Methods Guide for
Comparative Effectiveness Reviews

Conducting Quantitative Synthesis When Comparing Medical Interventions: AHRQ and the Effective Health Care Program



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

Comparative Effectiveness Reviews are systematic reviews of existing research on the effectiveness, comparative effectiveness, and harms of different health care interventions. They provide syntheses of relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. Strong methodologic 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 Comparative Effectiveness 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. Comments and suggestions on the *Methods Guide for Comparative Effectiveness Reviews* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

This document was written with support from the Effective Health Care Program at AHRQ.

None of the authors has a financial interest in any of the products discussed in this document.

Suggested citation: Fu R, Gartlehner G, Grant M, et al. Conducting Quantitative Synthesis When Comparing Medical Interventions: AHRQ and the Effective Health Care Program. In: Agency for Healthcare Research and Quality. Methods Guide for Comparative Effectiveness Reviews [posted October 2010]. Rockville, MD. Available at: http://effectivehealthcare.ahrq.gov/.

Conducting Quantitative Synthesis When Comparing Medical Interventions: AHRQ and the Effective Health Care Program

Authors:

Rongwei Fu[a*]
Gerald Gartlehner[b]
Mark Grant[c]
Tatyana Shamliyan[d]
Art Sedrakyan[e]
Timothy J. Wilt[f]
Lauren Griffith[g]
Mark Oremus[g]
Parminder Raina[g]
Afisi Ismaila[g]
Pasqualina Santaguida[g]
Joseph Lau[h]
Thomas A. Trikalinos[h]

[a]Oregon Evidence-based Practice Center, Department of Public Health and Preventive Medicine, Oregon Health & Science University, Portland, OR
[b]Danube University, Krems, Austria
[c]Technology Evaluation Center, Blue Cross Blue Shield Association
[d]Minnesota Evidence-based Practice Center, Division of Health Policy and Management, University of Minnesota, Minneapolis, MN
[e]Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD
[f]Minnesota Evidence-based Practice Center, Minneapolis VA Center for Chronic Disease Outcomes Research and the University of Minnesota Department of Medicine, Minneapolis, MN
[g]McMaster Evidence-based Practice Center, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada
[h]Tufts Evidence-based Practice Center and Center for Clinical Evidence Synthesis, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA

The views expressed in this paper are those of the authors and do not represent the official policies of the Agency for Healthcare Research and Quality, the Department of Health and Human Services, the Department of Veterans Affairs, the Veterans Health Administration, or the Health Services Research and Development Service.

Conducting Quantitative Synthesis When Comparing Medical Interventions: AHRQ and the Effective Health Care Program

Abstract

Objective

The objective is to establish recommendations for conducting quantitative synthesis or meta-analysis using study-level data in Comparative effectiveness reviews (CERs) for the Evidence-based Practice Center (EPC) program of the Agency for Healthcare Research and Quality (AHRQ).

Study Design and Setting

We focused on recurrent issues in the EPC program and the recommendations were developed using group discussion and consensus based on current knowledge in the literature.

Results

We first discussed considerations for deciding whether to combine studies, followed by discussions on indirect comparison and incorporation of indirect evidence. Then we described our recommendations on choosing effect measures and statistical models, giving special attention to combining studies with rare events, and on testing and exploring heterogeneity. Finally, we briefly present recommendations on combining studies of mixed design and on sensitivity analysis.

Conclusion

Quantitative synthesis should be conducted in a transparent and consistent way. Inclusion of multiple alternative interventions in CERs increases the complexity of quantitative synthesis while the basic issues in quantitative synthesis remain crucial considerations in quantitative synthesis for a CER. We will cover more issues in future versions and update and improve recommendations with the accumulation of new research to advance the goal for transparency and consistency.

Introduction

Comparative effectiveness reviews (CERs) are systematic reviews that summarize comparative effectiveness and harms of alternative clinical options, and aim to help clinicians, policy makers, and patients make informed treatment choices. Quantitative synthesis, or meta-analysis, is often essential for CERs to provide scientifically rigorous summary information. Quantitative synthesis should be conducted in a transparent and consistent way, and methodologies reported explicitly. Reasons for this were made clear during the controversy around the safety of rosiglitazone, where a systematic review that found increased risk for myocardial infarction[1] spurred heated debate on issues around choosing appropriate methods for quantitative syntheses;[2-4] and the subsequent Congressional hearing[5] brought these issues further into spotlight. This story highlighted the fact that basic issues in quantitative syntheses, such as choice of an effect measure or a model or how to handle heterogeneity, remain crucial considerations and are often the subject of controversy and debate.

A CER typically evaluates the evidence on multiple alternative interventions whereas most published meta-analyses compared one intervention with a placebo. Inclusion of multiple interventions increases the complexity of quantitative synthesis and entails methods of comparing multiple interventions simultaneously. Evaluation of multiple interventions also makes the assessment of similarity among studies and the decision to combine studies even more challenging. Presenting results of a meta-analysis from a CER in a way that is useful to decisionmakers is also a challenge.

The Evidence-based Practice Center (EPC) program of the Agency for Healthcare Research and Quality (AHRQ)[6] is the leading U.S. program providing unbiased and independent CERs. The goal of this article is to summarize our recommendations in conducting quantitative synthesis of CERs for therapeutic benefits and harms for the EPC program with the goal to improve consistency and transparency. The recommendations cover recurrent issues in the EPC program and we focus on methods for combining study-level effect measures. First, we discuss considerations for deciding whether to combine studies, followed by discussions on indirect comparison and incorporation of indirect evidence. Then we describe our recommendations for choosing effect measures and statistical models, giving special attention to combining studies with rare events; and on testing and exploring heterogeneity. Finally, we briefly present recommendations on combining studies of mixed design and on sensitivity analysis. This article is not a comprehensive review of methods.

The recommendations were developed using group discussion and consensus based on current knowledge in the literature.[7] EPC investigators are encouraged to follow these recommendations but may choose to use alternative methods if deemed appropriate. If alternative methods are used, the investigators are required to provide rationales for their choice, and if appropriate, to state the strengths and limitations of the chosen method in order to promote consistency and transparency. In addition, several steps in conducting a meta-analysis require subjective decisions, for example, the decision to combine studies or the decision to incorporate indirect evidence. For each subjective decision, investigators should fully explain how the decision was reached.

Decision To Combine Studies

The decision to combine studies to produce an overall estimate should depend on whether a meaningful answer to a well formulated research question can be obtained. The purpose of a meta-analysis should be explicitly stated in the methods section of the CER. The overall purpose of the

review is not in itself a justification for conducting a meta-analysis, nor is the existence of a group of studies that address the same treatments. Investigators should avoid statements such as "We conducted a meta-analysis to obtain a combined estimate of …" Rather, explain the reason a combined estimate might be useful to decision makers who might use the report or products derived from the report.

Study Similarity Is a Requirement for Quantitative Synthesis

Combining studies should only be considered if they are clinically and methodologically similar. There is no commonly accepted standard defining which studies are "similar enough." Instead, the similarity of selected studies is always interpreted in the context of the research question, and to some extent, is subjective. In addition, judging similarity among studies depends on the scope of the research question. A general question may allow inclusion of a broader selection of studies than a focused question. For example, it may be appropriate to combine studies from a class of drugs instead of limiting only to a particular drug, if the effect of the drug class is of interest, and the included studies are methodologically comparable.

Statistical Heterogeneity Does Not Dictate Whether or Not To Combine

Variation among studies can be described as:[8]

1. **Clinical diversity:** variability in study population characteristics, interventions and outcome ascertainments.
2. **Methodological diversity:** variability in study design, conduct and quality, such as blinding and concealment of allocation.
3. **Statistical heterogeneity:** variability in observed treatment effects across studies. Clinical and/or methodological diversity, biases or even chance, can cause statistical heterogeneity.

Investigators should base decisions about combining studies on thorough investigations of clinical and methodological diversity as well as variation in effect size. Both the direction and magnitude of effect estimates should be considered. These decisions require clinical insights as well as statistical expertise.

Clinical and methodological diversity among studies always exists even if a group of studies meet all inclusion criteria and seem to evaluate the same interventions in similar settings. Incomplete description of protocols, populations, and outcomes can make it impossible to assess clinical and methodological diversity among trials; nor does it always result in detectable statistical heterogeneity.[9] Further, evolving disease biology, evolving diagnostic criteria or interventions, change in standard care, time-dependent care, difference in baseline risk, dose-dependent effects and other factors may cause seemingly similar studies to be different. For example, the evolution of HIV resistances makes the HIV population less comparable over time, while the effectiveness of the initial highly-active antiretroviral therapy improves rapidly over time. These increased the complexity in the evaluation of clinical and methodological diversity.

Statistical tests of heterogeneity are useful to identify variation among effects estimates, but their performance is influenced by number and size of studies[10] or choice of effect measures.[11] As a general rule, however, investigators should *not* decide whether to combine studies based on the p-value of a test of heterogeneity. When there is a large amount of clinical and methodological diversity along with high statistical heterogeneity such that any combined estimate is potentially misleading, the investigators should not combine the studies to produce an overall estimate. Instead, investigators should attempt to explore heterogeneity using subgroup analysis and meta-regression

if there is sufficient number of studies (see section on Test and Explore Statistical Heterogeneity) or describe the heterogeneity qualitatively. However, combining clinically or methodologically diverse studies can make sense if effect sizes are similar, particularly when the power to detect variation is large. In this situation, investigators should describe the differences among the studies and population characteristics, as well as the rationale for combining them in light of these differences. Ultimately the decision will be judged on whether combining the studies makes sense clinically, a criterion that is qualitative and perhaps subjective. Examples to illustrate how to make appropriate decisions based on evaluation of different types of heterogeneity are helpful to guide the consistent implementation of these principles and need to be developed by the EPC program.

Indirect Comparisons and Consideration of Indirect Evidence

Multiple alternative interventions for a given condition usually constitute a network of treatments. In its simplest form, a network consists of three interventions, for example, interventions A, B, and C. Randomized controlled trials (RCT) of A vs. B provide direct evidence on the comparative effectiveness of A vs. B; trials of A vs. C and B vs. C would provide indirect estimates of A vs. B through the "common reference," C. The inclusion of more interventions would form more complex networks and involve more complex indirect comparisons.[12,13]

Consideration of Indirect Evidence

Empirical explorations suggest that direct and indirect comparisons often agree,[13-18] but with notable exceptions.[19] In principle, the validity of indirect comparison relies on the invariance of treatment effects across study populations. However, in practice, trials can vary in numerous ways including population characteristics, interventions and cointerventions, length of followup, loss to followup, study quality, etc. Given the limited information in many publications and the inclusion of multiple treatments, the validity of indirect comparisons is often unverifiable. Moreover, indirect comparisons, like all other meta-analyses, essentially constitute an observational study, and residual confounding can always be present. Systematic differences in characteristics among trials in a network can bias indirect comparison results. In addition, all other considerations for meta-analyses, such as choice of effect measures or heterogeneity, also apply to indirect comparisons.

Therefore, in general, investigators should compare competing interventions based on direct evidence from head-to-head RCTs whenever possible. When head-to-head RCT data are sparse or unavailable but indirect evidence is sufficient, investigators could consider indirect comparisons as an additional analytical tool.[20] If the investigators choose to ignore indirect evidence, they should explain why.

Approaches of Indirect Comparison

The naïve indirect comparison—where the summary event rate for each intervention is calculated for all studies and compared—is unacceptable. This method ignores the randomized nature of the data and is subject to a variety of confounding factors. Confounders will bias the estimate for the indirect comparison in an unpredictable direction with uncertain magnitude.[21]

An alternative approach of indirect comparison is to use qualitative assessments by comparing the point estimates and the overlap of confidence intervals from direct comparisons. Two treatments are suggested to have comparable effectiveness if their direct effects vs. a common intervention have the same direction and magnitude, and there is considerable overlap in their confidence intervals. Under this situation, the qualitative indirect comparison is useful by saving the resources of going through formal testing and more informative than simply stating that there is no

available direct evidence. However, the degree of overlap is not a reliable substitute for formal testing. It is possible that the difference between two treatment effects is significant when there is small overlap of confidence intervals. When overlap in confidence intervals is less than modest and a significant difference is suspected, we recommend formal testing.

Indirect comparison methods range from Bucher's simple adjusted indirect comparisons[15] to more complex multi-treatment meta-analysis (MTM) models.[12,13,22,23] When there are only two sets of trials, say, A vs. C and B vs. C, Bucher's method should be enough to get the indirect estimate of A vs. B. More complex network needs more complex MTM models. Currently the investigators may choose any of the MTM models, and further research is required to evaluate their comparative performance and the validity of the model assumptions in practice. However, whichever method the investigators choose, they should assess the invariance of treatment effects across studies and appropriateness of the chosen method on a case-by-case basis, paying special attention to comparability across different sets of trials. Investigators should explicitly state assumptions underlying indirect comparisons and conduct sensitivity analysis to check those assumptions. If the results are not robust, findings from indirect comparisons should be considered inconclusive. Interpretation of findings should explicitly address these limitations. Investigators should also note that simple adjusted indirect comparisons are generally underpowered, needing four times as many equally sized studies to achieve the same power as direct comparisons, and frequently lead to indeterminate results with wide confidence intervals.[15,17]

MTM models provide the ability to check and quantify consistency or coherence of evidence for complex networks.[12,13,22,23] Consistency or coherence describes the situation that direct and indirect evidence agrees with each other, and when the evidence of a network of interventions is consistent, investigators could combine direct and indirect evidence using MTM models. Conversely, they should refrain from combining multiple sources of evidence from an incoherent network where there are substantial differences between direct and indirect evidence. Investigators should make efforts to explain the differences between direct and indirect evidence based upon study characteristics, though little guidance and consensus exists on how to interpret the results.

Choice of Effect Measures

Effect measures quantify differences in outcomes, either effectiveness or harms, between treatments in trials (or exposure groups in observational studies). The choice of effect measures is first determined by the type of outcomes. For example, relative risk and odds ratio are used for a binary outcome and mean difference is for a continuous outcome. They could also be broadly classified into absolute measures—such as risk differences or mean differences—and relative measures—such as odds ratio or relative risk. The number needed to treat (NNT) or harm (NNH) may also be considered effect measures, though they are usually not considered for meta-analyses as the standard error is rarely calculated or reported and normal approximation does not apply to NNT and NNH.

Binary Outcomes

Three measures are routinely used in a meta-analysis: the relative risk (RR), odds ratio (OR), and risk difference (RD). Criteria used to compare these measures include consistency over a set of studies, statistical property, and interpretability.[24] No single measure excels in all criteria.

The RD is most easily understood by clinicians and patients, and most useful to aid decision making, though it tends to be less consistent than relative measures (RR and OR) across studies. It is a preferred measure whenever estimates of RD are similar across studies and appropriate to be

combined. Usually in such cases, the proportions of events among control groups are relatively common and similar among studies. When events are rare, we don't recommend RD because combined estimates based on RD are often biased and have conservative confidence interval coverage and low statistical power.[25] When RD is not appropriate, RR is preferred over OR because it is easier to interpret clinically. RR and OR are effectively equivalent for rare events. However, RR is not a reversible measure in terms that if the definition of an outcome event and nonevent is switched, for example, from death to survival, the estimate of RR will be affected substantially and RR for death is not the reciprocal of RR for survival. The precision of the estimated RR would be affected, too. For RD and OR, such switch has no major consequence as OR for death is the reciprocal of OR for survival and the switch only changes the sign of RD. Therefore, while the definition of the outcome event needs to be consistent among the included studies when using any measure, the investigators should be particularly attentive to the definition of an outcome event when using a RR.

The reported measures or study design could prescribe the choice of effect measures. Case-control studies only allow the estimation of an OR. For observational studies, usually only relative measures are reported from a model adjusted for confounding variables. In another situation, when a subset of included studies only report, say, RR, without reporting raw data to calculate other measures, the choice could be determined by the reported measure in order to include all studies in the analysis.

To facilitate interpretation when a relative measure (RR or OR) is used, we recommend calculating a RD or NNT/NNH using the combined estimates at typical proportions of events in the control group. We also encourage the calculation of NNT/NNH when using RD. Investigators should calculate a confidence interval for NNT/NNH as well.[26,27]

Note that both absolute and relative effect measures convey important aspects of evidence. We consider it good practice to report the proportion of events from each intervention group in addition to the effect measure.

Continuous Outcomes

The two measures for continuous outcomes are mean difference and standardized effect sizes. The choice of effect measure is determined primarily by the scale of the available data. Investigators can combine mean differences if multiple trials report results using the same or similar scales. Standardized mean difference (SMD) is typically used when the outcome is measured using different scales. SMD is defined as the mean difference divided by a measure of within-group standard deviation and several estimators of SMD have been developed including Glass's Δ, Cohen's *d* and Hedge's *g*. Hedge also proposed an unbiased estimator of the population SMD.[28] Hedge's unbiased estimator should be used whenever possible; otherwise, Hedge's *g* is generally preferred over Cohen's *d* or Glass's Δ. Standardized mean differences of 0.3, 0.5 and 0.8 are suggested corresponding to small, medium, and large referents[29] and widely used, though they were not anchored in meaningful clinical context.

For some continuous outcomes, a meaningful clinically important change is often defined and patients achieving such change are considered as "responders."[30] Understanding the relationship between continuous effect measures and proportion of "response" is nascent and not straightforward. Further research is necessary and we currently recommend against inferring response rate from a combined mean difference.

Count Data and Time to Events

Rate ratio is used for count data and often estimated from a Poisson regression model. For time to event data, the measure is hazard ratio (HR), and most commonly estimated from the Cox proportional hazards model. Investigators can also calculate HR and its variance if observed and expected events can be extracted,[31, 32] although this is often quite difficult.[33]

Choice of Statistical Model for Combining Studies

Meta-analysis can be performed using either a fixed or a random effects model. A fixed effects model assumes that there is one single treatment effect across studies. Generally, a fixed effects model is not advised in the presence of significant heterogeneity. In practice, clinical and methodological diversity are always present across a set of included studies. Variation among studies is inevitable whether or not the test of heterogeneity detects it. Therefore, we recommend random effects models, with exceptions for rare binary outcomes (discussed in more details under Combining Rare Binary Outcomes). We recommend against choosing a statistical model based on the significance level of heterogeneity test, for example, picking a fixed effect model when the p-value for heterogeneity is more than 0.10 and a random effects model when $P < 0.10$.

A random effects model usually assumes that the treatment effects across studies follow a normal distribution, though the validity of this assumption may be difficult to verify, especially when the number of studies is small. When the results of small studies are systematically different from those of the large ones, the normality assumption is not justified either. In this case, neither the random effects model nor the fixed effects model would provide an appropriate estimate[8] and we recommend not combining all studies. Investigators can choose to combine the large studies if they are well conducted with good quality and expected to provide unbiased effect estimates.

General Considerations for Model Choice

The most commonly used random effects model, originally proposed by DerSimonian and Laird,[34] does not adequately reflect the error associated with parameter estimation. A more general approach has been proposed.[35] Other estimates are derived by using simple or profile likelihood methods, which provide an estimate with better coverage probability.[36] Likelihood based random effects models also account better for the uncertainty in the estimate of between-study variance. All these models could be used to combine measures for continuous, count and time to event data, as well as binary data when the events are common. For OR, RR, HR and rate ratio, they should be analyzed on the logarithmic scale. For OR, a logistic random effects model is another option.[37] When the estimate of between-study heterogeneity is zero, a fixed effects model (e.g., the Mantel-Haenszel method, inverse variance method, Peto method (for OR), or fixed effects logistic regression) could also be used for common binary outcomes and provide similar estimate to the DerSimonian and Laird approach. Peto method requires that no substantial imbalance exists between treatment and control group sizes within trials and treatment effects are not exceptionally large.

A special case: combining rare binary outcomes. When comparing rare binary outcomes, few or zero events often occur in one or both arms in some of the included studies. The normal approximation of the binomial distribution does not hold well and choice of model becomes complicated. A fixed effects model is often more appropriate for rare events based on simulation study, even under the conditions of heterogeneity,[38] because it provides less biased results and

better coverage property of the 95 percent confidence interval. However, investigators should note that no method gives completely unbiased estimates when events are rare.

When event rates are less than 1 percent, the Peto OR method is the recommended choice if the included studies have moderate effect sizes and the treatment and control group are of relatively similar sizes. This method provides the least biased, most powerful combined estimates with the best confidence interval coverage.[25] Otherwise when treatment and control group sizes are very different or effect sizes are large, or when events become more frequent (5 percent to10 percent), the Mantel-Haenszel method (without correction factor) or a fixed effects logistic regression provide better combined estimates and are recommended.

Exact methods have been proposed for small studies and sparse data.[39,40] However, simulation analyses did not identify a clear advantage of exact methods over a logistic regression or the Mantel-Haenszel method even in situations where the exact methods would theoretically be advantageous.[25] Therefore the investigators may choose to use exact methods but we don't specifically recommend exact methods over fixed effect models discussed above.

Considerations of correction factor for studies with zero events in one arm. In a study with zero events in one arm, estimation of effect measures (RR and OR) or their standard errors needs the addition of a correction factor, most commonly, 0.5 added to all cells. However, a combined estimate can be obtained using the Peto method, the Mantel-Haenszel method, or a logistic regression approach, without adding a correction factor. It has been shown that the Mantel-Haenszel method with the 0.5 correction does not perform as well as the uncorrected Mantel-Haenszel method or logistic regression,[25] nor as well as the Mantel-Haenszel method with alternative correction factors.[38] Therefore, we advise against the use of the Mantel-Haenszel method with the 0.5 correction. The investigators could choose adding no correction factors or exploring alternative correction factors using sensitivity analyses.[38]

Studies with zero events in both arms. When both arms have zero events, the relative measures (OR and RR) are not defined. These studies are usually excluded from the analysis as they do not provide information on the direction and magnitude of the effect size.[25,38] Others consider including studies without events in the analyses to be important and choose to include them using correction factors.[41,42] Inferential changes were observed when including studies without events[41] but the DerSimonian and Laird approach and RD[41] were used, which have been shown to have poor performance for rare events. [25]

We recommend that studies with zero events in both arms should be excluded from meta-analyses of OR and RR. The Peto method, fixed effects logistic regression (Bayesian or not), and the Mantel-Haenszel method effectively exclude these studies from the analysis by assigning them zero weight. Instead, the excluded studies could be qualitatively summarized, as in the hypothetical example below (Table 1), by providing information on the confidence intervals for the proportion of events in each arm. On the other hand, when the investigators estimate a combined control event rate, the zero events studies should be included, and we recommend the random effects logistic model that directly models the binomial distribution.[43]

Table 1. Example of a qualitative summary of studies with no events in both groups

	Intervention A		Intervention B	
Studies with zero events in both arms	Counts	One sided 97.5% exact confidence interval for the proportion of events	Counts	One sided 97.5% exact confidence interval for the proportion of events
Study 1	0/10	(0, 0.31)	0/20	(0, 0.168)
Study 2	0/100	(0, 0.036)	0/500	(0, 0.007)
Study 3	0/1000	(0, 0.004)	0/1000	(0, 0.004)

Bayesian Methods

Both fixed and random effects models have been developed within a Bayesian framework for various types of outcomes. The Bayesian fixed effects model provides good estimates when events are rare for binary data.[38] When the prior distributions are vague, Bayesian estimates are usually similar to estimates using the above methods, though choice of vague priors could lead to a marked variation in the Bayesian estimate of between-study variance when the number of studies is small.[44] Bayesian random models properly account for the uncertainty in the estimate of between-study variance.

We support the use of Bayesian methods with vague priors in CERs, if the investigators choose Bayesian methods. The statistical packages such as WinBUGS provide the flexibility of fitting a wide range of Bayesian models.[45] The basic principle to guide the choice between a random effects and a fixed effect model is the same as that for the above non-Bayesian methods, though the Bayesian method needs more work in programming, simulation and simulation diagnostic.

Test and Explore Statistical Heterogeneity

Investigators should assess heterogeneity for each meta-analysis. Visual inspection of forest plots and cumulative meta-analysis plots[46] are useful in the initial assessment of statistical heterogeneity. A test for the presence of statistical heterogeneity, for example, Cochran's Q test, as well as a measure for magnitude of heterogeneity, e.g., the I^2 statistic,[11,47] is useful and should be reported. Further, interpretation of Q statistic should consider the limitations of the test that it has low power when the number of studies is small and could detect unimportant heterogeneity when the number of studies is large. A p-value of 0.10 instead of 0.05 could be used to determine statistical significance. In addition, the 95 percent CI for I^2 statistic should also be provided, whenever possible, to reflect the uncertainty in the estimate.[48]

Investigators should explore statistical heterogeneity when present. Presentation and discussion of heterogeneity should distinguish between clinical, methodological and statistical heterogeneity when appropriate. Subgroup analysis or meta-regression with sensitivity analyses should be used to explore heterogeneity. When statistical heterogeneity is attributable to one or two "outlier" studies, sensitivity analyses could be conducted by excluding these studies. However, a clear and defensible rationale should be provided for identifying "outlier" studies. As discussed earlier, tests of statistical heterogeneity should not be the only consideration for the decision to combine studies or of the choice between a random or fixed effects model.

Subgroup analysis and meta-regression. Meta-regression models describe associations between the summary effects and study-level data, that is, it describes only *between-study*, not *between-patient*, variation. Subgroup analysis may be considered as a special case of meta-regression and involve comparison of subgroups of studies, for example, by study design, quality rating and other

topic-specific factors such as disease severity. Investigators should note the difference between two types of study-level factors: (1) factors that apply equally to all patients in a study, e.g., study design, quality and definition of outcomes, and (2) study-level summary statistics of individual patient-level data, e.g., mean age, percentage of diabetic patients.[49-51] Meta-regression is most useful with the first type of study-level factors. A meta-regression on summarized patient-level factors may be subject to ecological fallacy,[51] a phenomenon in which associations present at the study level are not necessarily true at the patient level. Therefore, interpretation of meta-regression on summary data should be restricted to the study level.

We encourage the use of subgroup analysis and meta-regression to explore heterogeneity, to investigate the contribution of specific factors to heterogeneity and obtain combined estimates after adjusting for study level characteristics, when appropriate. A random effects meta-regression should always be used, to allow residual heterogeneity not explained by study level factors. Whenever possible, study level factors, including subgroup factors, considered in meta-regressions should be prespecified during the planning of the CER and laid out in the key questions, though the actual data may be known to some extent when the analyses are being planned for a meta-analysis. Variables that are expected to account for clinical or methodological diversity are typically included, e.g., differences in populations, or interventions, or variability in the study design. Good knowledge of the clinical and biological background of the topic and key questions is important in delineating a succinct set of useful and informative variables. Use of permutation test for meta-regression can help assess the level of statistical significance of an observed meta-regression finding.[52]

When interpreting results, investigators should note that subgroup analyses and meta-regressions are observational in nature and suffer the limitations of any observational investigation, including possible bias through confounding by other study-level characteristics. As a general rule, association between effect size and the study-level variables (either pre- or post-specified) should be clinically plausible and supported by other external or indirect evidence, if they are to be convincing.

Number of studies required for a meta-regression. There is no universally accepted optimal minimum number of studies that are required for a meta-regression. The Cochrane handbook[8] suggests a minimum of 10 studies for each study-level variable without providing justifications, although fewer as six studies have been used in applied meta-regression empirical research.[50] The size of the studies and the distribution of subgroup variables are also important considerations. With the understanding that any recommended number has an arbitrary element, we advise a slightly different rule of thumb than the Cochrane handbook that when the sizes of the included studies are moderate or large, there should be at least 6 to 10 studies for a continuous study level variable; and for a (categorical) subgroup variable, each subgroup should have a minimum of 4 studies. These numbers serve as the lower bound for number of studies that investigators could start to consider a meta-regression. They are not the numbers that are sufficient for significant findings. The greater the number of studies, the more likely that clinically meaningful result is to be found. When the sizes of the included studies are small, it would take a substantial number of studies to produce useful results. When the number of studies is small, investigators should only consider one variable each time.

Combining studies of mixed designs. In principle, studies from different randomized trial designs, e.g. parallel, cross-over, factorial, or cluster-randomized design, may be combined in a single meta-

analysis. Investigators should perform a comprehensive evaluation of clinical and methodological diversity and statistical heterogeneity to determine whether the trials should actually be combined, and consider any important differences between different types of trials. For cross-over trials, investigators should first evaluate whether the trial is appropriate for the intervention and medical condition in question. The risk of carryover and the adequacy of the washout period should be fully evaluated. Estimates accounted for within-individual correlation are best for meta-analysis. Similarly for cluster randomized trials, estimates accounted for intra-cluster correlation are best for meta-analysis. More discussion on combining studies of mixed randomized trial designs is provided in the online appendix.

In addition to randomized trials, CER also examines observational studies, especially for harms, adherence, and persistence.[53] Trial and observational evidence often agree in their results.[54-56] However, discrepancies are not infrequent.[57] Though there are several examples in the literature,[58,59] synthesis across observational and randomized designs is fraught with theoretical and practical concerns and much research is necessary to assess the consistency between clinical trials and observational studies and investigate the appropriateness of and develop statistical methods for such cross-design synthesis. Currently, we recommend against combining clinical trials and observational studies in the same meta-analysis.

Sensitivity Analyses

Completing a CER is a structured process. Investigators make decisions and assumptions in the process of conducting the review and meta-analysis; each of these decisions and assumptions may affect the main findings. Sensitivity analysis should always be conducted in a meta-analysis to investigate the robustness of the results in relation to these decisions and assumptions.[60] Results are robust if decisions and assumptions only lead to small changes in the estimates and do not affect the conclusions. Robust estimates provide more confidence in the findings in the review. When the results are not robust, investigators should employ alternative considerations. For example, if the combined estimate is not robust to quality rating, investigators should report both estimates including and excluding studies of lesser quality and focus interpretation on estimates excluding studies of lesser quality. Investigators may also exclude studies of lesser quality.

Investigators should plan sensitivity analysis at the early stage of a CER, including tracking decisions and assumptions made along the way. Decisions and assumptions that might be considered in the sensitivity analysis include population or study characteristics, study quality and methodological diversity, choice of effect measures, assumptions of missing data, and so on. When necessary, multiple decisions and assumptions can be considered simultaneously.

Concluding Remarks

In this article, we provided our recommendations on important issues in meta-analyses to improve transparency and consistency in conducting CERs. The key points and recommendations for each covered issue are summarized in Table 2. Compared with the *Cochrane Handbook*, which explains meta-analysis methods in more detail, we focused on selected issues that present particular challenges in comparative effectiveness reviews. Overall there is no fundamental inconsistency between our recommendations and *Cochrane Handbook* on covered issues. We adopted the categorization of heterogeneity from the *Cochrane Handbook*, but provided more discussion of considerations for the decision to combine studies. For the choice of effect measures and statistical models, we favored RD and RR for binary outcome, and explicitly recommended random effects model except for rare binary outcome. Our recommendations and those of the *Cochrane Handbook*

follow similar principles to test and explore heterogeneity though we proposed a slightly different rule on the number of studies adequate for meta-regression and distinguished between continuous vs. subgroup study level covariates.

Table 2. Summary of key points and recommendations for quantitative synthesis in Comparative Effectiveness Reviews

Decision to combine studies
1. The decision to combine studies should depend on whether a meaningful answer to a well formulated research question can be obtained.
2. Investigators should make decisions of combining studies based on thorough investigations of clinical and methodological diversity as well as variation in effect size.
3. Statistical tests of heterogeneity are helpful, but investigators should *not* make a decision on combining studies based *only* on tests of heterogeneity.
4. When there is a large amount of clinical and methodological diversity along with high statistical heterogeneity such that any combined estimate is potentially misleading, the investigators should not combine the studies.
5. Combining clinically or methodologically diverse studies may make sense if there is no real difference among effect sizes, particularly when the power to detect variation is large.
6. Reasons to combine or to not combine studies and steps taken to reach the decision should be fully explained.
7. The purpose of a meta-analysis should be explicitly stated in the methods section of the CER.
Indirect comparison
1. In the absence of sufficient direct head-to-head evidence and presence of sufficient indirect evidence, indirect comparisons can be considered as an additional analytic tool.
2. The unadjusted (naïve) indirect comparison method is not recommended in any case.
3. A qualitative indirect comparison may be useful to judge comparable effectiveness when there is a large degree of overlap in confidence intervals, but we recommend formal testing when significant difference is suspected.
4. Validity of the adjusted indirect comparison methods depends on the consistency of treatment effects across studies, and the appropriateness of an indirect comparison needs to be assessed on a case-by-case basis.
5. Adjusted indirect comparison methods, such as Bucher's method or mixed treatment comparison, should be used for indirect comparison.
6. Investigators should conduct sensitivity analysis to check the assumptions of the indirect comparison. If the results are not robust to the assumptions, findings from indirect comparisons should be considered as inconclusive.
7. Investigators should make efforts to explain the differences between direct and indirect evidence based upon study characteristics.

Table 2. Summary of key points and recommendations for quantitative synthesis in Comparative Effectiveness Reviews (continued)

Choice of effect measure
1. For dichotomous outcomes, RD is a preferred measure whenever appropriate. Otherwise, RR is preferred over OR.
2. A relative measure (RR or OR) instead of RD should be used when the events are rare.
3. When using a relative measure, risk differences and NNT/NNH should be calculated using the combined estimates at typical proportions of event in the control group. Calculation of NNT/NNH when using RD is also encouraged.
4. Calculation of NNT/NNH should include both point estimate and confidence interval.
5. Proportion of events from each intervention group should be reported in addition to the effect measure.
6. For continuous outcomes, mean difference should be used if results are reported using the same or similar scales and standardized mean difference should be used when results are reported in different scales.
7. For standardized mean difference, Hedge's unbiased estimator should be used whenever possible. Otherwise, Hedge's *g* is generally preferred over Cohen's *d* or Glass's Δ.
8. Rate ratio should be used for count data and hazard ratios for time-to-event data.
Choice of model
1. A random effects model is recommended since clinical and methodological diversity are inevitable among included studies.
2. A fixed effects model is recommended for rare binary events, and the choice of a fixed effects model depends on the event rate, effect size, and the balance of intervention groups.
3. For rare binary events: 3.1. Studies with zero events in one arm should be included in the analyses. 3.2. When event rates < 1%, the Peto OR method is recommended when no substantial imbalance exists between treatment and control group sizes within trials and treatment effects are not exceptionally large. In other situations, the Mantel-Haenszel method or a fixed effects logistic regression provides better combined estimates and are recommended. 3.3. For the Mantel-Haenszel method, a correction factor of 0.5 is not recommended but using no correction factor or alternative correction factors could be considered, and investigated in sensitivity analyses when necessary. 3.4. Studies with zero events in both arms should be excluded from the analyses but should be summarized qualitatively.
4. Use of Bayesian methods with vague priors in CERs is supported, if the investigators choose Bayesian methods.

Table 2. Summary of key points and recommendations for quantitative synthesis in Comparative Effectiveness Reviews (continued)

Test and explore heterogeneity
1. Visual inspection of forest plots and cumulative meta-analysis plots are useful in the initial assessment of heterogeneity.
2. Heterogeneity should be assessed for each meta-analysis and both measures of the statistical significance and magnitude of heterogeneity should be reported.
3. Interpretation of statistical significance (for Q statistics) should consider the limitations of the test and the 95% CI for the estimate of magnitude of heterogeneity should be provided, whenever possible.
4. Presentation and discussion of heterogeneity should distinguish between clinical diversity, methodological diversity, and statistical heterogeneity when appropriate.
5. Heterogeneity should be explored using subgroup analysis or meta-regression or sensitivity analyses.
6. When heterogeneity is caused by one or two "outlier" studies, sensitivity analyses are recommended by excluding such studies.
7. Meta-regression (including subgroup analyses) is encouraged to explore heterogeneity.
8. Pre-specified meta-regression based on the key questions should be used to explore heterogeneity as much as possible.
9. A random effects meta-regression should be used.
10. Meta-regression is observational in nature, and if the results of meta-regression are to be considered valid, they should be clinically plausible and supported by other external or indirect evidence.
Combining studies of mixed designs
1. If cross-over trials are appropriate for the intervention and medical condition in question, and there are no systematic differences between the two types of design, cross-over designs can be combined with parallel trials.
2. Meta-analysis of cross-over trials should use estimates from within-individual comparisons whenever available.
3. If cluster-randomization trials are appropriate for the intervention and medical condition in question, and there are no systematic differences between the different types of design, cluster-randomization trials can be combined with individual-randomized trials.
4. When available, effect measures from an analysis that appropriately accounts for the cluster design should be used for meta-analysis.
5. Clinical trials and observational studies should not be combined.
Sensitivity analyses
1. A CER with a meta-analysis should always include sensitivity analyses to examine the robustness of the combined estimates in relation to decisions and assumptions made in the process of review.
2. Planning of sensitivity analysis should start at the early stage of a CER, and investigators should keep track of key decisions and assumptions.
3. When necessary, multiple decisions and assumptions may be considered at the same time.

This article does not address every major issue relevant to meta-analyses. Other interesting topics, such as meta-analysis of individual patient data, meta-analysis of diagnostic tests, assessing bias including publication bias, as well as more specific issues such as how to handle different comparators, composite outcomes or selective reporting will be considered in future versions of the EPC methods guide for CER. Meta-analysis methods for observational studies including combining observational studies, assessing bias for observational studies, incorporation of both clinical trials and observational studies, and even indirect comparison of observational studies will also be topics for both future version of guidelines and future research. As in most research areas, quantitative synthesis is a dynamic area with a lot of active research going on. Correspondingly, development of

guidelines is an evolving process and we will update and improve recommendations with the accumulation of new research and improved methods to advance the goal for transparency and consistency.

References

1. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 2007;356:2457–2471.
2. Dahabreh IJ, Economopoulos K. Meta-analysis of rare events: an update and sensitivity analysis of cardiovascular events in randomized trials of rosiglitazone. Clin Trials 2008;5:116–120.
3. Diamond GA, Bax L, Kaul S. Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. Ann Intern Med 2007;147:578–581.
4. Shuster JJ, Jones LS, Salmon DA. Fixed vs random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death. Stat Med 2007;26:4375–4385.
5. Committee on Oversight and Government Reform. Hearing on FDA's Role in Evaluating Safety of Avandia. Available at: http://oversight.house.gov/index.php?option=com_content&view=article&id=3710&catid=44%3Alegislation&Itemid=1. Accessed May 31, 2010.
6. Agency for Healthcare Research and Quality. Evidence-based Practice Centers. Available at: http://www.ahrq.gov/clinic/epc/. Accessed May 31, 2010.
7. Helfand M, Balshem H. Principles for developing guidance: AHRQ and the effective health care program. J Clin Epidemiol 2010;63: 484-490.
8. Higgins, J. Cochrane handbook for systematic reviews of interventions. Available at: http://www.cochrane.org/resources/handbook/. Accessed May 31, 2010.
9. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–1558.
10. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med 1998;17:841–56.
11. Engels EA, Schmid CH, Terrin N, et al. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med 2000;19:1707–1728.
12. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105–3124.
13. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002;21:2313–2324.
14. Baker SG, Kramer BS. The transitive fallacy for randomized trials: if A bests B and B bests C in separate trials, is A better than C? BMC Med Res Methodol 2002;2:13.
15. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50:683–691.
16. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005;331:897–900.
17. Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. Health Technol Assess 2005;9:1–148.
18. Song F, Glenny AM, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. Control Clin Trials 2000;21:488–497.
19. Chou R, Fu R, Huffman LH, et al. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. Lancet 2006;368:1503–1515.
20. Ioannidis JP. Indirect comparisons: the mesh and mess of clinical trials. Lancet 2006;368:1470–1472.
21. Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003;326:472.
22. Dominici F, Parmigiani G, Wolpert R, et al. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. J Am Stat Assoc 1999;94:16–28.

23. Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 2006;101:447–459.
24. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21:1575–1600.
25. Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med 2007;26:53–77.
26. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452–454.
27. Schulzer M, Mancini GB. 'Unqualified success' and 'unmitigated failure:' number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment-induced adverse events. Int J Epidemiol 1996;25:704–712.
28. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. J Educ Stat 1981;6:107-128.
29. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: L. Erlbaum Associates; 1988.
30. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis 2005;64:29–33.
31 Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 1998;17:2815–2834.
32. Tierney J, Stewart L, Ghersi D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials 2007;8:16.
33. Duchateau L, Collette L, Sylvester R, et al. Estimating number of events from the Kaplan-Meier curve for incorporation in a literature-based meta-analysis: what you don't see you can't get! Biometrics 2000;56:886–892.
34. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–188.
35. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials 2007;28:105–114.
36. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001;20:825–840.
37. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med 1995;14:2685–2699.
38. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004;23:1351–1375.
39. Mehta CR. The exact analysis of contingency tables in medical research. Cancer Treat Res 1995;75:177–202.
40. Mehta CR, Patel NR. Exact logistic regression: theory and examples. Stat Med 1995;14:2143–2160.
41. Friedrich JO, Adhikari NK, Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. BMC Med Res Methodol 2007;7:5.
42. Sankey S, Weissfeld L, Fine M, et al. An assessment of the use of the continuity correction for sparse data in meta-analysis. Communications in statistics—Simulation and computation 1996;25:1031–1056.
43. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. J Clin Epidemiol 2008;61:41–51.
44. Lambert PC, Sutton AJ, Burton PR, et al. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med 2005;24:2401–2428.
45. The BUGS Project. WinBUGS. Available at: http://www.mrc-bsu.cam.ac.uk/bugs/. Accessed May 31, 2010.
46. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol 1995;48:45–57; discussion 9–60.
47. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–560.

48. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ 2007;335:914–916.
49. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. Lancet 1998;351:123–127.
50. Schmid CH, Lau J, McIntosh MW, 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;17:1923–1942.
51. Schmid CH, Stark PC, Berlin JA, et al. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol 2004;57:683–697.
52. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004;23:1663–1682.
53. Slutsky J, Atkins D, Chang S, et al. Comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2008; in press.
54. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–1892.
55. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–1886.
56. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286:821–830.
57. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–228.
58. Droitcour J, Silberman G, Chelimsky E. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. Int J Technol Assess Health Care 1993;9:440–449.
59. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. Stat Med 2000;19:3359–3376.
60. Olkin I. Re: "A critical look at some popular meta-analytic methods." Am J Epidemiol 1994;140:297–299; discussion 300–301.