Short Research Review

Performing and evaluating meta-analyses

Jakob Burcharth, MD, PhD, Hans-Christian Pommergaard, MD, PhD, *and*
Jacob Rosenberg, MD, DMSc, *Herlev, Denmark*

From the Centre for Perioperative Optimization, Department of Surgery, Herlev Hospital, University of Copenhagen, Herlev, Denmark

Literature reviews can be performed as narrative or systematic. The narrative review often is written by experts and generally provides an extensive overview of the literature; however, the narrative review does not provide transparency of the review process. In contrast, systematic reviews include a description of a predefined research question, a systematic search strategy, and a screening and selection strategy that uses predefined inclusion and exclusion criteria; the systematic review evaluates the quality of the included studies. Together, this type of review potentially allows the reader to reproduce the review process and add transparency and objectivity to the work. If some or all data from the included studies in a systematic review also are presented in a quantitative synthesis, it is named a meta-analysis.[1]

Meta-analyses are powerful tools for summarizing knowledge as well as estimating treatment effects with greater precision. Furthermore, meta-analyses are able to identify possible publications bias (eg, when small studies with undesirable findings are not being published).[2] The number of published meta-analyses are increasing dramatically, and more than 18% of all published meta-analyses indexed in MEDLINE were published in the year 2013 (Fig 1). Despite the usefulness of systematic reviews and meta-analyses in answering important clinical questions, these reviews are sometimes controversial and should be interpreted with care, because results can be misleading if the methodology is inappropriate. Criticism of the meta-analytic method includes that the method is based on data that are extracted and integrated from a number of independent studies instead of random sampling of data, which means that the results from a meta-analysis cannot test causality. This article introduces and should help to clarify the basic methodologic principles of performing and evaluating meta-analyses.

PLANNING AND REGISTERING A REVIEW

The Cochrane Handbook of Systematic Reviews of Interventions[2] describes the details of performing the specific steps of systematic reviews and meta-analyses, and the Preferred Reporting Items of Systematic Reviews and Meta-analysis guideline, which consists of a checklist and a 4-phased flowchart, describes the preferred content of each passage when reporting a systematic review and meta-analysis.[3] The Meta-analyses Of Observational Studies in Epidemiology checklist describes how to report meta-analyses of observational studies.[4] Following Preferred Reporting Items of Systematic Reviews and Meta-analysis or Meta-analyses Of Observational Studies in Epidemiology guidelines ensures systematic reporting and transparency in the review process.

When planning a systematic review and meta-analysis, it is advisable to publish a detailed protocol before commencing the review. By registering a protocol prospectively at the PROSPERO (International Prospective Register of Systematic Reviews) webpage,[5] it is possible to prevent multiple reviews addressing the same question. This registration process will also decrease the risk of publication bias in the event of negative results, because registration will bind review authors to the literature search, analysis plan, bias evaluation,

J.R. received lecture fees from Merck and Bard and research funding from Baxter Healthcare, Johnson & Johnson, and Bard.

Accepted for publication August 22, 2014.

Reprint requests: Jakob Burcharth, MD, PhD, Centre for Perioperative Optimization, Department of Surgery, Herlev Hospital, University of Copenhagen, Herlev 2730, Denmark. E-mail: jakobburcharth@gmail.com.

Surgery 2015;157:189-93.

0039-6060/$ - see front matter

http://dx.doi.org/10.1016/j.surg.2014.08.087

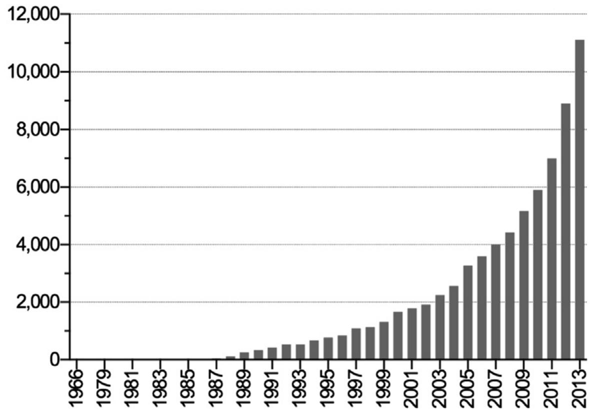


Fig 1. Number of published meta-analyses by year. A total of 52,626 meta-analyses are published in MEDLINE, and of those, 18% were published in the year 2013.

outcome selection, and reported outcomes. The registration is expected to be kept up-to-date, and despite strong encouragement from several opinion-leaders for the registration of reviews at PROSPERO,[6] this process is still voluntary.

The research question of a systematic review and meta-analysis can either be broad or narrow. A broad research question increases the chance of applying the findings to a wider population; however, it also increases the risk of too much variation between the included studies (eg, heterogeneity). A narrow research question can lead to difficulties in finding enough includable studies as well as generalizing the results. In each case, the research question and eligibility criteria of the includable studies should be clinically relevant and may be defined by PICO(S) (ie, Population, Intervention, Comparison, Outcome(s), and Study type).[7]

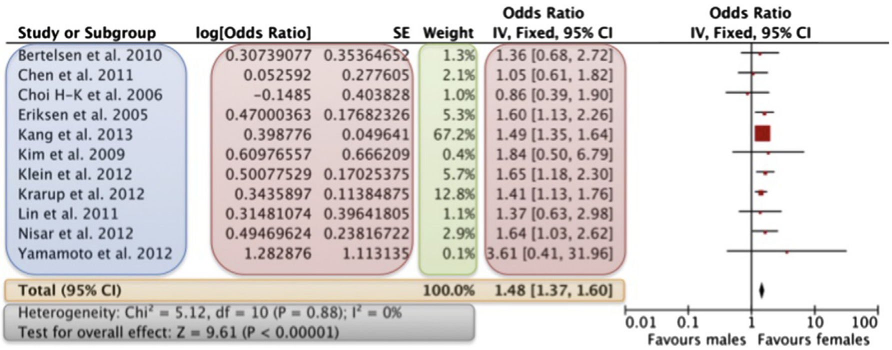
Risk of bias within studies. The degree of bias within the included studies determines the certainty of which conclusions can be drawn. Just as with the process of screening and study selection, the assessments of risk of bias of the included studies should be performed by at least 2 authors.

Depending on the study design of the included studies, more than 190 different tools have been developed to assess the risk of bias within the studies.[8] In evaluating randomized controlled trials, the Jadad score[9] and the Cochrane risk of bias tool[2] are the most commonly used tools to evaluate risk bias. When nonrandomized studies are being evaluated, the Newcastle-Ottawa scale or the Downs and Black checklist are the recommended tools to evaluate risk bias in these types of studies.[2]

Heterogeneity. Criticism of meta-analyses includes the possibility of inappropriately combined studies that are quite different, which leads to a risk of the results being incorrect reflections of the true effect. Only studies without major bias and with comparable designs, interventions, patients, and measures of outcome should be included and combined.

Studies that are brought together in reviews addressing a specified research question will differ inevitably with a degree of diversity, either because of clinical or methodologic differences among the included studies. In meta-analyses, this difference is termed heterogeneity.[2] Heterogeneity arises when the observed outcome effects determined from the included studies are more different than expected by random chance alone. Clinical differences between the studies can lead to heterogeneity if the outcome is affected by factors that vary across the studies (ie, different patient characteristics or different study interventions). Moreover, methodologic differences can lead to heterogeneity if the included studies are of different design. Assessment of heterogeneity in meta-analyses is crucial, because the results otherwise may be misleading.

It can be argued that some degree of heterogeneity always will exist in meta-analyses, whether or not it can be detected by statistical tests. This determination depends on how heterogeneity is measured and quantified, because heterogeneity can be determined in several different ways. Visually, a lacking overlap in the Forest plot of the horizontal lines representing confidence intervals of the included studies will indicate some degree of heterogeneity (Fig 2).[10] The RevMan program includes measures automatically of the heterogeneity in the Forest plots that determine whether there is a greater spread of the results between the studies than due strictly to chance. One of these heterogeneity measures is the I^2 (also called inconsistency).[2] The I^2 quantifies the degree of variability among the included studies by the use of the χ^2 and the degrees of freedom (dependent on the number of included studies in the meta-analysis) from the pooled estimate. The I^2 can range from 0 to 100% and is easily interpretable since 0% indicates no heterogeneity, and 100% indicates complete heterogeneity. Heterogeneity is generally considered high when $I^2 > 50\%$ and as being substantial when $I^2 > 75\%$.[2] Before reporting a meta-analysis, heterogeneity needs to be investigated and preferably be explained to determine if data can be combined and presented reliably. Normally the pooled estimate should not be reported if $I^2 > 75\%$; however, because no consistent rules exist regarding this



Study or Subgroup	log[Odds Ratio]	SE	Weight	Odds Ratio IV, Fixed, 95% CI
Bertelsen et al. 2010	0.30739077	0.35364652	1.3%	1.36 [0.68, 2.72]
Chen et al. 2011	0.052592	0.277605	2.1%	1.05 [0.61, 1.82]
Choi H-K et al. 2006	-0.1485	0.403828	1.0%	0.86 [0.39, 1.90]
Eriksen et al. 2005	0.47000363	0.17682326	5.3%	1.60 [1.13, 2.26]
Kang et al. 2013	0.398776	0.049641	67.2%	1.49 [1.35, 1.64]
Kim et al. 2009	0.60976557	0.666209	0.4%	1.84 [0.50, 6.79]
Klein et al. 2012	0.50077529	0.17025375	5.7%	1.65 [1.18, 2.30]
Krarup et al. 2012	0.3435897	0.11384875	12.8%	1.41 [1.13, 1.76]
Lin et al. 2011	0.31481074	0.39641805	1.1%	1.37 [0.63, 2.98]
Nisar et al. 2012	0.49469624	0.23816722	2.9%	1.64 [1.03, 2.62]
Yamamoto et al. 2012	1.282876	1.113135	0.1%	3.61 [0.41, 31.96]
Total (95% CI)			100.0%	1.48 [1.37, 1.60]

Heterogeneity: Chi² = 5.12, df = 10 (P = 0.88); I² = 0%
Test for overall effect: Z = 9.61 (P < 0.00001)

Fig 2. An example of a Forest plot from a study that evaluates sex as a risk factor for leakage of colon anastomosis.[10] The included studies are to the left in the *blue study* or *subgroup box* and are often named by the first author and year of publication. The results from the individual studies are seen in the *red boxes*, which in this example are based on the odds ratios and 95% confidence intervals from the included studies. The *green box* indicates the weight given to each study based on the 95% confidence intervals (narrow 95% confidence intervals are weighted greater). In this example, the random effects model was used. The *gray box* indicates the meta-analysis statistics (χ^2 and I^2 test) and the statistical significance for the overall effect (in this example, $P < .00001$). The overall effect size is shown in the *orange box* and is displayed as a combined odds ratio from the included studies and 95% confidence interval. To the right, the graphic presentation of the results is shown as a Forest plot. A *black line* and a *red box* represent each study horizontally. The *black horizontal lines* represent the 95% confidence interval, and the *red box* indicates the result of study and the weight of the study (*larger boxes* indicating greater weight of the study). If both the *horizontal line* and the box lie to the left of the *vertical line*, then the study shows a statistically significant effect in favor of males. If both the *horizontal line* and box lie to the right of the *vertical line*, then the study favors females. If the *horizontal line* crosses the *vertical line*, then the study is not statistically significant. The overall effect is depicted graphically by a diamond, with the size of the diamond showing the 95% confidence interval of the overall effect. If the diamond touches the *vertical line*, then the overall effect is not statistically significant.

concept of degree of heterogeneity, authors should predefine the limit of heterogeneity that is acceptable with regard to presenting a pooled outcome estimate.

Assessing the quality of outcomes. When evaluating the quality of narrative outcomes not appropriate for meta-analysis, the authors must evaluate the quality of the outcome subjectively. When evaluating the overall quality of the outcomes of meta-analyses, the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach can be used with a software tool called the GRADE profiler assessment.[2] The level of quality of an outcome can be graded as “high,” “moderate,” “low,” or “very low” based on the study methodology and grading factors. The GRADE approach uses several factors that can downgrade or upgrade the overall quality estimate. The factor that downgrade the overall quality estimate of a meta-analysis are: a moderate or high risk of bias (from the risk of bias tool), high degree of inconsistency (measured by I^2), indirectness (for example when a study is performed on a population subgroup, which limits the generalizability of the results), imprecision of the effect estimate (when the studies are small in size and have wide confidence intervals), and publication bias (visualized from the funnel plot). The factor that upgrade the overall quality estimate of are: a large or very large effect estimate ($0.5 > RR > 2$ or $0.2 > RR > 5$, respectively), confounding changes of the effect estimate that lower the effect estimate (all plausible confounding would decrease a demonstrated effect when results show no effect), or occurrence of a dose-response gradient (eg, increased risk of outcome when increased levels of exposure of a drug). Observational studies by definition start as low quality in the GRADE approach because of the risk of bias and confounders inherent the study design, in contrast to randomized controlled trials that start as high quality prior to upgrading or downgrading.

COMBINING DATA IN A META-ANALYSIS

A free, well-functioning software program for performing meta-analysis is the RevMan program which is freely available from the Cochrane collaboration.[2] In the RevMan program, data (categorical or continuous) are entered and combined statistically. The results from the analysis are presented visually by a Forest plot (Fig 2) and a funnel plot (Fig 3). Regardless of the statistical method used in this analysis, the study data will be weighed according to the variance of data (spreading of

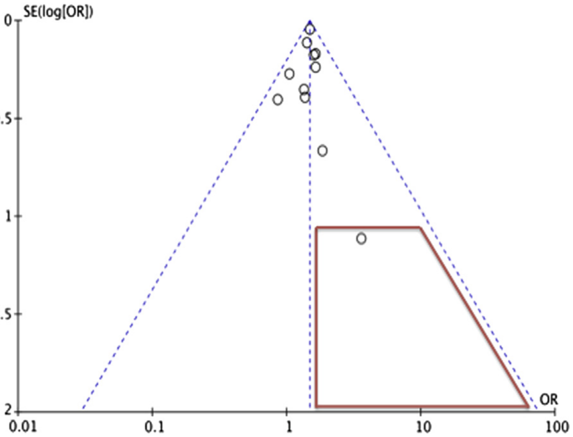


Fig 3. Funnel plots are used to visualize possible publication bias. In this funnel plot, the studies from the meta-analysis[10] are plotted on the horizontal axis by the odds ratio against the size of the study on the vertical axis (measured by the standard error of the odds ratio (SE(log)[OR]). A smaller spread of the data indicates larger studies, and the larger the study the further up the vertical axis the study will be depicted. The overall effect from the meta-analysis (Fig 2) was odds ratio 1.48 (95% confidence interval 1.37–1.60), which is depicted by the *vertical dotted line*. The *dotted diagonal lines* are +/− 1.96 × standard error of pooled estimate on each side and represent the triangular region where 95% of studies should lie if no bias or heterogeneity is present. It is assumed that the larger the studies (or the studies with the least spread of data), the closer to the true result they will be. The *red* marking illustrates where small negative studies would be located graphically. Absence of small negative studies is usually a usual sign of publication bias.

data). Less variance (large study sample size) will contribute more heavily to the overall estimate than a large variance (smaller study sample size).

There are 2 statistical models to choose from in a meta-analysis: the fixed effects model and the random effects (RE) model. The common assumption of both models is that there are enough similarities between the included studies of the meta-analysis to synthesize a pooled estimate. In choosing the fixed effects model for a meta-analysis, it is assumed that the study subjects, interventions, and outcomes do not differ between the studies. In choosing the RE model for meta-analysis, it is assumed that the subjects, interventions, and outcomes of the included studies differs. Any common effect size can, therefore, not be assumed, and the RE attempts to account for, but not explain, such differences. In cases with heterogeneity, the RE model should be chosen, because the RE model provides a more conservative estimate of the pooled effect estimate. In cases of no heterogeneity, the 2 models produce identical pooled estimates.

Table I. Methods for evaluating bias and quality of outcomes in meta-analyses

Evaluating bias	
Randomized controlled trials	• Cochranes risk of bias tool • Jadad score
Nonrandomized studies	• Newcastle Ottawa scale* • Downs and Black checklist*
Evaluating quality of outcomes	
Meta-analysis outcome	• GRADE approach
Narrative outcome	• Subjective evaluating by authors

*More than 190 different scores and scales have been developed to evaluate risk of bias in non-randomized studies[8]; however, the Newcastle-Ottawa scale or the Downs and Black checklist are recommended by the Cochrane Collaboration.[2]

Several methods have been developed for evaluating risk of bias and quality of outcomes in meta-analyses depending on the design of the study.

GRADE, Grading of Recommendations Assessment, Development and Evaluation.

The usual way for displaying data from a meta-analysis is by the use of graphs. The Forest plot depicts the details of the analysis by highlighting the study effect sizes, the pooled effect estimate (weighted average of study estimates), confidence intervals, and the estimates of heterogeneity (I^2; Fig 2). The funnel plot evaluates the presence of publication bias. In a funnel plot, the effect size is plotted versus a measure of its precision (ie, study size) (Fig 3). If no publication bias is present, the included studies are distributed symmetrically around the pooled effect size forming a funnel. In the case of publication bias, it will be expected that the studies are asymmetric with a lack of smaller studies with negative study findings (Fig 3). Analyses that do not take missing negative and smaller studies into account will tend to overestimate a potential treatment effect.

MAKING CONCLUSIONS OF META-ANALYSES

When appraising a review and a meta-analysis, it is imperative that the reader assess critically the methodology (research question, PICO(S), literature search, screening, study selection, and analysis plan) and whether it was predefined and how assessments of the risk of bias was performed (Table). The evaluation of the size of the pooled effect and the possible causes of heterogeneity are involved in the interpretation of the results. The objective conclusions of a meta-analysis should always be based in the degree of evidence from the included studies as well as assessments of the risk of bias, and it is important that the authors refer back to the original question asked.

REFERENCES

1. Glass GV. Primary, secondary and meta-analysis of research. Educ Res 1976;5:3-8.
2. Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011. Available from: www.cochrane-handbook.org.
3. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6: e1000097.
4. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. J Am Med Assoc 2000;283:2008-12.
5. PROSPERO webpage. Available from: http://www.crd.york.ac.uk/NIHR_PROSPERO/. Accessed April 30, 2014.
6. Davies S. The importance of PROSPERO to the National Institute for Health Research. Systematic Rev 2012;1:5.
7. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak 2007;7:16.
8. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:27.
9. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1-12.
10. Pommergaard HC, Gessler B, Burcharth J, Angenete E, Haglind E, Rosenberg J. Preoperative risk factors for anastomotic leakage after resection for colorectal cancer: a systematic review and meta-analysis. Colorectal Dis 2014;16:662-71.