Chapter 10
Meta-analysis, Evidence-Based Medicine, and Clinical Guidelines

Stephen P. Glasser and Sue Duval

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

R.A. Fisher. Presidential Address by Professor R.A. Fisher,
Sc.D., F.R.S. Sankhyā: The Indian Journal of Statistics
(1933–1960), Vol. 4, No. 1 (1938), pp. 14–17.
http://www.jstor.org/stable/40383882.

Abstract Meta-analysis refers to methods for the systematic review of a set of individual studies (either from the aggregate data or the individual patient data) with the aim to quantitatively combine their results. This has become a popular approach to attempt to answer questions when the results from individual studies have not been definitive. This chapter will discuss meta-analyses and highlight issues that need critical assessment before the results of the meta-analysis are accepted. Some of these critical issues include: publication bias, sampling bias, and study heterogeneity. Evidence-based medicine and clinical practice guidelines are dependent upon meta-analyses to guide their recommendations. Evidence-based medicine is an apt term to the extent that it advocates more reliance on clinical research than on personal experience or intuition; and, has led to a paradigm outlining the "level of evidence" that addresses a particular clinical question (also see Chap. 3). These "levels of evidence" are also utilized by clinical practice guidelines, but "as the number of available guidelines provided by a variety of sources has literally exploded, serious questions and controversies have arisen about how guidelines should be developed, implemented, and evaluated."

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT 638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S. Duval, Ph.D.
Cardiovascular Division, University of Minnesota Medical School, Minneapolis, MN, USA
e-mail: sueduval@umn.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_10,

Keywords Meta-analysis • Publication bias • Systematic review • Evidence-based medicine • Clinical practice guidelines

Introduction

Meta- is from Latin meaning among, with, or after; occurring in succession to, situated behind or beyond, more comprehensive, or transcending. This has led some to question if meta-analysis is to analysis as metaphysics is to physics (metaphysics refers to the abstract or supernatural), to which a number of article titles would attest, such as: "is a meta-analysis science or religion?" [1]; "have meta-analyses become a tool or a weapon?" [2]; "meta-statistics: help or hinderance?" [3] and, "have you ever meta-analysis you didn't like?" [4], or as Bangalore put it "a meta-analysis is like a sausage. God and the butcher know what goes in it and neither would ever eat any" [5]. Overviews, systematic reviews, pooled analyses, quantitative reviews and quantitative analyses are other terms that have been used synonymously with meta-analysis, but some distinguish between them. For example, pooled analyses might not necessarily use the true meta-analytic statistical methods, and quantitative reviews might similarly be different than a meta-analysis. Compared to traditional reviews, meta-analyses are often more narrowly focused, usually examine one clinical question, and necessarily have a strong quantitative component. Meta-analysis can be literature based and these are essentially, studies of studies. Said simply, meta-analysis is the statistical combination of two or more separate studies, with the potential advantages being improved precision and increased power. The majority of meta-analyses rely on published reports, however more recently, meta-analyses of individual patient (participant) data (IPD) have appeared.

The earliest meta-analysis may have been that of Karl Pearson in 1904, which he applied in an attempt to overcome the problem of reduced statistical power in studies with small sample sizes [6]. The first meta-analysis of medical treatment is probably that of Henry K Beecher on the powerful effects of placebo, published in 1955 [7]. But, the term meta-analysis is credited to Gene Glass in 1976 [8]. Only four meta-analyses could be found before 1970, 13 were published in the 1970s and fewer than 100 in the 1980s. Since the 1980s more than 10,000 meta-analyses have been published. Why this popularity of meta-analysis, and why do a meta-analysis in the first place? Individual studies attempt to make inferences by setting up experimental contrasts that pertain to the hypothesis at hand. Nevertheless, observed findings are subject to random variation that could lead the inference astray, and it is also difficult to test the consistency of findings across a variety of settings from a single study. The goal of a meta-analysis is to enhance inference by increasing power and by assessing the consistency of findings across studies; and in so doing one can more appreciate the degree of uncertainty in the research question, and the degree of heterogeneity between studies.

Definition

Meta-analysis refers to methods for the systematic review of a set of individual studies or patients (subjects) within each study, with the aim to quantitatively combine their results. Meta-analysis has become popular for many reasons, some of which include**:**

- The adoption of evidence-based medicine which requires that all reliable information is considered
- The desire to avoid narrative reviews which are often misleading or inconclusive
- The desire to interpret the large number of studies that have been conducted about a specific intervention
- The desire to increase the statistical power of the results by combining many smaller sized studies

Some definitions of a meta-analysis include:

- An observational study in which the units of observation are individual trial results or the combined results of individual patients (subjects) aggregated from those trials.
- A scientific review of original studies in a specific area aimed at statistically combining the separate results into a single estimate
- A type of literature review that is quantitative
- A statistical analysis involving data from two or more trials of the same treatment and performed for the purpose of drawing a global conclusion concerning the safety and efficacy of that treatment

One should view the steps in designing a meta-analysis the same way as one views the steps take in designing a clinical trial (unless one is performing an exploratory meta-analysis), except that most meta-analyses are retrospective and observational. Beyond that, a meta-analysis is like a clinical trial except that the units of observation may be individual subjects within trials, or individual trial results. Thus, all the considerations given to the strengths and limitations of clinical trials should be applied to meta-analyses (e.g. a clearly stated hypothesis, a predefined protocol, considerations regarding selection bias, etc.).

The reasons one performs a meta-analysis is to 'force' one to review all pertinent evidence, to provide quantitative summaries, to integrate results across studies, and to provide for an overall interpretation of these studies. This allows for a more rigorous review of the literature, and it increases sample size and thereby potentially enhances statistical power. That is to say, the primary aim of a meta-analysis is to provide a more precise estimate of an outcome (say a medical therapy in reducing mortality or morbidity) based upon a weighted average of the results from the studies included in the meta-analysis (Table 10.1). The concept of a 'weighted average' is an important one. In the most basic approach, the weight given to each study is the inverse of the variance of the effect; that is, on average, the smaller the variance, and the larger the study, the greater the weight one places on the results of that study. Because the results from different studies investigating different but hopefully similar questions are often measured on different scales, the dependent variable in a

Table 10.1 Some reasons to perform a meta-analysis

"Force" a rigorous literature review
Resolve uncertainty when reports disagree
Increase sample size
Enhance statistical significance of subgroup analyses
Enhance scientific credibility of some observations
May identify new research directions
May help put into focus a controversial study
Provide more precise effect size estimates
Allow one to assess variability between studies
Increase statistical power
May identify characteristics associated with particularly effective treatments
Allow for study of heterogeneity

Table 10.2 Comparison of expert reviews vs. meta-analysis

	Expert review	Meta-analysis
Question	Broad	Focused
Sources	Often not specified	Comprehensive
Search	Ad-hoc	Explicit
Selection	Often not specified	Criterion-based
Appraisal	Variable	Rigorous
Synthesis	Usually qualitative	Qualitative or quantitative
Inference	Sometimes evidence-based	Usually evidence-based

meta-analysis is typically some standardized measure of effect size. In addition, meta-analyses may enhance the statistical significance of subgroup analysis, and enhance the scientific credibility of certain observations.

Finally, meta-analyses may identify new research directions or help put into focus the results of a controversial study. As such, meta-analyses may resolve uncertainty when reports disagree, improve estimates of effect size, and answer questions that were not posed at the start of individual trials, but are now suggested by the trial results. Thus, when the results from several studies disagree with regard to the magnitude or direction of effect, or when sample sizes of individual studies are too small to detect an effect, or when a large trial is too costly and/or too time consuming to perform, a meta-analysis should be considered.

One should make the distinction between meta-analysis, a systematic review, and an expert review. Meta-analysis is quantitative and employs statistical methods to combine and summarize the results of several studies; a systematic review is the process for searching the literature appropriately, in order to find the relevant information. Expert reviews are broad and frequently biased summaries by a leading authority in a given field (Table 10.2).

Weaknesses

As is true for any analytical technique, meta-analyses have weaknesses. For example, they are sometimes viewed as more authoritative than is justified. After all, meta-analyses are retrospective repeat analyses of prior published data. Rather, meta-analyses should be viewed as nearly equivalent (if performed properly under rigid study design characteristics) to a large, multi-center study. In fact, meta-analyses are really studies in which the 'observations' are not under the control of the meta-investigator (because they have already been performed by the investigators of the original studies); the included studies have not been obtained through a randomized and blinded technique; and, one must assume that the original studies have certain statistical properties they may not, in fact, have. In addition, one must rely only on reported rather than directly observed values, unless an IPD meta-analysis is undertaken.

There are at least nine important considerations in performing or reading a meta-analysis (Table 10.3):

1. They are sometimes performed to confirm an observed trend (this is equivalent to testing before hypothesis generation)
2. The sample of studies included in a meta-analysis may not be representative
3. Publication bias
4. Difficulty in pooling across different study designs
5. Dissimilarities of control treatment
6. Differences in the outcome variables
7. Studies are reported in different formats with different information available
8. The issues surrounding the choice of fixed versus random modeling
9. Alternative modeling

Table 10.3 At least nine considerations when performing or reading a meta-analysis

Is it being done to confirm observed trends?
Pooling across studies is difficult
Sample bias
Publication bias
Control treatment dissimilarities
Differences in primary and secondary outcomes across studies
Differences in reporting outcomes
Weighting
Modeling

Meta-analyses are Sometimes Performed to Confirm Observed Trends (i.e. Testing Before Hypothesis Generation)

Frequently in meta-analyses, the conduct of the analysis is to confirm observed 'trends' in sets of studies; and, this is equivalent to examining data to select which statistical analyses should be performed, rather than the reverse. This is well known to introduce spurious findings. It is important to be hypothesis driven i.e. to perform planning steps in the correct order (if possible).

In planning the meta-analysis, the same principles apply as planning any other study. That is, one forms a hypothesis, defines inclusion and exclusion criteria, collects data, tests the hypothesis, and reports the results. But, as previously mentioned, just like other hypothesis testing, the key is to avoid spurious findings by keeping these steps in the correct order, and this is sometimes NOT the case for meta-analyses. For example, frequently the 'trend' in the data is already known; in fact, most meta-analyses are performed because of a suggestive trend. In Petitti's steps in planning a meta-analysis she suggests first addressing the objectives (i.e. state the main objectives, specify secondary objectives); perform a review; information retrieval; specify MEDLINE search criteria; and explain approaches to capture 'fugitive' reports (those not listed in MEDLINE or other search engines and therefore not readily available) [9].

The Sample of Studies Included in a Meta-analysis May Not Be Representative

As with sampling in clinical trials identifying studies to be considered for inclusion is in essence, defining the 'sampling frame' for the meta-analysis. The overall goal is to include all pertinent studies; and, several approaches are possible. One approach could be: 'I am familiar with the literature and will include the important studies'. With this approach, there may be a tendency to be aware of only certain types of studies and selection will therefore be biased. A more scientific and valid approach is where one uses well-defined criteria for inclusion and exclusion applying an objective screening (search) tool such as MEDLINE. Clearly defined keywords and MESH terms, clearly defined years of interest, and a transparent description of what the meta-investigator did must be included in any report. Also, the impact of the 'Search Engine' on identifying papers must be adequately reported to allow for study replication. Surprising to some is that there may be problems with using MEDLINE alone to screen for articles. Other searches can be done with EMBASE or PUBMED and seeking the help of a trained Biomedical Librarian is generally advisable. In addition, not all journals are included in these search engines and there is dependence on keywords assigned by authors and MESH terms by Medline indexers [10]. Further, searches may not include fugitive or grey literature, government reports, book chapters, proceedings of conferences, published dissertations, etc.

As previously stated, the included studies in a meta-analysis have not been obtained through a randomized and blinded technique, so that selection bias becomes an issue. Selection bias occurs because studies are 'preferentially' included and excluded and these decisions are influenced by the meta-investigators prior beliefs as well as the fact that studies are included based upon recognized 'authorities'. That is, investigator bias occurs because the investigators who conducted the individual studies included in the meta-analysis may have introduced their own bias.

It is necessary for a complete meta-analysis to go to supplemental sources for studies, such as studies of which authors are personally aware, studies referenced in articles retrieved by Search Engines, and searches of Dissertation Abstracts to name a few. The biggest limitation, however, is how to search for unpublished and unreported studies. This latter issue is clearly the most challenging (impossible?), and opens the possibility for publication bias and the "file-drawer" problem.

Publication Bias (and the File-Drawer Problem)

Publication bias is one of the major limitations of meta-analysis as it derives from the fact that for the most part, studies that are published have positive results, so that negative studies are underrepresented and if published take longer to appear in the literature ("pipeline effect"). Stated another way, publication bias results from the selective publication of studies based on the direction and magnitude of their results. As an example, Turner et al. found that 17 % of 24 FDA registered trials were unpublished and 3 of 4 of the unpublished trials failed to show benefit over placebo [11].

The pooling of results of published studies alone can lead to an overestimation of the effectiveness of the intervention, and the magnitude of this bias tends to be greater for observational studies compared to RCTs. In fact, positive studies are three times more likely to be published than negative ones and this ratio is even greater for observational studies. Thus, investigators tend not to submit negative studies (this is frequently referred to as the 'file-drawer' problem), journals do not publish negative studies as readily, funding sources may discourage publication of negative studies, negative studies that do get published are published in lower impact journals some of which might not be indexed in Medline or other databases. One also has to be wary of overrepresentation of positive studies because duplicate publication can occur. The scenario resulting in publication bias goes something like this: one thinks of an exciting hypothesis, examines the possibility in existing data, if significant, the findings are published, but if non-significant the investigator loses interest and buries the results (i.e. puts them in a file drawer). Even if one is 'honorable' and attempts to publish a non-significant study, often the editor/reviewer will bury the result for you, since negative results are difficult to publish. One then continues on to the next idea and forgets that the analysis was ever performed. The obvious result of this is that the literature is more likely to include mostly positive findings and thereby is biased toward benefit. Publication bias is equivalent to performing a screen to select patients who only respond positively to a treatment before performing a clinical trial to examine the efficacy of that treatment.

To moderate the impact of publication bias, one attempts to obtain all published and unpublished data on the question at hand. There are also tests for the presence of publication bias, and methods to estimate the impact of publication bias and adjust for it. It should be noted that publication bias is a greater problem in epidemiological studies than clinical trials, since it is difficult to perform a major RCT and not publish the results even if negative, while for epidemiologic studies negative results are much less likely to be published.

As mentioned, there are ways that one can determine the likelihood that publication bias is influencing the meta-analysis. One of the simplest methods is to construct a funnel plot, which is a scatter plot of individual study effects against a measure of precision within each study. In the absence of bias, the funnel plot should depict an inverted 'funnel' shape centered about the true overall mean which the meta-analysis is trying to estimate. This is because we expect a wider spread of effects among the smaller studies. If the funnel appears truncated, it is likely that a group of studies is missing from the analysis set. It should be kept in mind however that publication bias is but one potential reason for this 'funnel plot asymmetry', and for this reason, current practice is to consider other mechanisms for the missing studies, such as English language bias, clinical heterogeneity, and location bias to name a few [12].

There are a number of relatively simple quantitative methods for detecting publication bias in the literature, including the rank correlation test of Begg and the regression-based test of Egger et al. [13, 14]. The Trim and Fill method can be used to estimate the number of missing studies and to provide an estimate of the treatment effect after adjustment for this bias [15]. The mechanics of this approach are displayed in Fig. 10.1a, using a meta-analysis of the effect of gangliosides and mortality from acute ischemic stroke [16]. Although in this example, the effect size is not great, the striking aspect of the plot is that it appears that there are no negative effects of therapy. The question is whether that observation is true or if this is an example of publication bias where the negative studies are not represented. Figure 10.1b shows what happens when the asymmetric studies are 'trimmed' to generate a symmetric plot to allow estimation of the true pooled effect (in this example, the five rightmost studies are trimmed). These trimmed studies are then returned, along with their imputed or 'filled' symmetric counterparts. An adjusted pooled estimate and corresponding confidence interval are then calculated based on the now presumed complete dataset (bottom panel). The authors of this method stress that the main goal of such an analysis is to allow a 'what if' approach; that is, to allow sensitivity analyses to the missing studies, rather than actually finding the values of those studies per se. Heterogeneity, reporting bias, and chance may all lead to asymmetry or other shapes in funnel plots (box). Funnel plot asymmetry may also be an artifact of the choice of statistics being plotted. Reporting biases arise when the dissemination of research findings is influenced by the nature and direction of results. As noted by Sterne et al. [12], positive studies are more likely to be published, published

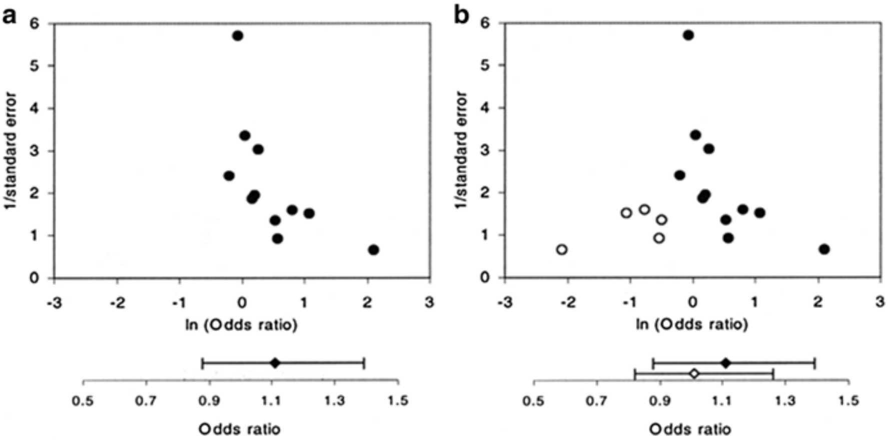

Fig. 10.1 (**a**) A of the studies included in the meta-analysis. (**b**) Filled "presumed" negative studies shown as unfilled circles, with the adjusted odds ratio calculated

rapidly, published in English, published more than once, published in high impact journals, and cited by others; while negative studies may be filtered, manipulated, or presented in such a way that they become positive. Reporting biases can have three types of consequence for a meta-analysis:

- A systematic review may fail to locate an eligible study because all information about it is suppressed or hard to find (publication bias)
- A located study may not provide usable data for the outcome of interest because the study authors did not consider the result sufficiently interesting (selective outcome reporting)
- A located study may provide biased results for some outcome—for example, by presenting the result with the smallest P value or largest effect estimate after trying several analysis methods (selective analysis reporting).

These biases may cause funnel plot asymmetry if statistically significant results suggesting a beneficial effect are more likely to be published than non-significant results. Such asymmetry may be exaggerated if there is a further tendency for smaller studies to be more prone to selective suppression of results than larger studies. This is often assumed to be the case for randomized trials. For instance, it is probably more difficult to make a large study disappear without a trace, while a small study can easily be lost in a file drawer. The same may apply to specific outcomes. For example, it is difficult not to report on mortality or myocardial infarction if these are outcomes of a large study. Smaller studies have more sampling error in their effect estimates. Thus even though the risk of a false positive significant finding is the same, multiple analyses are more likely to yield a large effect estimate that may seem worth publishing. However, biases may not act this way in real life; funnel plots could be symmetrical even in the presence of publication bias or selective outcome reporting for example, if the published findings point to effects in different

Table 10.4 Publication of studies based upon positive vs. negative results

Publication status of studies reviewed by the Central Oxford Research Ethics Committee (1984–1987)

	Statistically significant	Statistically non-significant
% Published	60	35
% Only presented	45	22
% Neither	15	42

Adapted from Easterbrook [17]

Table 10.5 Magnitude of effect size in published vs. registered studies

Meta-analysis of published vs. registered studies of treatment with alkylating agents for advanced ovarian cancer

	Published studies (n = 16)	Registered studies (n = 13)
Survival ratio	1.16	1.06
95 % CI	1.06–1.27	0.97–1.15
P-Value	0.02	0.24

Adapted from Easterbrook [17]

directions but unreported results indicate neither direction. Alternatively, bias may have affected few studies and therefore not cause glaring asymmetry.

Perhaps the best approach to avoid publication bias is to have a registry of all trials at their inception, that is, before results are available, thereby eliminating the possibility that the study results would influence inclusion into the meta-analysis. After a period of apathy, this concept is taking hold and a website (*clinicaltrials.gov*) is now available. But, to emphasize the importance of this, Table 10.4 points out an example of the publication status of studies that were statistically significant vs. those that were not; and Table 10.5 emphasizes the magnitude of outcome bias seen in this set of published vs. registered studies.

The effect of publication bias on meta-analytical outcomes was demonstrated by Glass et al. in 1979 [18]. They reported on 12 meta-analyses, and in every instance where it could be determined, found that the average experimental effect from studies published in journals was larger than the corresponding effect estimated from unpublished work (mostly from theses and dissertations), accounting for almost a 33 % bias in favour of the benefit. As a result, some have suggested that a complete meta-analysis should include attempts to contact experts in the field as well as authors of referenced articles for access to unpublished data. More recent estimates have suggested that the effect of publication bias accounts for 5–15 % in favour of benefit.

Some literature that is available but hard to find includes grey and fugitive literature. Grey literature refers to a body of materials that cannot be found easily through conventional channels, “but which is frequently original and usually recent”. The “Grey Information Functional Plan,” defines grey literature as foreign or domestic open source material that usually is available through specialized channels and may not enter normal channels or systems of publication, distribution, bibliographic control, or acquisition by booksellers or subscription agents. Examples of grey literature include technical reports from government agencies or scientific research groups, working papers from research groups or committees, and white papers. But, the identification and acquisition of grey literature poses difficulties for librarians and other information professionals for several reasons. Generally, grey literature lacks strict bibliographic control, meaning that basic information such as author, publication date or publishing body may not be easily discerned. Similarly, non-professional layouts and formats and low print runs of grey literature make the organized collection of such publications challenging compared to more traditional published media such as journals and books. Fugitive literature is literally the ones for which you have to hunt. On the World Wide Web, it is not always easy to hunt for specific information, particularly if you do not know where to begin. The following provides a partial list of websites that provide entry points for searching fugitive literature:

- http://www.google.com Meta search engine that searches across other engines
- http://www.healthfinder.gov Healthcare information from the USDHHS
- http://www.guidelines.gov/index.asp Summary Guidelines info from the AHRQ
- http://www.cdc.gov Healthcare information from the CDC

The Difficulty in Pooling Across a Set of Individual Studies and Heterogeneity

One of the reasons that it is difficult to pool studies is selection bias. Selection bias occurs because studies are 'preferentially' included and excluded and these are influenced by the meta-investigators prior beliefs as well as the fact that studies included are based upon recognized 'authorities'. That is, this type of bias occurs because the investigators who conducted the individual studies included in the meta-analysis may have introduced their own bias. In addition, there is always a certain level of heterogeneity of study characteristics included in a given meta-analysis so that as the cliché goes 'by mixing apples and oranges with an occasional lemon, ones ends up with an artificial product.' Glass argued this point rather eloquently as follows:

> '*…Of course it mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavor worthy of true scientists; comparing apples to apples is trivial.…*'
>
> *The same persons arguing that no two studies should be compared unless they were studies of the 'same thing' are blithely comparing persons within studies i.e. no two things can be compared unless they are the same…but if they are the same then they are not two things.'* Glass went on to use the classic paradox of Theseus's ship, which set sail on a 5-year journey. After nearly 5 years, every plank had been replaced. The question then i*s 'are Theseus and his men still sailing the ship that was launched 5 years earlier? What if as each plank was removed, it was taken ashore and repositioned exactly as it had been on the waters so that at the end of 5 years, there exists a ship on shore, every plank of which once stood exactly as it had been 5 years before. Is this new ship Theseus's ship, or is it the one still sailing?' The answer depends on what we understand the concept of 'same' to mean.*

Table 10.6 Some causes of heterogeneity

Some causes of heterogeneity
Differences in inclusion/exclusion criteria of the individual studies
Different control or treatment interventions (dose, timing, brand), outcome measures and definition, and different follow-up times
The reasons for withdrawals, drop-outs, cross-overs will likely differ between individual studies, as will the baseline status of the patients and the settings
The quality of the study design and its execution will likely differ

> *Glass goes on to consider the problem of the persistence of personal identity when he asks the question 'how do I know that I am the same person who I was yesterday, or last year…?'*
>
> Glass notes that probably there are no cells that are in common between the current organism called Gene Glass and the organism 40 years ago by the same name [19].

Recall that a number of possible outcomes and interpretations of clinical trials is possible. When one trial is performed, the outcome may be significant, and one concludes that a treatment is beneficial, or the results may be inconclusive leading one to say that there is not convincing statistical evidence to support a treatment benefit. But when multiple trials are performed other considerations present themselves. For example, when 'most' studies are significant and in the same direction one can conclude a treatment is beneficial, but when 'most' studies are significant in different directions one might question whether there are differences in the population studied or methods performed that warrant further consideration. The question that may then be raised is 'Could we learn anything by combining the studies?' It is this latter question that is the underlying basis for meta-analysis. Thus, when there is some treatment or exposure under consideration we assume that there is a 'true' treatment effect that is shared by all studies, and that the average has lower variance than the data themselves. We then consider each of the individual studies as one data point in a 'mega-study' and presume that the best (most precise) estimate of this 'true' treatment effect is provided by 'averaging' across studies. But, when is it even reasonable to combine studies? The answer to this latter question is that studies must share characteristics, including similar 'experimental' treatment or exposure, similar 'standard' treatment or lack of exposure, similar follow-up protocol, outcome(s) and patient populations. It is difficult to pool across different studies, even when there is an apparent similarity of treatments. This leads to heterogeneity when one performs any meta-analysis. The causes of study heterogeneity are numerous. Some of them are (Table 10.6):

– Differences in inclusion/exclusion criteria of the individual studies comprising the meta-analysis
– Different control or treatment interventions [dose, timing, brand], outcome measures and definition, and different follow-up times were likely to be present in each individual study

- The reasons for withdrawals, drop-outs, cross-overs will likely differ between individual studies, as will the baseline status of the patients and the settings for each study.
- Finally, the quality of the study design and its execution will likely differ

Heterogeneity of the studies included in the meta-analysis can be tested. For example, Cochran's Q is a test of homogeneity that evaluates the extent to which differences among the results of individual studies are greater than one would expect if all studies were measuring the same underlying effect and the observed differences between them were due only to chance. A measure of the proportion of variation in individual study estimates that is due to heterogeneity rather than sampling error, (known as I^2), is available and is the preferred method of describing heterogeneity [20]. This index does not depend on the number of studies, the type of outcome data or the choice of treatment effect. I^2 is related to Cochran's Q statistic and lies between 0 and 100 %, making it useful for comparison across meta-analyses. Most reviewers consider that an I^2 greater than 50 % indicates heterogeneity between the component studies. Rather sensitivity analysis to differences in study quality is more common. Sensitivity analysis describes the robustness of the results by excluding some studies such as those for example, of greater risk of bias and/or smaller studies.

Dissimilarities in Control Groups

Just as important as the similarity in treatment groups, is that one needs to take great caution to ensure that control groups between studies included in the meta-analysis are similar. For example, one study in a meta-analysis may have a statin drug vs. placebo, while another study compares a statin drug plus active risk factor management (smoking cessation, hypertension control, etc.) compared to placebo plus active risk factor management. Certainly, one could argue that the between study control groups are not similar (clearly they are not identical), and one can only surmise the degree of bias that would be introduced by including both in the meta-analysis.

Heterogeneity in Outcome

One might expect that the choice of an outcome to be evaluated in a meta-analysis is a simple choice. In many meta-analyses, it is not as simple as one would think. For example, consider a meta-analysis shown in Table 10.7. The range of effect has a risk differential from an approximately 60 % decrease to 127 % increase. One should reasonably ask whether the studies included in the meta-analysis should demonstrate approximately consistent results. Does it make sense to combine studies that are significant in different directions? If studies provide remarkably different estimates of treatment effect, what does an average mean? This particular scenario is used to further illustrate the use of sensitivity analyses in meta-analysis.

Table 10.7 Meta-analysis of stroke as a result of an intervention

Study	Estimate (95 % CI)	
1	1.12 (0.79–1.57)	Fatal and nonfatal first stroke
2	1.19 (0.67–2.13)	Hospitalized F/NF stroke
3	1.16 (0.75–1.77)	Occlusive stroke
4	0.64 (0.06–6.52)	Fatal SAH
5	2.27 (1.22–4.23)	Fatal and nonfatal stroke or TIA
6	0.40 (0.01–3.07)	Fatal stroke
7	0.97 (0.50–1.90)	Fatal and nonfatal first stroke
8	0.63 (0.40–0.97)	Fatal occlusive disease
9	0.97 (0.65–1.45)	Fatal and nonfatal stroke
10	0.65 (0.45–0.95)	Fatal and nonfatal first stroke
OVERALL	0.96 (0.82–1.13)	

A so-called 'influence analysis' is derived in which the meta-analysis is re-estimated after omitting each study in turn. It may be reasonable to consider excluding particular studies, or to present the results with one or two studies included and then excluded. Many analyses start out with the intention of producing quantitative syntheses, and fall short of this goal [21]. If the reasons are well argued, this can often be the most reasonable outcome.

Studies are Reported in Different Formats with Different Information Available

Since studies are reported in different formats with different information available, the abstraction of data can become problematic. There is no reason to anticipate that investigators will report data in a consistent manner. Frequently, differences in measures of association (odds ratio versus regression coefficients versus risk ratios, etc.) are presented in different reports which then forces the abstractor to try to reconstruct the same measure of association across studies. When abstracting information for meta-analyses, one must go through each study and attempt to collect the information in the same format. That is, one needs either a measure of association (e.g. an odds ratio) with some measure of dispersion (e.g. variance, standard deviation, confidence interval), or cell frequencies in 2×2 tables. If one wants to present a meta-analysis of subgroup outcomes, pooling may be even more problematic than pooling primary outcomes. This is because subgroups of interest are frequently not presented in identical categories.

The issue of consistency in the reporting of studies is a particular problem for epidemiological studies where confounders are a major issue. Although confounders are easily addressed by multivariable models, there is no reason to assume that authors will use the same models in adjusting for confounders. Another related problem is the possibility that there are multiple publications from a single population, and it is not always clear that this has occurred. For example, let's say that there is a publication reporting results in 109 patients. Three years later a report from the same or similar authors reports the results of a similar intervention in 500 patients. The question is, were the 500 patients all new, or did the first report of 109 patients get included in the 500 now being reported?

The Use of Random vs. Fixed Analysis Approaches

By far, the most common approach to weighting the results in meta-analyses is to calculate a 'weighted average' of the effects (e.g. odds ratios, risk ratios) across the studies. This has the overall goal of:

- Calculating an 'weighted average' measure of effect, and
- Performing a test to see if this estimated effect is different from the null hypothesis of no effect

In considering whether to use the fixed effect or random effects modeling approach, the fixed effect approach assumes that studies included in the meta-analysis are the only studies to which the inference will be applied, while the random effects approach assumes that the studies are a random sample of studies that may have occurred, and inference can be extended to "studies like these". The fixed effect model weights the studies by their 'precision' only. Precision is largely driven by the sample size and reflected by the widths of the 95 % confidence limits about the study-specific estimates. In general, when weights are assigned by the precision of the estimates they are proportional to (1/var(study)). This method assigns a bigger weight to a big and poorly-done study than it does to a small and well-done study. Thus, a meta-analysis that includes one or two large studies is largely a report of just those studies. Random effects models estimate a between study variance component, and incorporate that into the model. This effectively makes the contributions of individual studies to the overall estimate more uniform. It also increases the width of the confidence interval of the overall effect. The random effects approach is likely more representative of the underlying statistical framework and the use of the 'fixed' approach can provide an underestimate of the true variance and may falsely inflate power to see effects. Most older meta-analyses have used the fixed effect approach, while many newer meta-analyses are using the random effects approach since it is more representative of the 'real' world. A reasonable approach is to present the results from both models.

Assignment of Weights

Alternative weighting schemes have been suggested, such as weighting by the quality of the study, with points given based on the number of variables [22]. The problem with weighting is that one has started the meta-analysis in order to have an objective method to combine studies to provide an overall summary, and with weighting we are subjectively assigning weights to factors so that we can objectively calculate a summary measure. However, this aforementioned weighting is but one scheme and its use has been questioned by many experts in the field. Most meta-investigators now use fixed, random, or Bayesian approaches [23].

Statistical and Graphical Approaches

Forest Plot

The forest plot is a common graphical way of portraying the data in a meta-analysis. In this plot, the point is the estimate of the effect, the size of the point is proportional to the size of the study, and the confidence intervals around that point estimate are displayed (for example, an odds ratio of 1 means the outcome is not affected by the intervention under study). In Fig. 10.2, a hypothetical forest plot of log hazard ratios for each study, ordered by the size of the effect within each study is shown. At the bottom, a diamond shows the combined estimate from the meta-analysis.

An example of some of these aforementioned principles is demonstrated in a theoretical meta-analysis of six studies. For this ‘artificial’ meta-analysis, only multicenter randomized trials were included, and the outcome is total mortality. Tables 10.8a, 10.8b, and 10.8c, present the raw data, mortality rates and odds ratios.

The fundamental statistical approach in meta-analysis is similar to that of an RCT in that the hypothesis is conceived to uphold the null. According to the Mantel-Haenszel-Peto method, a technique commonly used when events are sparse, a 2×2 table is constructed for each study to be included, and the observed number for the outcome of interest is computed [24]. From that computation one subtracts the expected outcome had no intervention been given. If the intervention of interest has no effect, the observed minus the expected should be about zero; if the intervention is favorable (with the measure of association being the odds ratio-OR) the OR will be greater than 1 (as will its confidence limits). The magnitude of effect can be calculated in meta-analyses using a number of measures of association, such as the odds ratio (OR), relative risk (RR), risk difference (RD), and/or the hazard ratio (HR), to name a few. The choice is, to a great degree, subjective as discussed in Chap. 16, and briefly in section “Studies are reported in different formats with different information available” above.

One limited type of meta-analysis, and a way to overcome some of the limitations of meta-analysis in general, is to preplan them with the prospective registration of studies, as has been done with some drug developments. Berlin and Colditz present the potential uses of meta-analyses (primarily of RCTs) in the approval and postmarketing evaluation of approved drugs [25]. If a sponsor of a new drug has a program to conduct a number of clinical trials, and the trials are planned as a series with prospective registration of studies at their inception, one has a focused question (e.g. drug efficacy for lowering the total cholesterol), all patients are included (so no

Table 10.8a The raw data from the six studies included in the meta-analysis

Raw data						
	Treatment A			PLACEBO		
Study	Total no. of patients	No. dead	No. alive	Total no. of patients	No. dead	No. alive
1	615	49	566	624	67	557
2	758	44	714	771	64	707
3	317	27	290	309	32	277
4	832	102	730	850	126	724
5	810	85	725	406	52	354
6	2267	246	2021	2257	219	2038
Total	5599	553	5046	5217	560	4657

Table 10.8b The individual mortality rates from the six studies included in the meta-analysis

	ASPIRIN	PLACEBO	Aspirin-Placebo		
Study	Mortality rate	Mortality rate	Diff	SE of diff	P-value
1	.0797	.1074	−.0277	.0165	0.047
2	.0580	.0830	−.0250	.0131	0.028
3	.0852	.1036	−.0184	.0234	0.216
4	.1226	.1482	−.0256	.0167	0.062
5	.1049	.1281	−.0231	.0198	0.129
6	.1085	.0970	.0115	.0090	0.898

Table 10.8c The odds ratios from the six studies included in the meta-analysis

Odds ratios for the six trials				
Study	Log odds ratio	SE [log OR]	Odds ratio	CI on OR
1	−0.33	0.197	0.72	[0.49,1.06]
2	−0.38	0.203	0.68	[0.46,1.02]
3	−0.22	0.275	0.81	[0.47,1.38]
4	−0.22	0.143	0.80	[0.61,1.06]
5	−0.23	0.188	0.80	[0.55,1.15]
6	0.12	0.098	1.13	[0.93,1.37]

publication bias occurs), one then has the elements of a well-planned meta-analysis. In Table 10.9, Berlin and Colditz present their comparison of trials as they relate to four key elements of several types of clinical trials [23].

In designing a meta-analysis (or reading one in the literature) one should be certain that a number of details are included so the validity of the results can be weighed. Some of the considerations are: listing the trials included and excluded in the meta-analysis and the reasons for doing so; clearly defining the treatment assignment in each of the trials; describing the ranges of patient characteristics, diagnoses, and treatment assignment; and, addressing what criteria were used to decide that the studies analyzed were similar enough to be pooled. Finally, meta-analyses can provide more precise estimates of the effects of interventions, increase

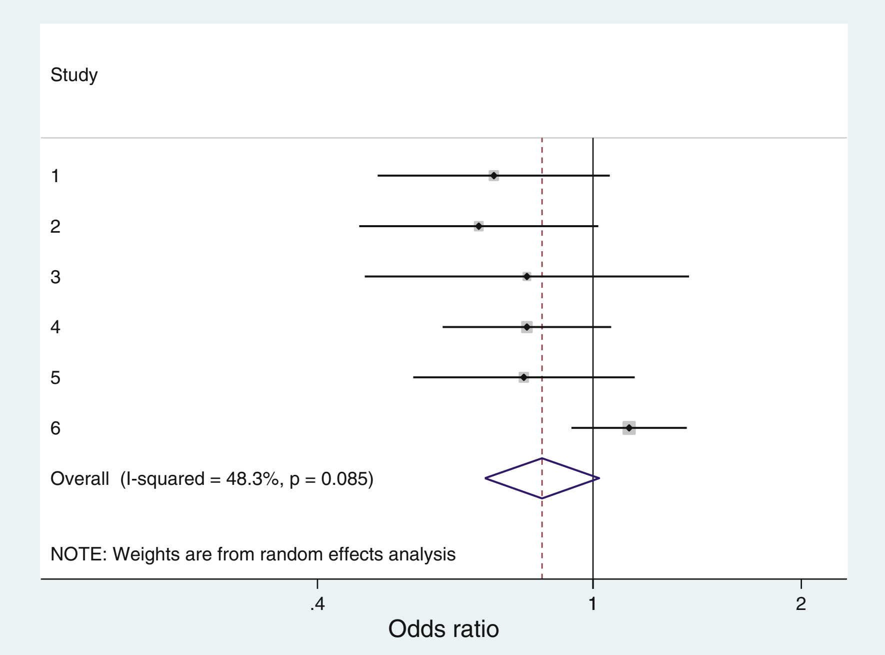

Fig. 10.2 Example of a forest plot

Table 10.9 Variables relating to publication bias, generalizability, and validity with different study approaches

Approach	Avoids publication bias	Generalizes across protocols	Generalizes across centers	Validity
Pre-planned	+++	+++	+++	++
LST	++	_	+++	++
Retrospective	_	++	++	+
2 RCTs	_	++	++	++
1 RCT	_	_	_	+

LST Large Simple Trial

statistical power, assess the amount of variability between studies, reach agreements when results from different studies are discordant, and identify study characteristics associated with particularly effective treatments (Table 10.10). Typically, analyses should include: the point estimate, 95 % confidence limits, a graphical display (forest plot), p values, a statistical test for heterogeneity, sensitivity analyses, and potential sources of bias (e.g. publication bias using the funnel plot).

As is true for clinical trials and the CONSORT Guidelines, there are guidelines for the reporting of meta-analyses: PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) is an update of QUORUM (QUality Of Reporting of Meta-Analyses), for meta-analyses of RCTs; and, Meta-analysis Of Observational

Table 10.10 What can meta-analyses provide?

Provide more precise estimates of the effects of interventions
Increase statistical power
Assess the amount of variability between studies
Reach agreements when results from different studies are discordant
Identify study characteristics associated with particularly effective treatments

Studies in Epidemiology (MOOSE) for meta-analyses of Observational studies see Chap. 2 [26]. In addition, a critical appraisal checklist for a systematic review has been developed under the guidance of the Critical Appraisal Skills Program [27].

Evidence-Based Medicine

> *'It ain't so much what we don't know that gets us into trouble as what we do know that ain't so' (Will Rogers)* (http://humrep.oxfordjournals.org)

Clinical Effectiveness, Clinical Governance, Risk Management, Benchmarking—Essence of Care, NHS Knowledge and Skills Framework, and Evidence-based Practice are but a few of the terms that have now become part of everyday practice for health professionals. Such terms appear to be open to interpretation and confusion. Evidence-based medicine was originally defined as the process of "*…integrating individual clinical expertise and the best available external clinical evidence from systematic research.*" [28] Meta-analysis and evidence-based medicine (EBM) arose together as a result of the fact that the traditional way of learning (the Historic Paradigm i.e. 'evidence' is determined by the leading authorities in the field from textbooks, review articles, seminars, and consensus conferences) was based upon the assumption that experts represented infallible and comprehensive knowledge. Numerous examples of the fallibility of that paradigm are present in the literature e.g.:

- Prenatal steroids for mothers to minimize risk of Respiratory Distress Syndrome (RDS)
- Treatment of eclampsia with magnesium sulfate vs. diazepam
- NTG use in suspected MI
- The use of diuretics for pre-eclampsia

In 1979 Cochrane stated 'It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or sub-specialty, updated periodically, of all relevant randomized controlled trials' [29]. The idea of EBM then was to devise answerable questions, track down the best evidence to answer them, critically appraise the validity and usefulness of the evidence, apply the appraisal to clinical practice, and to evaluate one's performance after applying the evidence into practice [30]. As such, EBM called for the integration of individual clinical expertise with the best available external evidence from systematic research (i.e. meta-analysis). One definition of EBM is the conscientious, explicit judicious use of current best available evidence in making decisions about the care of individual patients with the use of RCTs, wherever possible, as the gold standard [31]. EBM also incorporates the need to encourage patterns of care that do more good than harm.

It has been said, it is not that we are reluctant to use evidence-based approaches, it is that we may not agree on what the evidence is, so why shift to an EBM approach? The answers are many, but include the fact that the volume of new evidence can be overwhelming (this remains the clinician's biggest challenge), there is limited time available to keep up, up-to-date knowledge and clinical performance deteriorate with time; and, traditional CME has not been shown to improve clinical performance.

The necessary skills for EBM include the ability to precisely define a patient problem, ascertain what information is required to resolve the problem, the ability to conduct an efficient search of the literature with the selection of the most relevant articles, the ability to determine a study's validity, extract the clinical message and apply it to the patient's problem [32].

There are, of course criticisms of the EBM approach. For example, some feel that evidence is never enough i.e. evidence alone can never guide our clinical actions and that there is a shortage of coherent, consistent scientific evidence. Also, the unique biological attributes of the individual patient render the use of EBM to that individual, at best, limited. For many, the use of EBM requires that new skills be developed in an era of limited clinician time and technical resources. Finally, who is to say what the evidence is or that evidence-based medicine works? Some have asked, are those who do not practice EBM practicing 'non-evidence-based medicine'? Karl Popper perhaps summarized this best in a very thoughtful and insightful commentary, where he discussed the differences between evidence, truth, and knowledge when he noted that there are all kinds of sources of our knowledge but none has authority [33, 34]. *"Evidence is information that is used to approach truth, whereas truth is an infallible, unequivocal, immutable fact. The definition of knowledge…is typically used as a representation of a person's comprehension of a particular subject."* He further notes that *although truth is our ultimate desire it is likely unattainable, and that although evidence imbues us with knowledge, it does not affirm truth."* [34] As an example, RCTs use inductive reasoning to draw conclusions that are expressions of probability (not truth), but are often dubbed as truth. Baum cites Prasad et al. who define to "*signify the phenomenon of a new trial-superior to predecessors because of better design, increased power, or more appropriate controls-contradicting current clinical practice.*" [35] In Baum's study 212 original publications in the New England Journal of Medicine were reviewed, 124 of which made some claim with respect to medical practice. Of these 124 there were 16 reversals (13 %). That is "truth" was reversed 13 % of the time [35].

Baum ends with the following "…*through the medical systems endowment of the P value and the RCT with boundless unfounded power, the lay public and physicians alike have become confused. Conflicting publications are released nearly on a weekly basis, each of them being treated as gospel with its message being shouted from the rooftops by the media as well as the camera-adoring members of our profession.*

The fact that science is a process is ignored. Undecipherable statistical jargon cloaks the fact that medical evidence emanates not from truth, but instead the falsifiable proof (the rejection of the null hypothesis).” [35]

Evidence-Based Medicine is perhaps a good term to the extent that it advocates more reliance on clinical research than on personal experience or intuition. But, medicine has always been taught and practiced based upon available scientific interpretation. The question can then be asked is whether the results of a clinical trial hardly deserve the title *evidence* as questions arise about the statistical and design aspects, and data analysis, presentation, and interpretation contain many subjective elements as we have discussed in prior chapters. Thus, even if we observe consistency in the results and interpretation (a rare occurrence in science) how many times should a successful trial be replicated to claim proof? That is, whose evidence is *the evidence in evidence-based medicine*?

The five steps of EBM were first described in 1992 as follows [36]

1. The translation of uncertainty into an answerable question
2. Systematic retrieval of the best evidence available
3. A critical appraisal of the evidence (e.g. confounding, selection bias etc.)
4. Application of results into clinical practice (see Chapter on Implementation Research)
5. Performance evaluation

Several guidelines have been suggested as a way of assessing the quality of evidence and include the US Preventative Task Force, the UK National Health Service, and the GRADE Working Group. The US Preventive Services Task Force guidelines rank evidence about the effectiveness of treatments or screening (http://en.wikipedia.org/wiki/Levels_of_evidence):

Level I: Evidence obtained from at least one properly designed randomized controlled trial.

Level II-1: Evidence obtained from well-designed controlled trials without randomization.

Level II-2: Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group.

Level II-3: Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled trials might also be regarded as this type of evidence.

Level III: Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

UK National Health Service

The UK National Health Service uses a similar system with categories labeled A, B, C, and D. These levels are only appropriate for treatment or interventions; different types of research are required for assessing diagnostic accuracy or natural history and prognosis, and hence different "levels" are required. For example, the Oxford Centre for Evidence-based Medicine suggests levels of evidence (LOE) according to the study designs and critical appraisal of prevention, diagnosis, prognosis, therapy, and harm studies [34]:

- Level A: Consistent randomised controlled clinical trial, cohort study, all or none (see note below), clinical decision rule validated in different populations.
- Level B: Consistent retrospective cohort, exploratory cohort, ecological study, outcomes research, case-control study; or extrapolations from level A studies.
- Level C: Case-series study or extrapolations from level B studies.
- Level D: Expert opinion without explicit critical appraisal, or based on physiology, bench research or first principles.

Categories of Recommendations

In guidelines and other publications, recommendations for a clinical service are classified by the balance of risk versus benefit of the service *and* the level of evidence on which this information is based. The U.S. Preventive Services Task Force uses [35]:

- Level A: Good scientific evidence suggests that the benefits of the clinical service substantially outweigh the potential risks. Clinicians should discuss the service with eligible patients.
- Level B: At least fair scientific evidence suggests that the benefits of the clinical service outweigh the potential risks. Clinicians should discuss the service with eligible patients.
- Level C: At least fair scientific evidence suggests that there are benefits provided by the clinical service, but the balance between benefits and risks are too close for making general recommendations. Clinicians need not offer it unless there are individual considerations.
- Level D: At least fair scientific evidence suggests that the risks of the clinical service outweigh potential benefits. Clinicians should not routinely offer the service to asymptomatic patients.
- Level I: Scientific evidence is lacking, of poor quality, or conflicting, such that the risk versus benefit balance cannot be assessed. Clinicians should help patients understand the uncertainty surrounding the clinical service.

The Grading of Recommendations Assessment, Development and Evaluation (The GRADE Working Group)

A newer system was developed by the GRADE working group and takes into account more dimensions than just the quality of medical research. It requires users of GRADE who are performing an assessment of the quality of evidence, usually as part of a systematic review, to consider the impact of different factors on their confidence in the results. Authors of GRADE tables, divide the quality of evidence into four levels, on the basis of their confidence in the observed effect (a numerical value) being close to what the true effect is. The confidence value is based on judgments assigned in five different domains in a structured manner. The GRADE working group defines 'quality of evidence' and 'strength of recommendations' as two different concepts which are commonly confused with each other.

Systematic reviews may include randomized controlled trials that have low risk of bias, or, observational studies that have high risk of bias. In the case of randomized controlled trials, the quality of evidence is high, but can be downgraded in five different domains.

- Risk of bias: Is a judgment made on the basis of the chance that bias in included studies has influenced the estimate of effect.
- Imprecision: Is a judgment made on the basis of the chance that the observed estimate of effect could change completely.
- Indirectness: Is a judgment made on the basis of the differences in characteristics of how the study was conducted and how the results are actually going to be applied.
- Inconsistency: Is a judgment made on the basis of the variability of results across the included studies.
- Publication bias: Is a judgment made on the basis of the question whether all the research evidence has been taken to account.

In the case of observational studies, the quality of evidence starts out lower and may be upgraded in three domains in addition to being subject to downgrading.

- Large effect: This is when methodologically strong studies show that the observed effect is so large that the probability of it changing completely is less likely.
- Plausible confounding would change the effect: This is when despite the presence of a possible confounding factor which is expected to reduce the observed effect, the effect estimate still shows significant effect.
- Dose response gradient: This is when the intervention used becomes more effective with increasing dose. This suggests that a further increase will likely bring about more effect.

Meaning of the levels of quality of evidence as per GRADE

- High Quality Evidence: The authors are very confident that the estimate that is presented lies very close to the true value. One could interpret it as: there is very low probability of further research completely changing the presented conclusions.
- Moderate Quality Evidence: The authors are confident that the presented estimate lies close to the true value, but it is also possible that it may be substantially different. One could also interpret it as: further research may completely change the conclusions.
- Low Quality Evidence: The authors are not confident in the effect estimate and the true value may be substantially different. One could interpret it as: further research is likely to change the presented conclusions completely.

- Very Low Quality Evidence: The authors do not have any confidence in the estimate and it is likely that the true value is substantially different from it. One could interpret it as: New research will most probably change the presented conclusions completely.

Guideline panelists may make strong or weak recommendations on the basis of further criteria. Some of the important criteria are:

- Balance between desirable and undesirable effects (not considering cost)
- Quality of the evidence
- Values and preferences
- Costs (resource utilization)

Despite the differences between systems, the purposes are the same: to guide users of clinical research information on which studies are likely to be most valid. However, the individual studies still require careful critical appraisal.

In summary, the term EBM has been linked to three potentially false premises: that evidence has a purely objective meaning in biomedical science; that one can distinguish between what is evidence and what is lack of evidence; and that there is evidence-based, and non-evidence-based medicine. As long as it is remembered that the term evidence, while delivering forceful promises of truth, is limited in the sense that scientific work can never prove anything but only serves to falsify, the term has some usefulness. Finally, EBM does rely upon the ability to perform systematic reviews (meta-analyses) of the available literature, with all the attendant limitations of meta-analyses discussed above.

In a "tongue and cheek" article, Smith and Pell addressed many of the above issues in an article entitled "*Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomized controlled trials*" [37]. In their results section, they note that they were unable to find any RCTs of "parachute intervention". They conclude that:

> *only two options exist. The first is that we accept that under exceptional circumstances, common sense might be applied when considering the potential risks and benefits of interventions. The second is that we continue our quest for the holy grail of exclusively evidence-based interventions and preclude parachute use outside of a properly conducted trial. The dependency we have created in our population may make recruitment of the unenlightened masses to such a trial difficult. If so, we feel assured that those who advocate evidence-based medicine and criticize use of interventions that lack evidence-base will not hesitate to demonstrate their commitment by volunteering for a double blind, randomized, placebo controlled, crossover trail.* (See Fig. 10.3)

Isaacs has embellished this with a list for the basis of clinical decision making (Table 10.11) in which evidence is one, and then eminence, vehemence, eloquence, providence, diffidence, nervousness, and confidence round out the list [38]. For each they describe the bias as follows: eminence based medicine-"the more senior the colleague, the less importance he or she placed on the need for anything as mundane as evidence"; vehemence based medicine- is determined by the loudest colleague; eloquence based medicine is predicted on sartorial elegance as a powerful substitute for evidence; providence based medicine occurs when you have no clue what to

Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomised controlled trials

Fig. 10.3 Humorous example of evidence-based medicine (With permission: Smith and Pell [37]

Table 10.11 Humorous outline of the basis of clinical decision making

Basis for clinical decision making	Marker	Measuring device	Unit of measurement
Evidence	RCT	Meta-analysis	Odds ratio
Eminence	Grey hair	Luminometer	Optical density
Vehemence	Stridency	Audiometer	Decibels
Eloquence	Sartorial splendor	Teflometer	Adhesion score
Providence	Religious fervor	Genuflection angle	Piety units
Diffidence	Gloom level	Nihilometer	Sighs
Nervousness	Litigation phobia	Every conceivable test	Bank balance
Confidence	Bravado	Sweat test	No sweat

Adapted from: Isaacs and Fitzgerald [38]

do and you turn to God to give you a hand with decision making; diffidence based medicine is when nothing is done out of a sense of despair, however they further point out that this may be beneficial since doing something may be worse ("don't just do something, stand there" as the axiom goes); nervousness based medicine is decision making based on fear of litigation (here the only bad test is "the one you didn't think of ordering"); and finally, confidence based medicine, which the authors point out is restricted to surgeons.

Clinical Practice Guidelines

Another outcropping from evidence-based medicine is clinical practice guidelines. Guideline recommendations have become the standard of care, and quality of care is increasingly assessed on the basis of adherence to these recommendations. In 1990 the Institute of Medicine defined practice guidelines as *"systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances*" [39]. In an editorial, Gibbons et al. noted that "*as the number of available guidelines provided by a variety of sources has literally exploded, serious questions and controversies have arisen about how guidelines should be developed, implemented, and evaluated.*" [40] They go on to point out that guideline developers have been criticized for failing to control for conflicts of interest, for variable quality, and for failing to prove that guidelines benefit patients.

Despite the fact that guideline recommendations are being used to asses standard of care, clinical practice guidelines are recommendations (not rules or standards) about the care of patients with a specific condition and ideally are based upon the "best available evidence", but should always be tempered on the basis of individual patient circumstances and preferences. The American Academy of Family Practice defines guidelines as *"a recommendation issued for the purpose of influencing decisions about health interventions*." The "best available evidence" is generally considered evidence from systematic reviews ideally of randomized controlled trials. Although guidelines are intended for clinicians, they are (perhaps unfortunately) used by others to monitor physician practice and in medical-legal proceedings. These aforementioned uses sometimes do not recognize that guidelines are suggestions for care not mandates, and only apply to a percentage of patients with a condition and certainly not all patients. However, the impetus for practice guidelines is many and includes:

- Increasing/changing medical knowledge
- Rising health care costs unrelated to health outcomes
- Wide variations in clinical decisions
- Desire for evidence-based, outcomes-oriented clinical decisions

The reality in medicine is that there has been an explosion of knowledge technology, and of patient expectations and "just keepin' up" with the literature (much less reviewing older literature) is problematic. For example in 1998 there were at least 20,657 articles involving human beings. If one slept 4 h a night, spent 25 h a week seeing patients and 1 h a day on personal activities, and read three articles an hour in the remaining awake time, after 1 year one would be 3,800 additional articles behind. There is no question that the complexity of medical decisions is rapidly growing and that there is uncertainty and variability in medical practice and even the best-trained physician with the greatest experience is not perfect. The above-average physician has even more problems with consistency and accuracy. Thus, there is variability in clinical judgment, a question about the reliability of diagnostic judgment (If a doctor tells you that you have a disease, do you have it? If a doctor does not find a disease, are you well?), and physician decisions can be highly variable (it is well known that physicians can disagree with their peers who have reviewed the same patient, and that they can disagree with themselves when presented with the same patient records at two points in time). An example of this aforementioned disagreement is a study of four cardiologists presented with high-quality angiograms and asked to determine if stenosis in the proximal or distal left anterior descending artery was >50 %.

- The cardiologists disagreed on 60 % of the cases [41]
- Cardiologists looking at the same angiograms at two points in time disagree with themselves 8–37 % of the time [42]

It is also known that there is substantial geographic variability in the rates of procedures.

Guideline recommendations come from medical textbooks, review articles, meta-analyses, expert opinion and consensus panel recommendations, but whereas the US government was once the primary source of guidelines, this is now mostly the province of specialty and subspecialty societies with the exception of the US Preventive Service Task Force. There are instances where there is disagreement amongst the guidelines and although the disagreements are usually minor, the disagreements are certainly a barrier to their acceptance, although clinicians are most likely to accept the recommendations from their own specialty society (and least likely to accept recommendations from managed care organizations or insurance companies).

Some guideline panels use a grading system (discussed above) attached to their recommendations based on the strength of evidence leading to the recommendation.

Summary of Concerns About Guidelines

- Guidelines are often outdated by the time they are released. (Burn your textbooks, except this one, of course)
- Guidelines often emphasize peer consensus rather than outcome evidence
- Guidelines ignore patient preference.

Other Concerns

Evidence-based guidelines disregard effective treatments that have not been evaluated in systematic experimental studies. A treatment might get a low rating because it does not work *or* because it has not been evaluated in a randomized clinical trial. Evidence-based medicine assumes that untested treatments are ineffective. Finally, many clinicians view practice guidelines as “cook book medicine” with “not enough recipes in the cookbook” [43].

The limitations in the evidence EBM is nicely reviewed by Sniderman et al. in response to a commentary by Prasad who discusses the two medical world views of whether RCTs are needed to accept new practices [44]. Some of the limitations discussed by Sniderman et al. include:

For many clinical problems there simply is no RCT evidence
Other times multiple RCTs have been performed but the conclusions are in conflict
RCTs are limited in their generalizability

There are limitations in applying the results in a group of patients to the individual

In an attempt to overcome some of the above limitations meta-analyses are performed, but meta-analyses have their own set of limitations (see above)

There are limitations in the guideline process which is also developed to address some of the above problems (e.g. conflicts of interest, failure to ensure dissenting and minority viewpoints, the absence of a process to challenge the validity of specific conclusions that guidelines reach)

There can be a diminution of clinical reasoning as a result of guideline recommendations

References

1. Meinert CL. Meta-analysis: science or religion? Control Clin Trials. 1989;10:257S–63S.
2. Boden WE. Meta-analysis in clinical trials reporting: has a tool become a weapon? Am J Cardiol. 1992;69:681–6.
3. Oxman AD. Meta-statistics: help or hindrance? ACP J Club. 1993;118:A-1–13.
4. Goodman SN. Have you ever meta-analysis you didn't like? Ann Intern Med. 1991;114:244–6.
5. Bangalore S. Dueling data: separating the wheat from the statistical chaff. CardioSource WorldNews. 2012;Dec:14.
6. Pearson K. Report on certain enteric fever inoculation statistics. Br Med J. 1904;3:1243–6.
7. Beecher HK. The powerful placebo. JAMA. 1955;159:1602–6.
8. Glass G. Primary, secondary and meta-analysis of research. Educ Res. 1976;5:3–8.
9. Petitti DB. Approaches to heterogeneity in meta-analysis. Stat Med. 2001;20:3625–33.
10. Neveol A, Dogan RI, Lu Z. Author keywords in biomedical journal articles. AMIA Ann Symp Proc/AMIA Symp. 2010;2010:537–41.
11. Turner EH, Knoepflmacher D, Shapley L. Publication bias in antipsychotic trials: an analysis of efficacy comparing the published literature to the US Food and Drug Administration database. PLoS Med. 2012;9:e1001189.
12. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002. doi:10.1136/bmj.d4002.
13. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50:1088–101.
14. Egger M, Smith DG, Altman DG. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2000.
15. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56:455–63.
16. Candelise L, Ciccone A. Gangliosides for acute ischaemic stroke. Cochrane Database Syst Rev. 2001;4:CD000094.
17. Easterbrook. Publication bias in clinical research. Lancet. 1991;337:867.
18. Smith ML. Publication bias and meta-analysis. Eval Educ. 1980;4:22–4.
19. Glass G. Meta-analysis at 25. 2000. Available at: http://glass.ed.asu.edu/gene/papers/meta25.html
20. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Br Med J. 2003;327:557–60.
21. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. Br Med J. 2008;336:1413–5.
22. Chalmers TC, Smith Jr H, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials. 1981;2:31–49.

23. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2000.
24. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22:719–48.
25. Berlin JA, Colditz GA. The role of meta-analysis in the regulatory process for foods, drugs, and devices. JAMA. 1999;281:830–4.
26. Stroup DF, Berlin JA, Morton SC, Olkin L, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology. JAMA. 2000;283:2008–12.
27. Carroli A, Mackey ME, Bergel E. Critical appraisal of systematic reviews. The World Health Organization. www.casp-uk.net. Accessed 20 Aug 2013.
28. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312:71–2.
29. The Cochrane Library. Issue 2. Chichester: Wiley; 2007.
30. Evidence-Based Medicine. 1999. http://library.uchc.edu/lippub/fall99.PDF. Accessed 29 July 2013.
31. Panda A, Dorairajan LN, Kumar S. Application of evidence-based urology in improving quality of care. Indian J Urol. 2007;23:91–6. PMC2721549.
32. Uniformed Services University James A Zimble Learning Resource Center. Evidence-Based Medicine (EBM) Resources. 2000; Available from: www.lrc.usuhs.edu/lrcguides/?q=node/16
33. The Problem of Induction. 1953, 1974. Accessed at http://dieoff.org/page126.htm
34. Baum SJ. Evidence-based medicine: what's the evidence? Clin Cardiol. 2012;35:259–60. doi:10.1002/clc.21968.
35. Prasad V, Gall V, Cifu A. The frequency of medical reversal. Arch Intern Med. 2011;171:1675–6. doi:10.1001/archinternmed.2011.295.
36. Cook DJ, Jaeschke R, Guyatt GH. Critical appraisal of therapeutic interventions in the intensive care unit: human monoclonal antibody treatment in sepsis. J Club Hamilt Reg Crit Care Gr J Intensive Care Med. 1992;7:275–82.
37. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. Br Med J. 2003;327:1459–61.
38. Isaacs D, Fitzgerald D. Seven alternatives to evidence based medicine. Br Med J. 1999;319:1618–9.
39. Institute of Medicine. Clinical practice guidelines: directions for a new program, Committee to advise the public health service on clinical practice guidelines. In: Field MJ, Lohr KN, editors. US Dept of Health and Human Services. Washington, DC: National Academy Press; 1990.
40. Gibbons GH, Shurin SB, Mensah GA, Lauer MS. Refocusing the agenda on cardiovascular guidelines: an announcement from the National Heart, Lung, and Blood Institute. Circulation. 2013;128:1713–5. doi:10.1161/CIRCULATIONAHA.113.004587.
41. Zir LM, Miller SW, Dinsmore RE, Gilbert JP, Harthorne JW. Interobserver variability in coronary angiography. Circulation. 1976;53:627–32.
42. Detre KM, Wright E, Murphy ML, Takaro T. Observer agreement in evaluating coronary angiograms. Circulation. 1975;52:979–86.
43. Parmley WW. Clinical practice guidelines. Does the cookbook have enough recipes? JAMA. 1994;272:1374–5.
44. Prasad V. Why randomized controlled trials are needed to accept new practices: 2 medical worldviews. Mayo Clin Proc. 2013;88:1046–50. doi:10.1016/j.mayocp.2013.04.026.