Otolaryngology–Head and Neck Surgery (2010) 142, 6-14

INVITED ARTICLE

A practical guide to understanding systematic reviews and meta-analyses

J. Gail Neely, MD, Anthony E. Magit, MD, Jason T. Rich, MD, Courtney C. J. Voelker, MD, DPhil (Oxon), Eric W. Wang, MD, Randal C. Paniello, MD, Brian Nussenbaum, MD, and **Joseph P. Bradley, MD,** St. Louis, MO; and San Diego, CA

No sponsorships or competing interests have been disclosed for this article.

ABSTRACT

A systematic review is a transparent and unbiased review of available information. The published systematic review must report the details of the conduct of the review as one might report the details of a primary research project. A meta-analysis is a powerful and rigorous statistical approach to synthesize data from multiple studies, preferably obtained from a systematic review, in order to enlarge the sample size from smaller studies to test the original hypothesis and/or to generate new ones.

The objective of this article is to serve as an easy to read practical guide to understand systematic reviews and meta-analyses for those reading them and for those who might plan to prepare them.

A systematic review is the process of identifying and evaluating multiple studies on a topic using clearly defined methodology. It is similar to performing primary research, including describing the methods, collecting data, and performing analysis—not necessarily meta-analysis. This approach can be used for observational or interventional studies and is not limited to randomized clinical trials. Many literature reviews are not systematic and are known as narrative reviews, which may be incomplete, usually fail to describe how the literature was selected, and may be suspect of being highly biased by the author(s). The purpose of a systematic review is to present a transparent and unbiased review of available information.

Meta-analysis is a powerful and rigorous statistical approach to synthesize data from multiple studies, preferably obtained from a systematic review, in order to enlarge the sample size from smaller studies to test the original hypothesis and/or to generate new ones.[1] It has a long and distinguished history (Fig 1).[2] However, meta-analysis may be inappropriately applied when insufficient data are available. A recent study found that the majority of meta-analyses in the field of otolaryngology had methodological flaws.[3]

The objective of this article is to serve as an easy-to-read, practical guide to understanding systematic reviews and meta-analysis for those reading them and those who might plan to prepare them.

Systematic Review

The major components of a systematic review are as follows:[4] 1) define objective, 2) develop protocol, 3) develop search strategy, 4) search, 5) identify relevant papers, 6) screen papers, 7) select papers, 8) evaluate quality of papers and construct data extraction forms, 9) analyze and synthesize findings, 10) determine if meta-analysis is appropriate and select model, 11) report results without meta-analysis when data are insufficient for meta-analysis.

Upon completion of the systematic review, if the data are sufficient for further analysis, then the effect size is analyzed, meta-analysis is performed, and the results may be reported.

1. Define Objective

The first step is to define the specific objective of the review. Objective here usually means the driving hypothesis for the review. This step includes the construction of a research question in terms of specific variables of interest. The relevant variables are dependent upon the type of question asked.[5] The research question may focus upon 1) clinical findings, 2) etiology, 3) differential diagnosis, 4) diagnostic tests, 5) prognosis, 6) therapy, 7) prevention, 8) self-improvement, 9) costs. The variables of interest within these categories of questions contain all or most of the following four elements: 1) specific population and problem, 2) intervention (cause, exposure, prognostic factor, or treatment), 3) comparison intervention, 4) outcome of interest (Fig 2).

Received April 3, 2009; revised August 13, 2009; accepted September 16, 2009.

0194-5998/$36.00
doi:10.1016/j.otohns.2009.09.005

At this early stage, it is worthwhile to rapidly purview the literature in order to see what is available. The focus of this rapid, nonsystematic look is to 1) see what types of literature are available to address the question, 2) determine what support there is for the development of a testable hypothesis and how the hypothesis germane to the question has been assessed, and 3) determine what effect sizes might be gleaned from the literature; this is particularly important so that the proper data during the review might be extracted.

2. Develop Protocol

Just as is done in a primary research protocol, the specific characteristics of the patient/subject population, including potentially eligible subjects, and inclusion and exclusion criteria, are set in advance. In the case of systematic reviews, the individuals included in the "subject population" are the studies to be reviewed. The selection criteria for the studies focus upon the category of question being asked and the variables of interest germane to the question (see Define Objective above) and usually include the participants, interventions, and outcomes stated in each study. Because the idea of the systematic review is to attempt to synthesize a larger pool of results from smaller studies and thus generate a truer picture of nature's reality, the details within each study must fit the overall research question being asked of the review. Data extraction forms and article quality assessment tools are designed or selected at this stage.

Types of studies. If the study objective is the comparison of treatment interventions and there are many good randomized controlled trials (RCTs), the "gold standard" to reduce bias, then the type of studies for review may be limited to RCTs. On the other hand, if there are not sufficient RCTs or the question does not lend itself to that technique of subject group allocation, other types of studies may be more appropriate, such as cohort, case-control, or even case-series studies. For each type of question, it is important to define the levels of evidence most appropriate. These levels of evidence are outlined in an easily accessible table from the Oxford Centre for Evidence-based Medicine–Levels of Evidence.[6]

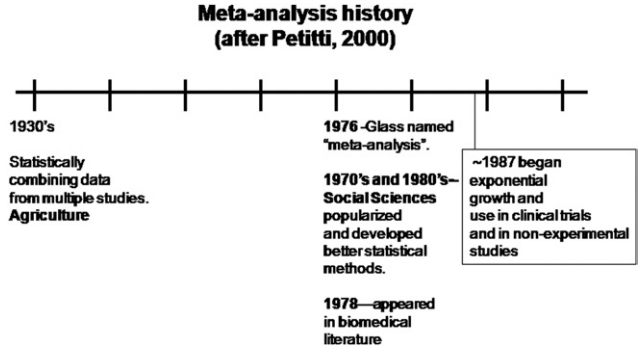

Figure 1 A timeline illustration of meta-analysis, after Petitti.[2]

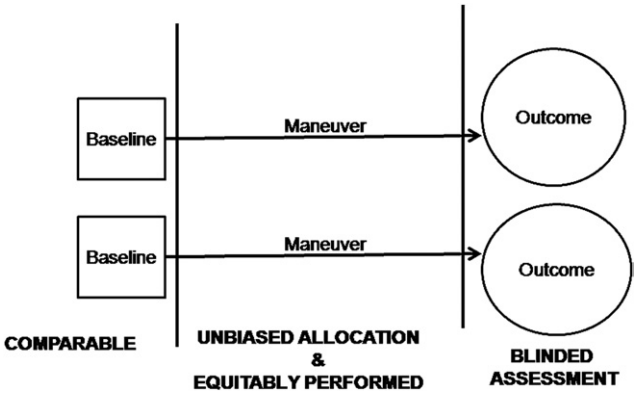

Figure 2 A graphic illustration of the primary components in a comparison study template. The *vertical lines* demarcate segments showing the fundamental requirements for quality.

It should be emphasized that the type of study relative to the question and the quality of the study, such as a well designed and conducted RCT, are the important issues here and not the statistical significance of the outcome. Because statistical significance is highly dependent upon the number in the study, to limit study selection to those that achieved statistical significance is to bias the results from the beginning. The whole point of a systematic review is to recognize that smaller studies not reaching significance may contain important facts. The objective of the review is to attempt to look at those facts and to synthesize the small sample data into a larger sample for analysis using meta-analysis techniques; or at least consider in detail the strengths and weaknesses of the studies as they may influence inferences about the hypothesis.

Participants within each study. The characteristics of the sample population within each study may vary greatly, such as degree of disease, age, sex, and confounding variables. When selecting the studies to review, the participant characteristics must be prospectively set. Depending upon the question and the hypothesis, these characteristics might need to be very tightly defined, or, conversely, may need to be extremely inclusive.

Interventions. Interventions, comparison interventions, risk factors, prognostic factors, exposures, etc, dependent on the question, often vary considerably among studies. It is useful to define these as broadly as possible; however, they must fit the hypothesis to be tested. During this phase of protocol development, practical considerations may require a restatement of the question/hypothesis that does not distort the fundamental question of interest.

Outcome(s). The focus of review is often the endpoint or outcome of interest. However, different studies may approach this using different outcome measures. When considering which studies to include, the outcomes part of the review protocol becomes crucial. Some questions require specific outcome assessment tool(s), and other questions are not so dependent upon the actual tool used. In prospectively setting this inclusion/exclusion criterion for the studies to review, the outcome measures must be consistent with the

question/hypothesis; sometimes conversion of results from different studies to a standard measure may be appropriate. For example, in determining the endpoint in treatments for Bell's palsy, the House-Brackmann (HB) facial grading scale might be used in one study, the Sunnybrook (SB) facial grading scale in another, and achieving normal facial movement without synkinesis may be used in another. In this circumstance, recovery to normal might be recorded for those with HB I, SB 100, and those determined to have achieved normal facial movement without synkinesis, and all three studies may be included in the meta-analysis.

Analysis. If meta-analysis is considered, the model of analysis should be stated in the protocol: fixed effect (note singular) and/or random effects (note plural) models (more on this later). The study weights, appropriate statistical tools, and conclusions from a meta-analysis depend upon the model selected. Both models can be used and the results compared.[2] If meta-analysis is not planned, an *a priori* plan as to how one plans to make cohesive sense from the review should be developed and included in the report. This can be quite complex and requires much thought prior to conducting the review. One method of structuring this plan is to approach it as though one were going to conduct a meta-analysis and then to explain what facts are available and what facts are missing. If progress is to be made, our knowing just how close or far we are from a scientifically informed decision is very important.

3. Develop Search Strategy

Sources. A major component of a systematic review is the identification of all relevant information as completely as possible. Dependent upon the intent of the review, the search strategy may include unpublished works and publications from all languages; this might be rational if the cost in dollars or human suffering would be significantly at stake if something were missed. However, most systematic reviews concentrate on published literature in one or a few languages. One step up from that is the attempted discovery of unpublished works that might be obtainable from known investigators in the field, abstracts, and conference proceedings. For simplicity, the discussion in this article will concentrate on published works.

In the current electronic era, computerized searches are the rule. It is expected that more than one search engine will be used in a thorough systematic review. There are multiple search engines; 32 are included in an incomplete list in the text by Portney and Watkins.[4] Additionally, hand searches for articles that are listed in the reference section of books and articles are usually considered appropriate due diligence. Often overlooked sources of articles are those excluded from previous systematic reviews of the topic; some of these might be appropriate to include. Another source of articles may be obtained by searching publications that seem off the beaten path in which some key articles discovered have been published, such as progress reports in the archives of agencies, or journals with which you might not be familiar, such as engineering or anthropology journals, to name a few. Some searches may include "grey literature,"[4] which refers to unpublished works or those available through noncustomary sources, such as working papers, theses, dissertations, and condition-specific associations. Consultation with a university or research organization–based research librarian is important to a well designed search strategy.

4. Conduct the Search

The search strategies used to mine the databases and other sources are fundamental to a successful search. Research librarians can be indispensible in this endeavor, but only if you, the author, sit with them as the strategies are developed and tested. You are the topic expert and they are the search experts; face-to-face collaboration is crucial for success. Specific search strategy chains should be developed, tested, edited, retested, and finalized prospectively. Time spent in this endeavor is well spent. One effective way to test and retest the strategies is to see if they efficiently detect all the papers you feel are important. The point of publishing these strategies is to assure the reader and other investigators that if they wished to corroborate your work, they could do so by using the same strategies. It also gives the reader some idea of the validity of the work. It might be reasonable to expect a published acknowledgement of the research librarian. An incomplete or haphazardly done search casts serious doubt on the work. In addition to the research librarian consultation, two reviewers usually independently search and combine the results for further discussion and perhaps further searching; the objective is to identify all potentially relevant papers for review. Subsequent levels of review may use the same or different reviewers.

5. Identify Relevant Papers (First Level of Review)

A list of studies identified in the search process is generated. The titles and abstracts are assessed by two reviewers to determine the relevance to the systematic review. The articles are separated into those to be included for further study and those to be excluded. If there are questions about the appropriateness of a study, these questionable papers are also included to pass to the next level of review.

6. Screen Papers (Second Level of Review)

The second level of review involves at least two reviewers for each relevant study being advanced beyond the first level of review. The complete article is briefly reviewed, and each reviewer must assure that the design *appears consistent* with the protocol of the systematic review. When there is a discrepancy between reviewers regarding the appropriateness of a study, the reviewers discuss their opinions and conflicts are resolved. The method of conflict resolution should be detailed prospectively in the protocol. Papers that pass this level of review move on to the third level of review.

7. Select Papers (Third Level of Review)

The third level of review also involves at least two reviewers and rigorously evaluates all the articles in detail to determine if the study subjects, interventions, risk factors, and outcomes are all *exactly consistent* with the protocol. These papers are selected for the systematic review. All previously excluded papers are kept in a separate file, to be discussed later in the review as to why they were excluded. A table of these excluded articles and reasons for exclusion can be helpful.

8. Evaluate Quality of Papers and Construct Data Extraction Forms

The selected studies are assessed for quality, and data are collected on a data extraction form. A minimum of two people independently perform this review; however, they initially collaborate and discuss the process in detail in one or more training sessions in which one or more articles are reviewed for parameters, criteria for evaluation, and the technique for completing the data extraction form. This training and the predetermined form assure that the reviewers collect the same information for easy review and analysis. An example data extraction form is available on the Cochrane website (http://www.cochrane-renal.org/docs/data_extraction_form.doc).

Biases that may affect the conclusion of the review are as follows: 1) publication bias (inherent in reviews based only on published studies that tend to overstate positive results; evidence of published negative studies helps reduce this concern), 2) selection bias (when study groups are differentially selected and are not equally comparable; randomization helps reduce this bias), 3) performance bias (this occurs if there are major differences in care among groups, other than the intervention being tested), 4) attrition bias (this may occur if there are more dropouts in one comparison group than the other), and 5) detection bias (this occurs if outcome assessments are different among comparison groups).[4]

Quality assessment of each study may be assessed by one of more than 25 published rating scales. The ratings for each study should be included in the systematic review report. Three rating scales discussed by Portney and Watkins[4] are the Jadad scale (instrument to measure the likelihood of bias),[7] the PEDro scale (physiotherapy evidence database scale),[8] and the QUADAS scale (quality assessment of diagnostic accuracy studies).[9] The very short and easy to use Jadad scale, which takes five minutes, compared very favorably with the classic, much more complex, scale of T. C. Chalmers, which takes about 60 minutes to complete.[10] The quality assessment of each article may be included in tabular format within the systematic review. This is usually the consensus between the two reviewers; however, testing inter-reviewer agreement and adding that to the systematic review report adds credibility. Agreement upon the quality assessment tool should be part of the protocol development. Figure 2 illustrates some general characteristics that are fundamental to the quality of a study. Groups to be compared must be similar, such that the major baseline variables within the groups are not different. Often, table 1 in most articles shows evidence that the variables are not statistically different. The interventions or maneuvers applied to the groups must be allocated in as unbiased a manner as possible; for example, randomly allocated. Interventions must also be equitably applied; for example, if testing operation A with operation B, operation A cannot be performed by a beginner and operation B performed by a seasoned surgeon. And finally, the assessment of outcomes must use the same assessment tool, and the assessor must be blinded to the intervention, and preferably so must the intervention be concealed from the subject; this is known as double blind or concealed.

9. Analyze and Synthesize Findings

Data extraction forms from the protocol help populate tables used at this stage. The fundamental components, illustrated in Figure 2, and others may vary among studies. It is likely that each paper will be somewhat different in the way the authors have assembled the potentially eligible subjects and have applied inclusion/exclusion criteria to obtain the baseline group(s) to be enrolled into their study. The interventions may vary among studies, and the outcome measures may differ. If the outcome assessments are not blinded, it may not be appropriate to include such a paper. This is an important issue that may require discussion if the paper is included or given as a reason if the paper is excluded. See Analysis in the Develop Protocol section above concerning prospectively designed plans for analysis.

Basic to a systematic review is one or more tables summarizing the important variables extracted in the review. The columns in such a table may include 1) author, 2) quality score, 3) number of subjects in total and within each arm, 4) important demographic baseline variables, such as sex counts and mean age $\pm$ standard deviation, 5) important disease classification/severity, 6) interventions, 7) outcome measures used, and 8) outcomes summary statements for each arm.[4] The specifics of the data collection columns in the tables vary depending upon the point of the study. For example, in Rosenfeld's systematic review of the natural history of untreated otitis media, Table 11-1 headings for "Spontaneous relief of AOM in children" include 1) author, 2) country, 3) age range, 4) diagnostic certainty, 5) symptomatic relief of pain and fever by days, and 6) clinical resolution (specifically defined in the legend).[11] The structure as well as headings of the tables may also vary among authors. For example, the tables in Shin et al's text on evidence-based otolaryngology are quite different than many publications.[12]

It is obvious from this that the initial design of the testable question to be answered by the review must consider all of these issues and practicalities. It is also obvious that some aggregates of studies for review will be very similar (i.e., homogenous), while others may vary consid-

erably in the way they approached the problem and the outcomes obtained, even to the point of showing major conflicts among studies (i.e., heterogeneous). It is this point exactly that both emphasizes the value of an unbiased, transparent systematic review of available information and determines the appropriateness of a mathematical/statistical synthesis of data, i.e., meta-analysis. The more the studies vary, the less appropriate is meta-analysis. However, more important is an objective analysis of the information in a narrative and tabular form, dissecting the individual and collective strengths and weaknesses of the papers and outlining what facts are needed in future works. It is just as important, perhaps more important, to conclude that there is insufficient evidence to answer the question. Because we are biased toward the idea that everything has an answer, finding no answer is difficult to admit, and more difficult to publish.

10. Determine if Meta-Analysis Is Appropriate and Select Model

Just as the studies may be heterogeneous in baseline variables and methods, the outcomes between studies may also be quite varied. Effect size of the outcomes (more on this later) is the central focus of determining if statistically synthesizing the various study data into a concise summary statement, meta-analysis, is feasible. Thus, qualitative logic and quantitative statistical tests of heterogeneity are factors determining whether meta-analysis is appropriate, and if so, under what model the results will be configured and analyzed.[2,4,13]

Synthesis of data in meta-analysis is conducted according to one of two statistical models (fixed effect model versus random effects model), depending upon the underlying assumptions. The fixed effect (singular) model assumes that there is one true effect size, thus a fixed or common effect, and all the actually observed effect sizes in the studies vary only because of sampling error in trying to find the one true effect size in a single universal parent population. The variance in this meta-analysis model is limited to within-study variance. The summary effect size in the meta-analysis is understood to be an estimate of this one true effect size.

The random effects (plural) model assumes that there are many true effect sizes and that each study is an estimate of one of these effect sizes in statistically different universal parent populations. The actual observed effect sizes represent a random sample of all of these underlying true effect sizes; each study has a different true effect size. Thus, the variance in the meta-analysis is both the within-study and between-study variances.

The model chosen is primarily dependent upon the objective of the review. If the objective is to determine the common effect size for the population studied, assuming all the studies are looking at the same universal parent population, then the fixed effect model is appropriate. However, generalizing to other populations is inappropriate. If the objective from the beginning is to generalize to a range of scenarios, then the random effects model might be better.[13] It may be possible to get some idea statistically of the between-studies variance; if this is great and statistically significant, then the assumption is that each study comes from a separate universal parent population, and a random effects model may be better. However, the objective of the meta-analysis may need to override the statistical tests for heterogeneity/homogeneity; there are several of these. To determine what model best serves the objectives and the available data requires consultation with a statistician and is far beyond the scope of this article. For simplicity, the examples and discussions in this article use the fixed effect model.

11. Report Results without Meta-Analysis when Data Are Insufficient for Meta-Analysis

Systematic reviews without meta-analysis should be reported when there are insufficient or excessively heterogeneous data. In such a report, a detailed description of the raw data, tables, and the reasons meta-analysis is inappropriate are important. When available, the various effect sizes are important to report; this can be very useful in planning future primary studies on the topic.

Effect Size

The effect size, also known as treatment effect, is a standardized estimate of the magnitude of the difference between groups, independent of sample size, and is fundamental to meta-analysis.[4,13-16] An effect size may be determined for either type of question, comparison or correlation, and includes the central tendency, like the mean, and spread of the data, like the standard deviation. Effect sizes may be calculated from continuous scale data, such as means or correlations, ordinal scale data, such as intensity grades, or binary scale data, such as risk ratios or odds ratios.[4,13] For meta-analysis, an effect size index, a statistic that uses a standardized value that is unit free and independent of the number of subjects, is set for each type of question so that the results of various studies can be synthesized.

Comparison Questions Effect Size Index

Effect size index in studies comparing two groups using continuous scales is the "standardized" difference between two means, such as Cohen's d, which may be calculated by the difference between means divided by the standard deviation of either group (or the pooled standard deviations of both groups, which is better). An easy online calculator by Becker may be used for these calculations.[15] Effect sizes also use ratios, such as risk ratios and odds ratios; again for simplicity, these will not be discussed.

Correlation Questions Effect Size Index

The effect size index in correlation studies is known as effect size correlation and is calculated as the correlation

Table 1
Individual study raw data

	Group A			Group B		
Studies	Mean	SD	n	Mean	SD	n
1	61	28	30	51	27	30
2	59	28	30	49	26	30
3	59	26	60	53	24	60

between the independent variable classification and the dependent variable individual scores. Using the SPSS (SPSS Inc, Chicago, IL) statistical program, CORR procedure, effect size correlations can be computed. An easy online calculator by Becker may be used for these calculations.[15]

Meta-Analysis

It is important to know that meta-analysis does not pool all the subjects from various studies into a single sample as though all these subjects were in a single large study. What it does do is use each study's effect size index as if it were an individual subject's data point, i.e., each study contributes its effect size to the overall data set of the review and thus this synthesis has a central tendency, spread of data, and confidence interval.[4] Tables 1 and 2 show an example meta-analysis synthesis of some real data from our laboratory.

For studies reporting continuous data, the effect size index may be Cohen's d or Hedges's g, defined as the difference between means divided by the standard deviation (SD) of one of the groups, or the pooled standard deviation of both groups.[13] Cohen's d, used in this paper, converts the difference between means of two groups into standard deviations. For example, a difference between means of 10.27 might seem fairly large; however, a calculated Cohen's d of only 0.38 between the two groups makes this difference seem rather small. Interpretation of Cohen's d is roughly d = 0.2 small, d = 0.5 medium, d = 0.8 large.[17]

For studies reporting proportions or ratios, such as absolute or relative risk ratios or odds ratios, the effect size index is the ratio, and the 95% confidence interval (95% CI) about effect size index is that which is calculated for the ratios per se. The 95% confidence interval about Cohen's d takes a bit more calculation. See Table 3 for the formulae used for individual studies and for meta-analysis synthesis into an overall summary effect size.

Weighting Effect Sizes

Because sample sizes and other variables vary considerably across studies, not all studies contribute equally to the synthesized overall effect size in meta-analysis. Therefore, it is necessary to weight the individual study effect size; note that the examples herein are those for a fixed effect model using continuous variables. One way to weight a study effect size is to simply use the number of subjects in each study. However, it is generally understood that using only the study sample size does not take into account important variances among studies. The usual practice is to weight each study effect size using the inverse variance method, so named because the weight of the study (w) is the inverse of the squared standard error (SE); thus, $w = 1/SE^2$ for each study effect size.[18] Because the standard error is a direct index of the precision of the parameter, in this case the effect size, the weighting reflects the precision of the study. See Table 3 for more explanation of weight. Because the basic currency of meta-analyses is effect size, when tabulating the systematic review results for meta-analysis, it is useful to record for each study 1) the single estimate of effect and 2) its standard error.[19]

Use of Effect Size Weighting in Meta-Analysis

Table 1 shows real raw data measuring facial recovery using the Sunnybrook Facial Grading Scale. Appreciable differences were not expected; however, these data are sufficient

Table 2
Fixed effect model calculations for synthesis

Study	Effect size (Y) Cohen's d	SE_d	95% CI d or M = ± 1.96×SE_d or 1.96×SE_M	Variance (V_Y)	Weight (W_i)	Relative weight % (W_i/SumW)	WY
1	0.358	0.260	− 0.152 to 0.868	0.0677	14.7635	25%	5.2853
2	0.383	0.261	− 0.128 to 0.894	0.0679	14.7299	25%	5.6416
3	0.207	0.183	−0.152 to 0.566	0.0335	29.8402	50%	6.1769
Sums					59.336	100%	17.1038
Summary effect size (M) (M = sumWY/sumW)	0.288		0.034 to 0.543				
Variance of M (V_M) V_M = 1/sumW				0.0169			
Standard error of M ($SE_M = \sqrt{[V_M]}$)		0.1298					

Table 3
Formulae relative to comparison of two independent groups with continuous variables using fixed effect model[13]

Individual studies		
Item	Descriptor	Formula
Effect size (Y_i) in an individual study = Cohen's d	Standardized mean difference in an individual study; d = difference between means/standard deviation	$d = \frac{\bar{X}_1 - \bar{X}_2}{S_{within}}$
Within groups pooled standard deviation (S_{within}) in an individual study	Rather than using one group standard deviation (SD), it is better to pool the standard deviations of the groups in an individual study (S_{within})	$S_{within} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$
Variance Y (V_{Yi}) here = variance of d (V_d) in an individual study	Variance is one measure of the variability of scores within a sample;[4] in this case the variability of d	$V_{Yi} = V_d = \frac{n_1+n_2}{n_1 * n_2} + \frac{d^2}{2(n_1+n_2)}$
Standard error of Y (SE_{Yi}) in an individual study; here = standard error of d (SE_d)	Standard error represents the standard deviation of the sampling distribution of means if the same thing were repeated many times; in this case, the sampling error of the standardized mean difference if theoretically infinitely sampled	$SE_{Yi} = \sqrt{V_{Yi}}$ or $SE_d = \sqrt{V_d}$
Lower limit of CI (LL_Y)	Lower limit of 95% confidence interval of effect size (Y) in an individual study	$LL_Y = \bar{Y} - 1.96(SE_Y)$
Upper limit of CI (UL_Y)	Upper limit of 95% confidence interval of effect size (Y) in an individual study	$UL_Y = \bar{Y} + 1.96(SE_Y)$
Weight of effect size (W_i) in an individual study	Using inverse variance model for weight in an individual study; V_{Yi} is within study variance	$W_i = \frac{1}{V_{Yi}}$ Also $W_i = \frac{1}{SE_{Yi}^2}$ remembering $SE_{Yi} = \sqrt{V_{Yi}}$: thus, if one squares both sides of the equation, one gets $V_{Yi} = SE_{Yi}^2$
Weighted effect size (W_iY_i) in an individual study	This product allows the eventual computation of the overall summary effect size that is reflective of the weights of the studies	$W_iY_i = W_i * Y_i$
Relative weights %	Each study weight divided by the sum of weights, reported as a percent	$\% = \frac{W_i}{\sum W_i} \times 100$
Meta-analysis synthesis		
Summary effect size (M)	The overall summary effect size in the meta-analysis (M) = sum of (effect sizes multiplied by weight) divided by sum of weights	$M = \frac{\sum W_iY_i}{\sum W_i}$
Variance of summary effect (V_M)	Variance of the summary effect using reciprocal of the sum of weights	$V_M = \frac{1}{\sum W_i}$
Standard error of summary effect (SE_M)	Square root of the summary effect variance	$SE_M = \sqrt{V_M}$

Table 3
(*continued*)

Individual studies		
Item	Descriptor	Formula
Lower limit of CI about M (LL_M)	Lower limit of 95% confidence interval of summary effect size	$LL_M = M - 1.96(SE_M)$
Upper limit of CI about M (UL_M)	Upper limit of 95% confidence interval of summary effect size	$UL_M = M + 1.96(SE_M)$

Y, generic symbol of effect size used in descriptions of meta-analysis, which is variously determined (in these examples, Y is calculated as Cohen's d); n_1, number of subjects in group one; n_2, number of subjects in group two; S_1, standard deviation (SD) in group one; S_2, standard deviation (SD) in group two; $*$, multiply; $\surd$, square root of everything that follows under the symbol; x^2, whatever x is squared; *subscript i*, individual study; *subscript M*, summary effect in the meta-analysis; n_1+n_2-2, generally represents the degrees of freedom; $\bar{Y}$, mean of assumed normal distribution of Y's.
Small s usually represents standard deviation; however, in these formulae by Borenstein et al,[13] capital S is used for standard deviation.

to illustrate an effect size index and weighting for continuous variables and to illustrate the importance of the spread of the data, not just the difference between means. These data also show the relative contribution each trial adds to the overall results and ultimately what the final meta-analysis shows. Using the raw data from Table 1, the calculated results in Table 2, and the formulae in Table 3, one can get a rather clear picture of what meta-analysis is and how the process works. Interestingly, these data were not expected to show a statistically significant difference, and individually the trials did not show a difference; however, the synthesis of these results did show a statistically significant difference.

Forest Plot

A very nice two-page review of the origins, name, and evolutionary use of forest plots is available.[20] Forest plots, so named because they look like a forest of lines, are the classic way meta-analyses are currently presented (Fig 3). The vertical line represents the no-difference line, zero for continuous variables and usually 1 for proportions. The horizontal lines demarcate the upper and lower limits of the confidence intervals. The squares illustrate the relative weight of each study, which is related to the number (n) in the study; but, as discussed above, it is not the number in the study per se. Thus, a forest plot of the data in Table 2 would have the square in trial 3 approximately twice the size of trials 1 and 2; in this case, the number in the study versus the weight of the study does not differ that much in choosing the size of the squares. The overall summary effect size and confidence interval are illustrated by a diamond.

Sensitivity Analysis

Because studies usually have differing numbers of subjects, often begin with sample populations with differing baseline variables, have somewhat varied interventions, and their outcomes may be measured differently, sensitivity analysis is used to see if the conclusion of the meta-analysis will change if some studies are removed. If the overall conclusions do not change with the elimination of one or more studies, the meta-analysis is considered robust. If, on the other hand, the conclusions change significantly, the meta-analysis is much less robust, and the reasons why this is the case become very important to explore. In selecting studies to exclude, the numbers in the study are obvious places to start; for example, excluding the largest study. Other important sources of variability among studies for sensitivity analysis are the inclusion/exclusion criteria or methods. Testing the analysis along these lines can be revealing. For example, if studies were included in which outcome assessments were not blinded, i.e., the allocation of interventions were inadequately concealed from those that might influence the outcome recorded, such as the subject, and especially those measuring the outcome, there could be a major

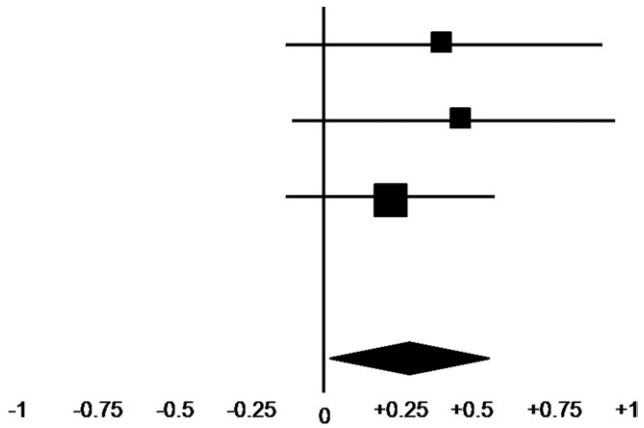

Figure 3 Illustration of a forest plot of data in text tables. The sizes of *squares* are proportional to the weight of each study, and the confidence intervals are indicated by the *horizontal lines*. The *diamond* represents the summary effect size value of the overall meta-analysis. The *vertical bar* demarcates no difference between interventional arms.

difference between blinded and unblinded studies, which is usually the case.

Quality Review of Meta-Analyses

Oxman et al[21] developed a checklist for the assessment of the quality of systematic reviews and meta-analyses; this list was again reemphasized in a recent excellent text on the applications of clinical research to practice.[4] The checklist includes the following major points for evaluation: 1) Was the question clear? 2) Were inclusion/exclusion criteria specified and appropriate? 3) Was the validity of included studies assessed by the reviewers? 4) Was it reasonable to perform a meta-analysis? 5) Were the overall results clearly expressed, and were specific effect sizes included? 6) Were point estimates and confidence intervals reported? 7) Can the results be applied to practice with assurances of benefit/harm/cost?

Summary

A systematic review is a transparent and unbiased review of available information. The published systematic review must report the details of the conduct of the review as one might report the details of a primary research project. A meta-analysis is a powerful and rigorous statistical approach to synthesize data from multiple studies, preferably obtained from a systematic review, in order to enlarge the sample size from smaller studies to test the original hypothesis and/or to generate new ones.

For those interested in studying systematic reviews and meta-analysis further, two books, *Meta-analysis, Decision Analysis, and Cost-effectiveness Analysis, Methods for Quantitative Synthesis in Medicine,* 2nd edition,[2] and *Introduction to Meta-analysis*,[13] and two computer programs, Comprehensive Meta-analysis Version 2,[16] and Review Manager (RevMan) Version 5,[22] are excellent sources for information and practice. If the reader plans to perform a systematic review and meta-analysis, consultation with a research librarian and a statistician familiar with meta-analytical techniques is recommended.

Author Information

From the Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine (Drs Neely, Rich, Voelker, Wang, Paniello, Nussenbaum, and Bradley), St. Louis, MO; and the Division of Otolaryngology-Head and Neck Surgery, Pediatric Otolaryngology, University of California at San Diego School of Medicine (Dr Magit), San Diego, CA.

Corresponding author: J. Gail Neely, MD, Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, 660 S. Euclid Avenue, Box 8115, St. Louis, MO 63110.

E-mail address: neelyg@ent.wustl.com.

Author Contributions

J. Gail Neely, primary author; **Anthony E. Magit,** primary author, reader; **Jason T. Rich,** secondary author, reader; **Courtney C. J. Voelker,** secondary author, reader; **Eric W. Wang,** secondary author, reader; **Randal C. Paniello,** secondary author, reader; **Brian Nussenbaum,** secondary author, reader; **Joseph P. Bradley,** secondary author, reader.

Disclosures

Competing interests: None.

Sponsorships: None.

References

1. Feinstein AR. Principles of medical statistics. New York: Chapman & Hall/CRC; 2002.
2. Petitti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis. Methods for quantitative synthesis in medicine, 2nd ed. New York: Oxford University Press; 2000.
3. Rudmik LR, Walen SG, Dixon E, et al. Evaluation of meta-analyses in the otolaryngological literature. Otolaryngol Head Neck Surg 2008;139:187–94.
4. Portney LG, Watkins MP. Foundations of clinical research: applications to practice, 3rd ed. Upper Saddle River (NJ): Pearson Prentice Hall; 2009.
5. Sackett DL, Richardson WS, Rosenberg W, et al. Evidence-based medicine. How to practice and teach EBM. New York: Churchill Livingstone; 1997.
6. Centre for Evidence Based Medicine. Oxford Centre for Evidence-based medicine: levels of evidence (March 2009). Available at: http://www.cebm.net/index.aspx?o=1025. Accessed July 18, 2009.
7. 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.
8. Physiotherapy Evidence Database. PEDro scale (last modified March, 1999). Available at: http://www.pedro.org.au/scale_item.html. Accessed July 18, 2009.
9. Whiting P, Rutjes AA, Dinnes J, et al. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess 2004;8:1–234.
10. Ohlsson A, Lacy JB. Poster presentation: Quality assessments of randomized controlled trials: an evaluation by the Chalmers versus the Jadad method (Poster 4). In: Cochrane Colloquium at Oslo, Scientific presentations and Posters. Collaboration C, editor. Oslo, Norway; 1995.
11. Rosenfeld RM. Natural history of untreated otitis media. In: Rosenfeld RM, Bluestone CD, editors. Evidence-based otitis media. Saint Louis: B.C. Decker, Inc; 1999. p. 157–77.
12. Shin JJ, Hartnick CJ, Randolph GW. Evidence-based otolaryngology. New York: Springer Science+Business Media; 2008.
13. Borenstein M, Hedges L, Higgins J, et al. Introduction to meta-analysis. Chichester, West Sussex: John Wiley & Sons; 2009.
14. Becker LA. Lecture notes of effect size. Available at: http://web.uccs.edu/lbecker/Psy590/es.htm. Accessed July 18, 2009.
15. Becker LA. Effect size calculators. http://web.uccs.edu/lbecker/Psy590/escalc3.htm. Accessed July 18, 2009.
16. Borenstein M, Hedges L, Higgins J, et al. Comprehensive meta-analysis version 2. Englewood (NJ): Biostat; 2005.
17. Cohen J. Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale (NJ): L. Erlbaum Associates; 1988. p. xxi, p. 567.
18. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions version 5.0.1 [updated September 2008]. Oxford: Wiley-Blackwell; 2008.
19. Wilson DB. Overview of meta-analytic data analysis: lecture notes overheads. In: Manassas, VA: David B. Wilson. George Mason University, Administration of Justice Department; 1996.
20. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001;322:1479–80.
21. Oxman AD, Cook D, Guyatt G. Users' guide to the medical literature: VI. How to use an overview. JAMA 1994;272:1367–71.
22. Review Manager (RevMan) [computer program]. Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration; 2008.