk of an event from 78 percent to 76 percent. Therefore, when reporting a RD, the underlying event risks from each study should be reported as well, and investigators should comment on the clinical significance of the RD. Furthermore, the proportion of event for each intervention group usually increases with the increase of study duration and the estimated RD may increase accordingly. While it is not recommended to combine studies using RD when baseline risks are different among studies, when it is appropriate to combine RD, investigators should be clear about the length of followup periods of included studies. For example, for a group of studies with about 3 months' followup, the risk of an event in the intervention group in an average of 3 months is 5 percent higher than the risk in the placebo group.

RR (and OR) provide estimates that are less likely to vary over different populations and study durations, compared with RD. RR is interpreted as the ratio of probabilities of an event between two intervention groups. Therefore, a RR of 2 means a twofold increased risk of an event in patients receiving a treatment compared with those not receiving the treatment. For example, in a study examining the adherence to prescribed inhalers for patients with chronic obstructive pulmonary disease, patients on tiotropium were twice as compliant as patients using ipratropium (RR: 2.0; 95% CI, 1.8–2.3).[23] Likewise, a meta-analysis of crystalline silica, subjects exposed to crystalline silica were shown to have a twofold incidence of lung cancer compared with those not exposed to crystalline silica (RR: 2.0, 95% CI, 1.8–2.3).[24]

Alternatively, investigators could present results as a relative risk reduction or relative risk increase, especially when the RR is below 2. For example, a CER on second-generation antidepressants compared discontinuation due to adverse events between venlafaxine and the class of selective serotonin reuptake inhibitors (SSRIs), and the combined RR was 1.36 (95% CI, 1.09–1.69).[25] This finding could be expressed as a relative risk increase, that is, venlafaxine had a 36 percent higher risk of causing discontinuation due to adverse events than SSRIs as a class.

Similarly, if a combined RR is 0.74 to compare an intervention to the placebo, the finding could be interpreted as that the risk of the intervention is 36 percent less. However, investigators must be aware that the meaning of RR is not symmetric around 1. For example, the RR of 0.5 of dying is not the same as RR of 2 of not dying (living); whereas the OR calculation is valid.

Although ORs have mathematical advantages over RRs, they are more difficult to interpret because they describe the ratio of the odds of an event among those exposed to an intervention to the odds among those not exposed, and odds is not intuitive to communicate the magnitude of risk. Mathematically one could choose either the RR or OR metric in the analyses of data and their results would be similar when the event rates are low. Investigators should avoid the common misinterpretation of treating odds and odds ratios as risks and relative risks, especially when event risks are high (> 10%). This misinterpretation could lead to an overstatement of the actual effect size. For example, a survey designed to examine physician diagnostic practices for patients with chest pain noted a statistically higher rate of cardiac catheterizations for men than for women (OR 1.7, 95% CI, 1.1–2.5),[26] causing concerns in the media about gender disparities. Schwartz et al. reanalyzed the same data using RRs and found that the gender disparities is actually small (RR 1.07 95% CI 1.01–1.16).[27]

To facilitate interpretation when RR or OR is used, we recommend calculating a RD or number need to treat (NNT) or number needed to harm (NNH) and the corresponding 95% confidence interval using the combined estimates at typical proportions of events in the control group, to provide enough information for readers to assess the clinical relevance. For the above comparison between venlafaxine and SSRIs, given that a typical proportion of discontinuation is 8 percent for the SSRI group, the corresponding NNH to prevent one additional discontinuation is 35 (95% CI, 18–139).

NNTs and NNHs are frequently used because they portray the absolute effect of an intervention that is believed to be intuitive.[28] NNTs and NNHs themselves do not reflect variations attributable to underlying event rates; and they do not have a standardized unit of time. Therefore, when NNTs or NNHs are presented, investigators should report these measures with an appropriate time frame and make clear that they are based on an average estimate. For example, one correct interpretation of a NNT of 10 over 3 years could be that “On average, 10 patients would have to be treated for 3 years with treatment A to observe one fewer event after 3 years”.[29] A different and less used way to interpret a NNT (or NNH) would be as a treatment frequency. For example, a NNT of 100 could be presented as 10 in 1,000 treated people will benefit from treatment. If substantial variations in NNTs (NNHs) exist based on different event rates, dosages, or subgroups, then investigators should report them separately for each group. However, the use of NNT and NNH is not universally recommended.[30] Empirical studies have questioned whether this metric is really intuitive to patients.[31]

Finally, the terms “risk difference” or “relative risk” themselves can be confusing however if they refer to a beneficial outcome. Investigators should avoid the use of “risk” when reporting beneficial outcomes. Instead, investigators could interpret the results in terms of the probability of the beneficial outcome directly. For example, if a meta-analysis produced a RD of 10 percent when combining studies comparing the effectiveness of a drug vs. placebo to achieve a 50 percent pain reduction, it could reported as that comparing with the placebo group, the probability of achieving a 50 percent pain reduction was 10 percent higher in the treatment group. When RR is used, substituting “relative risk” with “relative benefit” may help readers avoid confusion with contradicting terminology. For example, the term "relative benefit" was used in a systematic review on the efficacy and safety of second-generation antidepressant to

describe the beneficial response to treatment.[32] For the outcome of being a responder, the result was reported as "suggested a modest additional treatment effect (relative benefit, 1.10 [95% CI, 1.01–1.22]) for sertraline compared with fluoxetine."[32]

Continuous Outcomes

The weighted mean difference (WMD) and the standardized mean difference (SMD) or effect size can be used for meta-analyses of continuous data. WMD can be used when outcome measurements in all trials are assessed on the same scale, and easily interpreted as the mean difference between two comparison groups. The summary effect has the same unit as the scale employed in the included studies. For example, in a meta-analysis of differences in points on the Montgomery-Asberg Depression Scale (MADRS) between escitalopram and citalopram,[25] the WMD was estimated to be 1.51 (95% CI, 0.58–2.45). This finding can be interpreted as escitalopram having an additional treatment effect of 1.51 points on the MADRS, or escitalopram having a 1.51 higher points on the MADRS. Although this finding was statistically significant, the clinical significance of a difference of 1.51 points must be determined independently.

Standardized mean difference or effect size meta-analyses can be used if the same outcome was assessed on different measurement scales. Results, however, are expressed in units of standard deviations, rather than in units of any measurement scales and can be difficult to interpret. For example, Hansen et al.[33] combined functional outcomes measured on different scales in placebo-controlled studies of Alzheimer's drugs using standardized mean different and the combined estimate was 0.25 (95% CI 0.13, 0.37) for trials less than 24 weeks, and 0.29 (95% CI 0.22, 0.36) for trials more than 24 weeks. Although these results were interpreted as small based on the most widely used classification, where standardized effect sizes of 0.2, 0.5 and 0.8 are suggested corresponding to small, medium, and large referents,[34] the clinical significance of the additional treatment effect of Alzheimer's drugs compared with placebo is difficult to determine. This is an inherent problem of using standardized mean difference where currently there is no better interpretation available. To facilitate interpretation, the investigators could consider calculating an approximation of mean difference on the included measurement scales by multiplying the standardized effect sizes by the combined standard deviation for each included scale.

Time to Event Data and Count Data

Hazard ratio (HR) is the measure typically used for time to event data. Interpretation of HR is similar to RR, so a HR of 2 could also be interpreted as a twofold increased risk of an event in patients receiving a treatment compared with those not receiving the treatment. However, there is a subtle difference between HR and RR where RR is a ratio of two probabilities and HR is a ratio of two hazard rates (instantaneous risk). Such distinction is not important for the clinical implication of the results and informing patients, clinicians and health policy makers. Rate ratio (RR) is the measure typically used for count data, and as the term indicates, the ratio of two rates. For a rate ratio of 2, it means the rate of an event in patients receiving a treatment is 2 times the rate of an event in patients not receiving the treatment. Similar to binary outcome, we recommend reporting both the event rate for each treatment arm and the rate ratio. The estimate of rate takes into account both the number of new cases, and followup time of population. Its interpretation depends on the selection of the time unit. For example, a rate of 0.097/person-years could be expressed as 0.008/person-months, or 97/1000 person-years. It is essential in

presenting incidence rates with appropriate time units. For clarity, the numerator is often expressed as a power of 10.

Similar to relative risk, investigators should calculate NNT/NNH based on combined hazard ratio or rate ratio while incorporating the time frame associated with such calculations. Smeeth et al.[30] provided a good example, and calculated NNT with statins to prevent one cardiovascular event and mortality over 5 years. Although they combined studies to achieve a summary NNT, they also presented NNTs for individual studies with varying baseline risks (Table 1). The combined NNT to prevent one death was 20 over 5 years. NNTs of individual studies, however, ranged from 8 to 28 corresponding to different baseline risks. A 95% CI was provided for each NNT from the combined estimates, and we recommend providing a 95% CI for all estimates. A similar table could also be used for reporting relative risk and NNT from binary data with minor modifications. For example, for the column of baseline risk, it could be replaced with control rate (proportion of event in the control group) if the event rate is not available. If it is appropriate to use risk difference to combine data, the columns of rate ratio could be replaced by risk difference to report the results.

Table 1. Number needed to treat with statins to prevent one cardiovascular event in 5 years

Trials	Number of Subjects	Baseline Risk of CHD Mortality per 100 Person-Years	Rate Ratios			Number Needed To Treat (5 years)		
			Total Mortality	CHD Mortality	All CV Events	Total Mortality	CHD Mortality	All CV Events
Primary Prevention								
AFCAPS/TexCAPS	6,605	0.1	1.04	1.36	0.69	167*	1,000*	28
WOSCOPS	6,595	0.4	0.78	0.67	0.7	118	182	28
Secondary Prevention								
Scandinavian simvastatin survival study trial	4,444	1.6	0.71	0.59	0.64	33	31	8
CARE	4,159	1.2	0.92	0.81	0.75	133	95	11
Long-term intervention with pravastatin in ischemic disease	9,014	1.4	0.78	0.77	0.8	41	64	17
Combined Effects (95% CI)			0.80 (0.74 to 0.87)	0.73 (0.66 to 0.81)	0.74 (0.71 to 0.77)	113 (77 to 285)	500 (222 to -)**	20 (17 to 25)

Adapted by permission from BMJ Publishing Group Limited. BMJ. Smeeth L, Haines A, Ebrahim S, vol. 318, pp. 1548-51, 1999.
CHD = coronary heart disease; CV = cardiovascular
* AFCAPS/TexCAPS study reported a nonsignificant increased total and CHD mortality in the intervention group. Numbers needed to treat are derived from the lower limit of the 95% CIs of the risk differences in event rates to illustrate the lower limit within which the numbers might lie.
**No upper number needed to treat can be calculated as the upper 95% CI of pooled absolute risk difference is greater than zero. In these circumstances, the number needed to treat is a number needed to harm.

Key Points

- Investigators should present summary effects in a way that makes it easy for readers to interpret and apply these findings appropriately.
- Investigators should interpret results accordingly based on the type of measure and data.
- For binary outcomes, report underlying event rates along with the effect measure used in the meta-analysis.
- For binary outcomes, consider calculating number need to treat (NNT) or number needed to harm (NNH) and the corresponding 95% confidence interval to provide information for readers to assess the clinical relevance.
- For binary outcome, NNTs and NNHs should be interpreted as "on average" within a specific time frame. If NNTs (NNHs) differ substantially based on control event rates, dosages, or subgroups, they should be presented separately. A confidence interval should be presented for each NNT or NNH.
- If ORs are used in a meta-analysis, results should be interpreted in terms of odds. Only when the event rate is low (< 10%), the OR can be interpreted approximately in the same way as RRs.
- If meta-analysis using standardized effect sizes is performed, standard deviations could be used to convert standardized effect sizes back to a unit on a specific scale to facilitate the interpretation.
- For time to event data and count data, investigators should also calculate NNT/NNH while incorporating the timeframe associated with such calculations.

Instructions for Reporting the Quantitative Synthesis of Studies

The purpose of the following summary of headings (Tables 2 and 3) for reporting quantitative syntheses of studies is to ensure some degree of uniformity in how EPCs present CER methods and results. The summary is not entirely prescriptive because CERs do not have to include all headings at all times. Rather, if a review touches upon an area encompassed by a heading or subheading, then the heading or subheading should be included in the review.

Reporting of elements pertaining to the heading or subheading should be done in accordance with the explanations provided in the "required reporting" column of the table below. For additional information, the section of the quantitative chapter that discusses the pertinent issues is identified.

For example, if the authors decide to conduct a meta-analysis, then they will have to include a heading in the methods section of their report that pertains to "method of combining studies." Under this heading, they will have to describe and justify the statistical procedure used to combine effect measures from individual studies. In the results, a graphical summary of individual and combined study effect estimates will have to be provided in accordance with the recommendations enumerated below.

If the review does not touch upon a specific area, then no mention of the associated heading or subheading is necessary. If no meta-analysis is conducted, then the authors would not have to

include a heading about methods of combining studies. The exact titles of headings and subheadings are left to the discretion of authors.

Table 2. Summary of headings for reporting the quantitative synthesis of studies: methods section

Headings	Subheadings	Required Reporting
Rationale to combine	Clinical heterogeneity	Specify important clinical characteristics which may differ among studies (e.g., intervention, dosage, baseline disease severity, length of followup) and how they will affect the decision to combine. Define the threshold for acceptable differences in clinical characteristics which could be combined in a meta-analysis based on the scope of the research question. For example, for length of followup, define the range of lengths of included studies that could be combined in one meta-analysis.
	Methodological heterogeneity	Specify important methodological characteristics which may differ among studies (e.g., mechanism of randomization, extent and handling of withdrawals and losses to follow up) and how they will affect the decision to combine. Define the threshold acceptable differences in methodological characteristics which could be combined in a meta-analysis based on the scope of the research question.
Criteria for selecting outcomes for combining	Outcome definitions	Specify whether outcome definitions or the way outcomes were measured differed among studies. Specify whether surrogate outcomes or composite endpoints were used. If observational studies are included, describe the definition and measurement of confounding factors/effect modifiers considered in the analyses of individual studies.
	Primary vs. secondary outcomes	Specify whether outcomes were primary or secondary outcomes in the original studies. Specify benefit and harm outcomes clearly.
	Outcome assessment in RCTs	Specify whether ITT, per protocol, last observation carried forward, etc. was used to handle outcomes in each study. If estimates from different outcome definitions were combined, then subgroup and/or sensitivity analyses should also be undertaken.
Types of studies included	Study design	Specify what type of study designs are being combined (e.g., RCT [crossover, cluster randomized, factorial], observational [cohort, case-control, cross-sectional]).
	Rationale for inclusion of observational studies	If observational studies are included, then provide a rationale (e.g., to broaden generalizability, to examine longer followup periods, inadequate data from RCTs, etc.).
Explanation of choice of effect measure		Specify what type of outcome data is being combined (e.g., dichotomous, continuous, ordinal, counts, time to event) and the measure(s) of effect chosen (e.g., RR, OR, RD, HR, mean difference, standardized mean difference). This should be done for each outcome considered. If the study design allows a choice of effect measure then choose the one that best answers the research question and provide a rationale for that choice.

Table 2. Summary of headings for reporting the quantitative synthesis of studies: methods section (continued)

Headings	Subheadings	Required Reporting
Methods for combining study estimates	Statistical procedure and justification of model chosen	First specify whether direct or indirect comparisons are being made.
	Direct comparison	For direct comparison, describe and justify the statistical model used to combine effect measures (e.g., random effects model, fixed effects model, Bayesian model).
	Indirect comparison	If indirect comparisons are being made, clearly state the rationale. Describe the methods used for indirect comparison and specify the analyses done to ensure the validity and robustness of results from indirect comparison.
	Special considerations	For rare binary outcome, describe and justify the statistical methods used.
Statistical heterogeneity	Statistical tests	Specify how statistical heterogeneity is assessed and the criteria used to identify "important" heterogeneity.
	Quantifying heterogeneity	Specify methods used to quantify statistical heterogeneity (e.g., I^2).
	Exploring heterogeneity	Specify the methods used to explore important clinical, methodological, or statistical heterogeneity (e.g., meta-regression, control rate meta-regression, subgroup analysis). Distinguish between prespecified and post hoc analysis. The exploratory nature of these analyses should be clear.
Sensitivity analyses		Specify what sensitivity analyses are being done and how they relate to key decisions and assumptions made in the systematic review.

Table 3. Summary of headings for reporting the quantitative synthesis of studies: results section

Headings	Recommendations
Descriptive study information	Include information for each study describing the sample size, intervention, outcome, study design, target population, study population, baseline risk and other important PICOS study characteristics that are related to clinical, methodological or statistical heterogeneity. Sponsorship of the studies and reported conflict of interest should also be reported.
Level of evidence and quality of the studies	Specify the level of evidence given feasibility of different designs to investigate the research question. Specify the scale to estimate the quality of the study and how internal and external validity of the studies are assessed.
Graphical summary of individual and combined study estimates	For each outcome present tables or a graphical representation of the data (forest plot) including: The comparison type, sample size for each study, weight given to each study (or represented by the size of plot symbol), measure of effect and confidence interval for each study, and a summary measure of effect and confidence interval for all studies combined. A p-value for a test and quantification of statistical heterogeneity should be included in the figure or in the figure legend. If study results are not quantitatively combined, a forest plot without a summary estimate can still be provided.
Reporting of individual and combined study estimates	Provide interpretation for the individual and combined study estimates based on the type of data and choice of effect measure. Provide interpretation for results from test and exploration of heterogeneity. If additional analyses were conducted (e.g., sensitivity analysis), report the results of all additional analyses undertaken.

References

1. Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews, Version 1.0 [Draft posted Oct. 2007]. Rockville, MD; Agency for Healthcare Research and Quality. http://effectivehealthcare.ahrq.gov/repFiles/2007_10DraftMethodsGuide.pdf. Accessed September 23, 2012.
2. Fu R, Gartlehner G, Grant M, et al. Conducting Quantitative Synthesis When Comparing Medical Interventions: AHRQ and the Effective Health Care Program. In: Methods Guide for Comparative Effectiveness Reviews [posted October 2010]. AHRQ Publication No. 10(12)-EHC063-EF. Rockville, MD; Agency for Healthcare Research and Quality. Chapters available at: http://effectivehealthcare.ahrq.gov/.
3. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011; 64:1187-97.
4. Palliative chemotherapy for advanced or metastatic colorectal cancer. Colorectal Meta-analysis Collaboration. Cochrane Database Sys Rev. 2000;(2):CD001545.
5. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med. 1998;17:841-56.
6. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21:1539-58.
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177-88.
8. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20:825-40.
9. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. Stat Meth Med Res. 2001;10:277-303.
10. Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. J Clin Epidemiol. 2004;57:1124-30.
11. Freemantle N, Calvert M, Wood J, et al. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003; 289:2554-9.
12. Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I, et al. Validity of composite end points in clinical trials. BMJ. 2005; 330:594-6.
13. Pogue J, Thabane L, Devereaux PJ, et al. Testing for heterogeneity among the components of a binary composite outcome in a clinical trial. BMC Medical Research Methodology. 2010;10:49.
14. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ. 2007; 334:786.
15. Cordoba G, Schwartz L, Woloshin S, et al. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ. 2010; 341:c3920.
16. Waksman R, Ajani AE, White RL, et al. Intravascular gamma radiation for in-stent restenosis in saphenous-vein bypass grafts. N Eng J Med. 2002; 346:1194-9.
17. Glasziou PP, Irwig LM. An evidence based approach to individualizing treatment. BMJ. 1995;311:1356-9.
18. Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. Lancet. 1998;351:123-7.
19. Schmid CH, Lau J, McIntosh M, et al. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med. 1998;17:1923-42.
20. McIntosh M. The population risk as an exploratory variable in research synthesis of clinical trials. Stat Med 1997;15:1713-28.
21. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med. 1997;16:2741-58.

22. Pagliaro L, D'Amico G, Sorenson TIA, et al. Prevention of first bleeding in cirrhosis a meta-analysis of randomized trials of nonsurgical treatment. Ann Intern Med. 1992;117, 59:70.

23. Breekveldt-Postma NS, Koerselman J, Erkens JA, et al. Enhanced persistence with tiotropium compared with other respiratory drugs in COPD. Resp Med. 2007;101:1398-405.

24. Smith AH. Lopipero PA. Barroga VR. Meta-analysis of studies of lung cancer among silicotics. Epidemiol. 1995;6:617-24.

25. Gartlehner G, Gaynes BN, Hansen RA, et al. Comparative benefits and harms of second-generation antidepressants: background paper for the American College of Physicians. Ann Intern Med. 2008;149:734-50.

26. Schulman KA, Berlin JA, Harless W, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. N Engl J Med. 1999; 340:618-26.

27. Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effect of race and sex on physicians' referrals for cardiac catheterization. N Engl J Med. 1999; 341:279-83.

28. Laupacis A, Sackett DL, Roberts RS, An assessment of clinically useful measures of the consequences of treatment. N Engl J Med. 1988; 318:1728-33.

29. Hutton JL. Number needed to treat: properties and problems. J Royal Statist Soc A. 2000;163:403-19.

30. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading. BMJ. 1999;318:1548-51.

31. Sheridan SL, Pignone MP, Lewis CL. A randomized comparison of patients understanding of number needed to treat and other common risk reduction formats. J Gen Intern Med. 2003;18:884-92.

32. Hansen RA, Gartlehner G, Lohr KN, et al. Efficacy and safety of second-generation antidepressants in the treatment of major depressive disorder. Ann Intern Med. 2005;143:415-26.

33. Hansen RA, Gartlehner G, Lohr KN, et al. Functional outcomes of drug treatment in Alzheimer's disease: a systematic review and meta-analysis. Drugs Aging. 2007;24:155-67.

34. Cohen J. Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.