2.1 Introduction

The support of medical decisions comes from several sources. These include individual physician experience, pathophysiological constructs, pivotal clinical trials, qualitative reviews of the literature, and, increasingly, meta-analyses. Historically, the first of these four sources of knowledge largely informed medical and dental decision makers. Meta-analysis came on the scene around the 1970s and has received much attention. What is meta-analysis? It is the process of combining the quantitative results of separate (but similar) studies by means of formal statistical methods. Statistically, the purpose is to increase the precision with which the treatment effect of an intervention can be estimated. Stated in another way, one can say that meta-analysis combines the results of several studies with the purpose of addressing a set of related research hypotheses. The underlying studies can come in the form of published literature, raw data from individual clinical studies, or summary statistics in reports or abstracts.

More broadly, a meta-analysis arises from a systematic review. There are three major components to a systematic review and meta-analysis. The systematic review starts with the formulation of the research question and hypotheses. Clinical or substantive insight about the particular domain of research often identifies not only the unmet investigative needs, but helps prepare for the systematic review by defining the necessary initial parameters. These include the hypotheses, endpoints, important covariates, and exposures or treatments of interest. Like any basic or clinical research endeavor, a prospectively defined and clear study plan enhances the expected utility and applicability of the final results for ultimately influencing practice or policy.

After this foundational preparation, the second component, a systematic review, commences. The systematic review proceeds with an explicit and reproducible protocol to locate and evaluate the available data. The collection, abstraction, and compilation of the data follow a more rigorous and prospectively defined objective process. The definitions, structure, and methodologies of the underlying studies must be critically appraised. Hence, both “the content” and “the infrastructure” of the underlying data are analyzed, evaluated, and systematically recorded. Unlike an informal review of the literature, this systematic disciplined approach is intended to reduce the potential for subjectivity or bias in the subsequent findings.

Typically, a literature search of an online database is the starting point for gathering the data. The most common sources are MEDLINE (United States Library of Medicine) and EMBASE (Excerpta Medica database) for a comprehensive search of the medical literature. The search terms and strategies are crucial in identifying relevant studies. After the search, a critical appraisal of the studies is conducted to ensure that they meet the inclusion criteria and are methodologically sound. The data from these studies are then extracted and combined using meta-analytic techniques to estimate the overall effect size and its confidence interval. This process allows for a more robust and evidence-based answer to the research question.

The meta-analysis provides a more precise estimate of the effect than any single study could have provided on its own. This is because the variability among study results is reduced, leading to a more stable and reliable estimate. The overall effect size can then be compared to existing knowledge, whether based on individual studies or previous meta-analyses, to assess the consistency of findings. If the results are inconsistent or if the meta-analysis fails to provide a convincing estimate of the effect, the conclusions may need to be qualified or modified.

However, it is important to recognize the limitations of meta-analyses. One of the main limitations is the potential for heterogeneity among the studies included. If the studies are too diverse in terms of design, patient characteristics, or intervention methods, combining their results may not yield meaningful or reliable findings. Another limitation is the potential for publication bias, where studies with statistically significant results are more likely to be published than those with nonsignificant results. This can lead to an overestimation of the true effect if only published studies are included in the meta-analysis.

Despite these limitations, meta-analyses remain a valuable tool in evidence-based medicine, providing a more comprehensive and reliable summary of the available evidence. They enable researchers and practitioners to make more informed decisions, improve patient care, and optimize clinical outcomes.
Medicine database), EMBASE (medical and pharmacologic database by Elsevier publishing), CINAHL (cumulative index to nursing and allied health literature), CANCERLIT (cancer literature database), and the Cochrane Collaborative [9]. These online resources have increased the likelihood that investigators will have ready access to all the available results from research directed at a common endpoint.

The third aspect is the actual meta-analysis. Meta-analysis is the statistical synthesis of the results of the studies included in the systematic review. Meta-analysis is the process of combining the quantitative results of separate (but similar) studies by means of formal statistical methods in order to increase the precision of the estimated treatment effect. It is generally used when individual trials yield inconclusive or conflicting results. It may also be used when several trials asking similar questions have been conducted and an overall conclusion is needed.

While the main focus of this chapter will be meta-analysis, it cannot be completely isolated from several prerequisites assessed in the systematic review. For example, the studies must address a common question. The eligibility criteria of the underlying studies must be well established. Evaluation techniques for endpoints must be reasonably consistent across the studies. In the clinical setting, when making comparisons between a treatment and control, the underlying studies must be properly randomized. Exploratory meta-analyses and meta-regressions may examine the associations between interventions, covariates, and secondary events [1].

After providing some historical context, the key definitions, elements, and output of meta-analytic methods will be reviewed with illustrative examples. Understanding these allows one to appreciate better the advantages and limitations of this approach to interpreting results from the literature.

### 2.1.1 Background

This idea of combining results from several studies is not new. Karl Pearson in 1904 established the first identifiable formal technique for data pooling [16]. He examined the correlation coefficients between typhoid and mortality by inoculation status among soldiers in various parts of the British Empire. As a summary measure, Pearson calculated the arithmetic mean of the correlation coefficients across five two-by-two contingency tables. In 1931, Tippett described a method for evaluating the likelihood of a significant effect from the ordered $p$-values observed across studies [22]. Subsequently in 1932, Fisher established a procedure for combining $p$-values from various studies asking a similar question [10]. In the hope of making a definitive statement about a treatment effect, he examined the null hypothesis of no treatment effect over all the studies vs. the alternative research hypothesis of a pooled treatment effect. A fairly straightforward statistical procedure, Fisher’s method is still used to combine the study results. If we assume that there are $k > 1$ studies being examined for the superiority of some form of intervention, call it A vs. not having A, then each of these studies has a $p$-value associated with it. Call them $p_1, p_2, \ldots, p_k$. Fisher established that the statistic, $2 \sum_{i=1}^{k} \log(p_i)$, has a chi square distribution with $2k$ degrees of freedom (df). Fisher computed $Y$ and compared that value to the tabled chi square value on $2k$ df for an alpha level of 0.05. If $Y$ was greater than that tabled value, then the null hypothesis would be rejected at the 0.05 level. Hence, the intervention, A, would be declared effective or superior to not having A. As an example of Fisher’s method, suppose there were five studies comparing some form of intervention, A, to a control, and the $p$-values for the five studies were $p_1 = 0.07$, $p_2 = 0.06$, $p_3 = 0.045$, $p_4 = 0.035$, and $p_5 = 0.05$. $Y$ would have the value 29.843 which is greater than the tabled value of 18.31 on $2 \times 5 = 10$ df. Thus A is effective overall from the five studies at the alpha = 0.05 level. Extensions of Fisher’s methods with weighted $p$-values were subsequently developed.

As the statistical methods for pooling data across studies evolved, the concept of combining the treatment effects has largely replaced combining the $p$-values. Further, the influence of each trial in the combined estimate of treatment effect is now typically weighted. Most commonly, each study’s contribution to the overall treatment effect is weighted by the inverse of the variance of the estimated treatment effect for that particular study. Since variance is largely a function of sample size, studies with large sample size will have smaller variance (and larger inverse variance.) Hence, larger studies will generally have a greater influence on the overall estimated treatment effect across studies.

The term “meta-analysis” was coined by Glass in 1976 who stated that, “Meta-analysis refers to the analysis
of analyses...the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” [11]. Meta-analytic techniques have continued to evolve and be refined with seminal advances such as DerSimonian and Laird’s development of the random-effects model and the advancement of meta-regression methods to explain across study heterogeneity [6]. The application of meta-analysis to enhance clinical decision making, craft guidelines, and inform policy has rapidly flourished and been widely embraced.

Every clinician in training knows the old saw, “For every study cited, an equal and opposite study can be found.” When studies address a similar question and the results do not agree, the role of meta-analysis is to hopefully put all the results together and reach a definitive conclusion about the positive, negative, or inconclusive impact of an intervention. Whether gathering aggregate data from the literature or pooling raw individual patient data (IPD) from multiple studies, the motivation is to reach a definitive conclusion about a therapy when several studies involving that intervention have been conducted. Tamoxifen treatment of early stage breast cancer provides a classic example [7]. Most of the individual separate studies had a favorable but statistically insignificant treatment effect. Several had unfavorable results albeit statistically insignificant. The beneficial mortality effect of tamoxifen therapy could not be definitively established until the patient-level data was combined from all the trials in a meta-analysis.

“Real world” limitations such as funding, available patients, access to testing and follow-up resources, and logistical limitations often inherently preclude adequately powered mega-trials to answer many investigative questions. Hence, methods to properly combine studies to obtain more precise results have understandable utility and appeal. Most disciplines, whether oncology, cardiovascular health, or dental interventions, tend to employ similar endpoints within their respective fields. Mean effects, response rates, odds ratios, time to events (survival, disease-free survival), correlations, etc. are common measures. As stated above, most meta-analyses are performed from data gathered in the literature or from pooled raw data. One has to understand that studies contributing to a meta-analysis are all asking a similar question, but all are not necessarily asking the same question. For example, if one wanted to examine whether receiving a statin after the first myocardial infarction reduces the chance of death within 5 years, several studies have addressed this issue. One may involve one type of statin vs. a placebo control, while another may be a different statin vs. a fibrate. They do not involve the same active treatment or the same control. However, they are asking a similar question in that does receiving a statin in some form vs. not receiving a statin reduce the risk of death over a certain period of time.

When one pools data from various trials it enhances the power. Small or moderate, but potentially meaningful differences can be detected that individual trials could not definitively establish. A clinical trial may be inconclusive in that it did not demonstrate statistical superiority of a treatment or intervention. Sometimes, one refers to this as a “negative” trial. However, that may not be a proper term. One learns something from all trials. One may prefer the term “inconclusive” for a study demonstrating neither clear benefit nor harm. Studies not having sufficient sample size or long enough follow-up to attain statistical significance to reach a definitive endpoint is commonplace because of resource limitations. Obviously, inconclusive results in a clinical trial do not necessarily mean the treatment is ineffective. Unfortunately, dismissal of an inconclusive trial as “negative” may contribute to publication bias. Since “positive” trials are more apt to receive prompt well-placed publication than the inconclusive trials, overly optimistic impressions of the treatment effect may arise from this publication bias. Diligent search for and location of both unpublished trials and those in less prominent journals is warranted. The meta-analyst must carefully evaluate this less formal data. This additional information gathered from more studies or subjects may lead to more accurate estimation of the treatment effect. Also, several studies may have a more heterogeneous sample than a single study. The result from this more diverse population may have more generalizability to unselected populations. Thus hopefully, one gathers a more global solution to a question or hypothesis based on most, if not all, existing evidence from all relevant studies.

This is particularly true in oncology where treatment effects are often small and contradictory results are likely to occur. The advantages of a meta-analysis of the pooled raw data from studies, if it can be obtained, are that, one can have more descriptive capabilities and apply covariate adjustment as well as have a rich database for future analysis as longer follow-up is accrued.
Most meta-analyses, however, are done using summary data from published articles, abstracts, or reports. The advantages are that data is easily gathered in statistical summary form and there are statistical techniques available for accurately combining the results. It may be necessary to summarize the results from older studies, where the data is lost, inaccessible, or too expensive to retrieve. A meta-analysis of the aggregate data from several trials may be the first step in motivating the commitment of effort and resources in pursuing the raw data for more detailed analyses. Methods have been developed and validated for mixed models combining IPD from some trials with aggregate data from other trials when the IPD is unavailable. Granted there are limitations such as inadequate details in the publications, protocol violations, undocumented patient compliance rates, and insufficient follow-up in longitudinal trials. Fortunately, in recent times, publications for the most part receive a good statistical review and the results are generally reliable.

### 2.1.2 Overview of Statistical Terms and Approaches

An appreciation of the terms and methods of meta-analysis assures deeper understanding of the ultimate results and their proper interpretation and limitations. Our strategy in this section is to present two meta-analyses, the first involving continuous data in Sect. 2.1.1 and the second involving categorical data in Sect. 2.1.2. We introduce the terms and analysis approaches with some attention to the limitations of which the readers may wish to be aware.

Meta-analyses deal with common endpoints which are well defined, easily measurable, and have been measured in a consistent way over all studies over time. Rather than combining p-values like Fisher, each individual study’s result is translated into an effect size (ES). ESs may be the differences in the mean effect of treatment or intervention, A, vs. the mean effect of treatment or intervention, B. Other ES examples are response rates (proportion of successes on A vs. proportion of successes on B) and time to event endpoints such as survival, time to progression, time to second myocardial infarct (MI), etc. ES can also be expressed in terms of odds ratios (OR), relative risk (RR), or hazard ratios (HR). The methodology of meta-analysis is to examine the ES from each study and then to statistically combine these ES to determine a common overall ES from all the studies and test the statistical significance of this combined statistic, ES.

#### 2.1.2.1 Continuous Data

In the first example, meta-analysis examines continuous endpoint data such as LDL cholesterol level. In general terms, we wish to know if the mean reduction in LDL cholesterol is greater with treatment A than with treatment B. There are seven studies \( (k = 7) \) in this example, where A is the treatment of interest and its comparison group is represented by B. The standardized ES \( \delta_i \) in each individual study in this application is:

\[
\delta_i = (\mu_{Ai} - \mu_{Bi})/\sigma_i.
\]

This difference in the population means \( \mu_{Ai} \) and \( \mu_{Bi} \) with treatment A compared to treatment B in the ith study is then standardized by dividing it by the population standard deviation \( \sigma_i \) in the ith study. This standard ES may vary depending on how the standard deviation is defined. For example, it can be the standard deviation of group A or group B (the reference or control group) or the pooled standard deviation of the two populations [12]. We will assume the pooled estimate for our example. The population ES \( \delta_i \) as defined by the population parameters \( \mu_{Ai}, \mu_{Bi}, \) and \( \sigma_i \) is of course estimated by the realizations of these parameters in the observed data of the ith study.

\[
ES_i = (Y_{Ai} - Y_{Bi})/S_i,
\]

In our case, each study will have an estimate of the ES, where \( Y_{Ai} \) minus \( Y_{Bi} \) is the difference in mean observed LDL cholesterol levels between treatment A and B in the particular study divided by the pooled standard deviation of the cholesterol levels in both A and B groups. The standardized ES becomes unitless or “scale free.” The pooled ES will be calculated from these 7 ES from the seven studies (see Table 2.1). The values, \( n_a \) and \( n_b \), in the first column of Table 2.1 are the number of subjects on each treatment from A and B, respectively. The middle column of Table 2.1 is the ES from each of the seven studies. One can see that in studies 1, 2, 3, 6, and 7, the ES is negative indicating a treatment effect in favor of A. This is not true for studies 4 and 5.
The last column gives the p-value for testing the ES = 0 in each study. Only studies 2 and 7 have a significant effect in favor of A. The overall ES in the last row is −0.143 showing a greater mean reduction in cholesterol on A. However, the overall p-value is 0.114 in the last column indicating no significant effect overall.

There is some discussion as to the accuracy of equation (2) in estimating the true population ES represented by equation (2.1) [12]. When the combined sample size of A and B within a single study is less than 50, the ES of (2) may overestimate the true population ES. Some authors [12] have suggested a correction factor (CF) with theoretical justification of 

$$\text{CF} = (1 - 3/(4N - 9))$$

multiplied by ES of (2.2) or CFxES to more closely approximate the true population ES where N is the sum of the sample sizes of the two groups, A and B, within a study ($N = n_A + n_B$). Clearly, as N becomes large this CF approximates one. This CF adjustment was made to the middle column of Table 2.1, although it was hardly necessary as all our sample sizes are fairly large.

The individual and overall study data from such meta-analyses is typically displayed in a Forest Plot as shown in Fig. 2.1. The Forest plot is a graphical summary of the meta-analytic statistics. The title, “std diff in means and 95% CI” is the ES with the 95% confidence intervals for each study. The dark rectangles represent the ES of each study found in the middle column of Table 2.1. Note that the larger the area of the rectangle the larger the sample size. Also, the horizontal lines of the boxes represent the 95% confidence interval of the ES. The rectangles to the left of the null line (treatment difference between A and B equals 0) favor treatment A. Those to the right of the null line favor treatment B. The horizontal lines that overlap the 0 vertical line indicate statistical nonsignificance at the 0.05 level for that particular study. This is consistent with the last column of Table 2.1. The diamond at the bottom of the Forest plot is the overall ES of the seven studies. The vertical points of the diamond overlap or are equal to the value, −0.143, favoring treatment, A, and the horizontal points are the lower and upper limits of the 95% confidence interval. The 95% confidence interval for the overall population ES is (−0.320, 0.034), which includes 0 and thus overlaps the null line in the Forest Plot. Hence, the overall combined results are not statistically significant. The actual p-value = 0.114.

### 2.1.2.1.1 Heterogeneity

Now, having understood the contents of Table 2.1 and Fig. 2.1, we can discuss and demonstrate one of the issues in meta-analysis called heterogeneity. When doing a meta-analysis, the issue of heterogeneity across the studies may affect the study results. This is simply the variation among the studies. In terms of a statistical hypothesis, we can write the null hypothesis: $H_0$: no heterogeneity or that the treatment effect is the same in all $k$ (in our case $k=7$) studies. The alternative hypothesis is: $H_1$: the treatment effect varies over the $k$ studies included in the meta-analysis, thus giving

### Table 2.1 Effects sizes and p-values for the seven LDL studies

<table>
<thead>
<tr>
<th>Study name ($n_A, n_B$)</th>
<th>Effect size $(Y_A - Y_B)/S$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(32, 26)</td>
<td>−0.160</td>
<td>0.545</td>
</tr>
<tr>
<td>2(26, 33)</td>
<td>−0.717</td>
<td>0.008</td>
</tr>
<tr>
<td>3(32, 43)</td>
<td>−0.261</td>
<td>0.266</td>
</tr>
<tr>
<td>4(37, 41)</td>
<td>0.324</td>
<td>0.156</td>
</tr>
<tr>
<td>5(37, 38)</td>
<td>0.432</td>
<td>0.064</td>
</tr>
<tr>
<td>6(35, 40)</td>
<td>−0.112</td>
<td>0.629</td>
</tr>
<tr>
<td>7(48, 38)</td>
<td>−0.627</td>
<td>0.005</td>
</tr>
<tr>
<td>Overall</td>
<td>−0.143</td>
<td>0.114</td>
</tr>
</tbody>
</table>

Fig. 2.1 Forrest plot of the seven LDL studies
heterogeneity. All we are doing here is testing for consistency among the ESs across the studies. If we reject $H_0$, then we have heterogeneity. This is important because heterogeneity implies that there is some underlying difference or differences among the included studies. We often want to know what this difference is. We would like to be able to understand the cause and explain this heterogeneity or variance. Note in Table 2.1 that 5 of the ES are negative or favor treatment A and 2 of the ES are positive or favor treatment B. Note also in Fig. 2.1 that 5 of the rectangles are to the left of the vertical 0 line and 2 are to the right. We want to know if statistically, this is a sufficient enough discrepancy to cause us to be concerned about heterogeneity.

So, the next task is to formally test for significant heterogeneity among the studies at some prespecified level of significance, typically $\alpha=0.05$ or 0.10. So, in addition to some visual indication of heterogeneity in the Forest Plot, we have a formal test statistic called the $Q$ statistic. The $Q$ statistic is used to determine the degree of heterogeneity across studies. The benchmark we use is that if $Q$ is close to 0 then there is no heterogeneity. The formula for calculating $Q$ involves calculating a function of the difference between each study ES and the overall mean ES of all studies combined [13]. If the ES from each study is close to the overall mean ES, showing little or no difference among the ES, then the $Q$ statistic will be close to zero indicating little or no difference among the ESs. The $Q$ has a chi square distribution [13] on $k-1$ df or the number of studies minus one. Like other statistics we calculate, the $Q$ statistic is compared to a critical value for the chi square distribution. The null hypothesis is that there is no heterogeneity in the combined analysis or $H_0$: $Q=0$ vs. the alternative hypothesis, $H_1$: $Q\neq0$ implying that there is significant heterogeneity. The critical value of the $k$ studies for a chi square distribution on $71=6$ df at an alpha level of 0.05 is 12.59. The calculated value of $Q$ from our data is 19.76. Clearly, the calculated $Q$ is much greater than the critical value. Therefore, we reject the null hypothesis and determine that there is significant heterogeneity among our studies. As a matter of fact, the exact $p$-value is 0.003.

This does not change our conclusion that there is no significant treatment effect due to A. However, it does alert us that we should try to determine the source of the heterogeneity. Some possible sources of heterogeneity may be the clinical differences among the studies such as patient selection, baseline disease severity, differences in the treatment applied, management of intercurrent outcomes (toxicity), patient characteristics, time point or conditions of measurements, or others. One may also investigate the methodological differences between the studies such as the mechanism of randomization, extent of withdrawals, and follow-up failure in longitudinal studies. One should also note that heterogeneity can still be due to chance alone and the source not easily detected. Quantifying heterogeneity can be one of the most troublesome aspects of meta-analysis. It is important, because it can affect the decision about the statistical model to be selected (fixed or random effects described below). If significant heterogeneity is found, then potential moderator variables can be found to explain this variability. It may require concentrating the meta–analysis on a subset of studies that are homogeneous within themselves.

Another statistic used to quantify heterogeneity is the $F$ index which quantifies the extent of heterogeneity from a collection of ESs by comparing the $Q$ statistic to its expected value assuming homogeneity, that is to its degrees of freedom, $(df=k-1)$. When the $Q$ statistic is smaller than its df, then $F$ is truncated to 0. The $F$ index can easily be interpreted as the percentage of heterogeneity in the system or the amount of the total variation accounted for by the between studies variance. For example, a meta-analysis with $F=0$ means that all variability in ES estimates is due to sampling error within studies. On the other hand, a meta-analysis with $F=60$ means that 60% of the total variability among ESs is caused not by sampling error, but by true heterogeneity among studies. In our seven-study case, $F=68.63$ or 68.6% of the variability among them is due to heterogeneity across the underlying studies.

2.1.2.1.2 Meta-Regression

If one finds significant heterogeneity with a significant $Q$ statistic and fairly large $F$, say greater than 0.55, one may use what is referred to as a “meta-regression” to determine significant variants or causes of among study variability [15]. Such sources of variability are given in the previous paragraphs (for example, patient characteristics, baseline disease severity, management of inter current outcomes or toxicity, etc.), and are sometimes referred to as covariates. The meta-regression approaches model the heterogeneity among study
treatment effects. Simply stated, the dependent variable is the ES, such as OR’s, RR, standardized mean differences, etc. from each study and is regressed on the covariates of interest. The suspected covariates and their strength of association with the differences in ES is determined with usual regression statistics such as the estimated regression coefficients or the $\beta$-values, the $p$-values, and the $R^2$-squares. Those covariates which are determined to have a statistically significant association with the dependent variable, ES, are assumed to be the major causes of heterogeneity. The assumption in using a meta-regression is that all the variables of interest suspected of being the sources of heterogeneity are available in all publications so they can be included in the meta-regression. Absence of data about important covariate variables in the literature can be a major limitation in assessing heterogeneity in meta-analyses with meta-regression. For example, in one study or publication, one may know the average age of individuals on each treatment or intervention, while in another study, only age groupings are given, and those groupings may be so restrictive as to not permit extrapolation of a meaningful average across treatment groups.

Other issues with meta-regression have been noted as well [21]. That is to say, the associations derived from meta-regressions are observational, and have a less rigorous interpretation than the causal relationships derived from randomized comparisons. This applies particularly to our example above when averages of patient characteristics in each study are used as covariates in the regression. Data dredging is the term used to describe the main shortcoming in reaching reliable conclusions from meta-regression. It can be avoided only in the strict statistical sense by prespecifying the covariates that will be investigated as potential sources of heterogeneity. However, in practice, this is not always easy to achieve as one has no guarantee that this information will be available in a reasonable form for coding before undertaking the literature search or systematic review. Only in the case of doing a meta-analysis of the raw patient-level data, if available, may this be possible.

Further statistical issues of meta-regression are also at hand. In a meta-analysis, the unit of analysis is the study, so the regression performance is determined by the number of studies in the meta-analysis. Thus the power of the regression is limited, and in most cases, as in the meta-regression from the literature, one could not expect much power depending on the number of covariates one considers in the model. Also with a restricted sample size, investigations of confounding interactions could be a challenge. One could derive weights for a weighted regression by study size. However, the usual inverse variance technique used for weighted regression would be difficult. Also, in the case of a meta-regression, most of the regressions performed use linear regression with no attempt to check the underlying assumptions of linearity or normality. Despite all these challenges, if one is determined to get a handle on the sources of heterogeneity, meta-regression is the tool of choice particularly with prespecified covariates. Again, the clinician’s or investigator’s substantive insight into the domain of the research can significantly enhance methodological planning for the systematic review and subsequent meta-analysis.

### 2.1.2.1.3 Funnel Plots to Measure Bias

Another useful diagnostic tool in examining meta-analyses is the funnel plot. The underlying studies’ effect estimates are plotted on the horizontal axis against sample size in ascending order on the vertical axis. These may be useful in assessing the validity of meta-analyses. This is often referred to as “publication bias.” The funnel plot is based on the fact that precision in estimating the underlying treatment effect will increase as the sample size of the included studies increases. Results from small studies will scatter widely at the bottom of the graph. The spread should narrow at the top of the graph among studies with larger sample sizes. In the absence of bias, the plot will resemble a symmetrical inverted funnel. Conversely, if there is bias, funnel plots will often be skewed and asymmetrical. Funnel plots are usually constructed as plotting the ES vs. the sample size or the ES vs. the standard error.

In meta-analysis, we can use a linear regression approach [8] to measure funnel plot asymmetry on the ES. The ES is regressed against the estimate’s precision, the latter being defined as the inverse of the standard error. The regression equation is: $ES = a + b*precision$, where $a$ is the intercept and $b$ is the slope. If the regression line passes through the origin, the intercept $a = 0$, then the funnel plot is symmetric. The null hypothesis is $H_0: a = 0$ which assumes symmetry of the effects about the null value of 0 in the case of
continuous data. Since precision depends largely on sample size, small trials will be close to zero on the horizontal axis and vice versa for larger trials. If the null hypothesis is rejected, the intercept of the regression line is different from 0 indicating bias in the relationship between precision and ES. We demonstrate this concept for our seven studies in which the funnel plot is not really very informative. Although symmetrical (see Fig. 2.2 which is the standard error on the vertical axis and the ES or “std diff in means” on the horizontal), it is rather flat because there are few studies with approximately the same sample size. However, the symmetry is preserved as shown by the plot and the regression results. The test of the null hypothesis that the intercept is equal to zero is not significant at the 0.05 alpha level indicating that the line goes through the origin. See the vertical mid line on Fig. 2.2. The actual $p$-value is 0.5194 indicating that there is not any bias in the relationship between ESs and precisions. Thus there is no publication bias. However, a limitation of this test is that it may not be accurate for small sample sizes, i.e., small number of studies in the meta-analysis. Figure 2.3 is a funnel plot of 14 studies addressing the same issue as our seven study plot, and one can see that with enough studies, the plot does take a nice funnel shape indicating no bias. Detailed discussions of funnel plots and test of symmetry can be found throughout the literature [3, 8] pertaining to meta-analyses.

### 2.1.2.2 Categorical Data

We next consider a second type of meta-analyses i.e., when data are categorical. This can involve statistics such as ORs, proportion of response, or relative risk ratios. In Table 2.2 are the results from six clinical studies comparing an intervention A with a control group B noted by the subscripts. $O$ represents the observed number of responses in the treated, A group, and the control, B group. $N$ represents the total sample size in each group, A and B. The odds ratio $OR= [O_A / (N_A - O_A)] / [O_B / (N_B - O_B)]$ or the odds of response vs. no response in group A vs. group B is noted in the sixth column. For example, in study 1, one interpretation could be that the odds of a responses vs. a nonresponse is 1.3 times more likely in group A than in group B showing superiority of intervention A to control B. We see that in studies 1, 2, 3, and 6, there is superiority of A to B with respect to achieving a higher odds of response and the opposite is true in studies 4 and 5. The $p$-values for each study is seen in the last column and the overall $p$-value using the Mantel–Haenszel statistic is 0.001, showing superiority of A over B with respect to the endpoint. Recall that the null hypothesis being tested in this context is that the $OR=1$ vs. the alternative hypothesis where the $OR\neq1$. The overall $OR=1.223$ with 95% confidence limits $(1.094, 1.366)$. The $p$-value=0.001. Figure 2.4 is the Forest plot with 95% confidence intervals on the ORs. Notice the larger rectangle for study 6 indicating a larger sample size than the other studies.

![Funnel Plot of Standard Error by Std diff in means](image-url)

**Fig. 2.2** Funnel plot of the 7 LDL studies
Just as in the previous case, not only do we have the treatment comparison effects, but also we can generate the other statistics testing for heterogeneity and publication bias. The \( Q \) statistic is 8.249 with a \( p \)-value = 0.143 indicating no heterogeneity. The \( I^2 \) statistic is 38.389 indicating that about 38% of the variance in heterogeneity is due to the between-studies variation. The funnel plot is not shown here. However, Egger’s test for symmetry yields a \( p \)-value = 0.019 indicating possible publication bias. Examining the statistics (not shown here) indicates that the larger studies 1, 2, 3, and 6 do in fact have smaller standard errors than do the smaller studies, 4 and 5. Thus this test of symmetry may not be very informative in this case as we mentioned above, as the sample size (number of studies) is rather small.

**Table 2.2** Six clinical studies involving categorical data

<table>
<thead>
<tr>
<th>Study</th>
<th>( O_A )</th>
<th>( N_A )</th>
<th>( O_B )</th>
<th>( N_B )</th>
<th>OR</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>392</td>
<td>512</td>
<td>358</td>
<td>498</td>
<td>1.277</td>
<td>0.090</td>
</tr>
<tr>
<td>2</td>
<td>286</td>
<td>483</td>
<td>249</td>
<td>482</td>
<td>1.358</td>
<td>0.018</td>
</tr>
<tr>
<td>3</td>
<td>164</td>
<td>427</td>
<td>297</td>
<td>848</td>
<td>1.157</td>
<td>0.235</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>152</td>
<td>36</td>
<td>137</td>
<td>0.808</td>
<td>0.439</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>154</td>
<td>51</td>
<td>172</td>
<td>0.724</td>
<td>0.202</td>
</tr>
<tr>
<td>6</td>
<td>334</td>
<td>1107</td>
<td>274</td>
<td>1116</td>
<td>1.328</td>
<td>0.003</td>
</tr>
<tr>
<td>Overall</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.223</td>
</tr>
</tbody>
</table>

**Fig. 2.3** Funnel plot of 14 studies showing symmetry

**Fig. 2.4** Forest plot of the six studies of Table 2.2
2.1.2.3 Time to Event Data

Figure 2.5 is a survival plot comparing two interventions, A and B. The vertical axis is the probability of remaining free of the event or failure and the horizontal axis is the time. Often, data from survival studies, generally known as the time to event analyses, are presented as survival curves as pictured in this figure. One can see in this study that the curve for intervention, A, is above the curve for B indicating a slower rate of failure over time for A as compared to B. In order to do a meta-analysis on aggregate data such as this, one has to know the hazard ratio (HR) or the log(HR), where HR is the ratio of hazards. The hazard function is a function associated with a survival curve or experience. The hazard function is basically the failure (death) rate over time. It can be thought of as the odds of failure over time. It is the chance of failure at any point in time. Just as the hazard function is a failure rate on a particular treatment, the HR is a measure of relative treatment effect. For example, suppose we have two groups (A or B), \( H_A \) is the hazard associated with A and \( H_B \) is the hazard associated with B. In Fig. 2.5, we wish to compare their survivals. Our null hypothesis is that the two survivals are not different or the hazard ratio, \( HR = 1 \) i.e.,

\[
H_0: HR = H_A/H_B = 1 \quad \text{(null hypothesis)}
\]

\[
H_1: HR = H_A/H_B \neq 1 \quad \text{(alternative hypothesis)}
\]

Thus the \( HR = [\text{Failure rate (new treatment)}/\text{Failure rate (control group)}] \).

If the \( HR < 1 \), then the outcome favors treatment A. This indicates that A has smaller hazard or chance of failure over time as compared to B. If \( HR > 1 \) then the outcome favors treatment B. Thus \( HR = 1 \) implies no difference between treatment and control groups. Meta-analysis combining several time-to-event studies determines whether treatment A is superior to its comparison group. If we do not have the raw patient data, an aggregate data approach may be required for some or all included studies. With meta-analysis of randomized controlled trials with time-to-event outcomes, the challenge is that there may be variation in the reporting of survival analyses [23]. Often, no single method for extracting the log(HR) estimate will suffice. Methods have been described for estimating the log(HR) from survival curves [23]. These methods extend to lifetables. In the situation where the treatment effect varies over time and the trials in the meta-analysis have different lengths of follow-up, heterogeneity may be evident. In order to assess whether the hazard ratio changes with time, several tests have been proposed and compared [23]. All issues discussed above for the OR applies to the hazard ratio. One can test for treatment effect, heterogeneity, and publication bias. As mentioned, one of the greatest sources of heterogeneity would be the differing length of follow-up across the studies included in the meta-analysis.

2.1.3 Other Tests in Meta-Analysis

In Sect. 2.0 above, we have discussed the background and the major tests for determining the overall treatment effect and its significance as well as heterogeneity in meta-analysis. There are other topics that one can pursue which are informative and will point out precautions that one must take in discussing meta-analyses.

2.1.3.1 Test for Interaction

One possible test is that of interaction. This involves grouping studies according to some characteristic of the study; for example, is the treatment effect homogeneous over the groups such as age categories, modalities, etc.? Let us suppose for the six studies of Table 2.2, that we believe the treatment effect to be nonhomogeneous over the age groups in our studies. In particular, we might suspect that the treatments may perform
Overview, Strengths, and Limitations of Systematic Reviews and Meta-Analyses

27

2.1.3.1 Test for Trend

Another important test to examine meta-analytic results is testing for a trend in the treatment effect. The technique is to split the data of each study according to some ordinal characteristic of the subjects which is suspected to influence the treatment effect; e.g., age, nodal status, disease stage, or severity, etc. Is there a systematic tendency of the treatment effect to increase/decrease with the severity of disease? Again, unless one has the raw data, or the ordinal categorization is obvious from the aggregate or published results, this may be impossible to do.

2.1.3.2 Sensitivity Analysis and “File-Drawer Effect”

One of the challenges in a meta-analysis is finding all of the studies on the subject. A critical issue in meta-analysis is what has been dubbed “the file-drawer effect.” That is, a study is conducted without a significant result, and it is not published. The time, effort, and resources necessary to publish the study are not expended particularly because it may be difficult to receive publication in a respected journal. It is very tempting for someone with inconclusive, nonsignificant data to quietly put it in a file drawer, intending to write it up when one gets some free time. Other more promising projects take priority.

The reason why publication bias is so important to a meta-analysis is that even if there is no real treatment effect, 5% of studies will show a significant positive result at the \( p < 0.05 \) level. That is, there is a 5% probability of getting a significant result if the null hypothesis is actually true. Counterbalancing inconclusive studies which would contribute to a more accurate overall ES estimate may be absent. Suppose one in fact performs a meta-analysis and achieves a significant effect, to evaluate the sensitivity of the result to publication bias, an estimate of how many unpublished, nonsignificant studies would reduce the overall effect to nonsignificance can provide some intuitive sense of the robustness of the result. Obviously, if the number of studies required to negate the observed effect is very large, you can be more confident that your meta-analysis’s significant treatment effect is not arising differently in the older than younger age groups. Assuming that we have the information to stratify the individual study results by age, we calculate OR comparing treatment A to treatment B in four age categories (Table 2.3). Note that overall, the results for Table 2.2 indicated superiority of A over B at \( p = 0.001 \). When the data was combined from all six studies in Table 2.3 and broken into age categories, we see that A is superior to B in ages 20–50 as the OR are all greater than one. Note that for age 50 or more, the OR is less than one, indicating superiority of B over A in the oldest group. In the last row of Table 2.3, we list the \( p \)-value for testing significant interaction of treatment with age which is \( p = 0.02 \) indicating significant interaction. This identifies a limitation of doing a meta-analysis on aggregate data. Unless one had the raw data from the six studies, or information on age categories and response was available in all 6 published articles, one would not be able to detect this potentially meaningful interaction. Hence, one might not discover that the overall result of the meta-analysis may not be true in a clinically important subgroup.

<table>
<thead>
<tr>
<th>Age category</th>
<th>Range</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20–30</td>
<td>1.23</td>
</tr>
<tr>
<td>2</td>
<td>31–40</td>
<td>1.04</td>
</tr>
<tr>
<td>3</td>
<td>41–50</td>
<td>1.09</td>
</tr>
<tr>
<td>4</td>
<td>51+</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Interaction – \( p = 0.02 \)

A meta-analysis was done on several studies to compare two interventions A and B. Was there a significant trend in response based on disease severity? We assume that we can group the data results from a number of studies by three classes of disease severity: low, medium, and high. When analyzing the data, we note that the OR of A vs. B (OR > 1) with respect to achieving a response overall was statistically significant, \( p \leq 0.05 \). When combining all the studies, we note a trend in the OR as 3.65, 1.72, and 1.05 for low, medium, and high disease severity respectively. Note the decreasing effect of A as the severity increases. A test for trend was significant at \( p = 0.01 \) indicating a statistically significant trend in the effect of A, with greater effect in less severe cases.
from the “file-drawer effect.” For example, in Table 2.2, at least 10 more studies with a sample size of at most 200 with nonstatistically significant results would be required to reduce the observed result to nonsignificance. Hence, we can be pretty confident that the results are not highly sensitive to publication bias. Sensitivity analysis, therefore, helps determine the robustness of the meta-analytic result.

Another form of sensitivity analysis is to reexamine the result when one or more studies are excluded. The reasons for doing this may include suspicion that one or more of the studies has undue influence on the overall result. Therefore, the question may arise as to the effect of certain studies in the meta-analysis that one believes may be over influencing the result, and if they were removed, would the effect be diminished? For example, in a meta-analysis of a longitudinal or time to event outcome, omit those studies with less follow-up than others to determine if the results still hold. Additionally, one may wish to perform a sensitivity analysis where studies judged to be of lower quality are excluded. Strategies have been employed to weight the studies by the presence of prespecified characteristics generally associated with optimal study design and conduct [5].

### 2.1.4 Statistical Techniques to Account for Study Size and Quality

In medicine and dentistry, one needs to properly weight all the evidence available to make sound principled decisions. Similarly, the studies in a meta-analysis can and should be weighted by study size (inverse variance) and perhaps other factors such as study quality.

#### 2.1.4.1 Weighing of Evidence

Now that we have discussed what a meta-analysis does and some of the issues involved, we are keenly aware of the fact that not all studies will behave in the same way. That is, they are of different sizes, perhaps different precision of the estimates of the effects across the studies and differing amounts of information such as subject characteristics from study to study. The ultimate goal of the meta-analysis is to calculate a “weighted average” of the effects of interest such as OR, risk ratio, hazard ratio, or standardized mean effect across the studies.

While there are various potential measures of ES, the overall goals are the same:
- Calculating a “weighted average” measure of effect
- Performing a test to see if this estimated overall effect is different from the null hypothesis of no effect.

The general concept is simple: determine the average effect across studies and assess its significance with a p-value. The mathematics is also straightforward.

The main challenge is to choose the type of weights to assign. By far, the most common approach is to weight the studies by the “precision” of the estimates. Precision is largely driven by the sample size and reflected by the widths of the 95% confidence limits about the study-specific estimates. Another way of considering precision is the variance of the effect within the study. The larger the variance, the less precise the estimate of the effect (i.e., OR, risk ratio, hazard ratio, or standardized mean effect). The smaller the variance, the more precise the estimate of the effect. The precision is defined in several ways, but has the same impact. The precision is sometimes defined as the inverse of the variance, inverse of the standard deviation, or inverse of the standard error of the mean. In any case, the smaller the variance, standard error, standard deviation, etc. usually as a result of a larger sample size within a study, the greater the precision.

When weights are assigned by the precision of the estimates, they are proportionate (1/(variance of the study)). This is the “statistician’s” approach. A possible problem, however, is that it could assign a bigger weight to the large and poorly-done study than it does to the small and well-done study. A meta-analysis that includes one or two large studies is largely a report of just those studies. However, for the most part, this is a reasonable assignment of weights. So, if we assume that we have chosen our studies well, we want the larger study to impact more on the meta-analysis than a smaller study. Some authors have assigned weights to studies based on the study quality or the amount of information provided about the study design, analysis, and study presentation. This is a scoring method and the studies with the most detail receive a higher score.
which is factored into the weighting \([5]\). The most widely used is the Jadad scale (Oxford quality scoring system) that assigns a score of 0–5 based on the study characteristics of randomization, blinding, and withdrawals/dropouts \([14]\).

### 2.1.4.2 Random Effects

In Sect. 2.1.2, we discussed the \(Q\) statistic as a measure of the heterogeneity across studies in a meta-analysis. Heterogeneity in the underlying studies can be a challenge when interpreting the results of a meta-analysis. The diverse nature of the patient populations, designs, and overall methods employed in the underlying studies can lead one to question the applicability of the results to a specific individual or group of patients. Also, the confidence interval surrounding the estimated overall effect, if derived from just the variance within each study, is too small when heterogeneity is present across the studies. Because of the varying sample sizes and underlying patient populations, it has been pointed out \([6]\) that each study has a different sampling error associated with it. The inverse variance weights defined above account for the within study variance but do not account for the between studies variation. With these fixed-effects methods of meta-analysis, one of the assumptions is that there is no variation between studies. However, if all the studies are not sampling from the same parent population, there may well be additional variance between the studies. In other words, one possible explanation for the heterogeneity of ESs among the studies is that each individual study in a meta-analysis may be sampling from a different parent distribution.

To allow for the randomness between the studies, we want to account for both the within and between study variation. Unlike fixed-effects models, random effects meta-analysis methodologies account for both within and between study variations. This allows one to adjust for the heterogeneity among the studies and gives a more accurate interpretation of the results when integrating studies. Several authors \([6]\) have done an excellent job of describing this process and constructing what is known as a random effects model. The convenience of the random effects model is that we need not worry about heterogeneity and discussion of the \(Q\) statistic and \(F\) with regard to the estimation of the confidence interval around the overall ES. The additional between study variance has been adjusted for in the calculations of the confidence interval surrounding the treatment effect. The estimate of the between study variance is calculated in the \(\tau\) statistic. The square of this statistic is added to the variance of each individual study variance before calculating the inverse variance \([1/(s^2 + \tau^2)]\).

This is not necessarily reassuring to some investigators who wish to know the degree of heterogeneity even before the random model is invoked. This does not necessarily prevent one from computing the \(Q\) and \(F\) statistic as described above in Sect. 2.1.2. Hence, many investigators will compute the results of the meta-analysis with a fixed-effects approach to explicitly and formally examine heterogeneity. If heterogeneity is found, typically a random effects model will then be invoked as the more sound estimate of the summary treatment effect and the surrounding confidence interval.

While there is a convenience to the random effects model, it generates wider 95% confidence intervals. This is true because the 95% confidence interval on the summary ES takes into account two sources of variation, the between study as well as the within study sampling variation. To demonstrate this, note from the seven-study example in Sect. 2.1.2 and Table 2.1 and Fig. 2.1, that the ES = −0.143 and the 95% confidence interval for the overall population ES is (−0.320, 0.034). For the random effects model, the ES = −0.154 and the 95% confidence interval for the overall population ES is (−0.476, 0.169). The results are the same in that there is no significant treatment effect, \(p=0.350\). However, the confidence interval in the random context is almost twice as large (1.8 times) as that for the fixed effects confidence interval. As another example, note the results of Sect. 2.1.2 for the categorical data in Table 2.2. The overall OR = 1.223 with 95% confidence limits (1.094, 1.366) for the fixed effects analysis. For the random effects analyses, the OR is 1.187 with 95% confidence interval, (1.019, 1.382). The results remain statistically significant \((p=0.028)\), but the confidence interval width in the random context is 1.33 times that for the fixed effects confidence interval. Given this, if no significant heterogeneity is detected between the ESs of the underlying studies, a fixed effects model may be appropriate. When a fixed effects assumption is appropriate, there is greater power to detect a significant overall treatment effect than with the random effects model.
2.1.4.3 Random Effects vs. Fixed Effects

The question naturally arises as to which is the better approach – a random effects model or a fixed effects model? One might determine that the more conservative approach is to consider the random effects since the confidence intervals are wider. However, an excellent discussion of this approach and its controversy has been put forth by several authors [18]. They point out that “conservative” in this context is generally taken to mean that random-effects approaches tend to have higher estimated variances than fixed effects models due to the inclusion or addition of the between study variation, which results, as we saw above, in wider confidence intervals. They point out that, when random-effects summaries are conservative in this context, they need not be conservative in other, equally important considerations. Specifically, random-effects ES can be farther from the null value and can have smaller \( p \) values, so they can appear more strongly supportive of causation or prevention than fixed-effects statistics or ES. The presence of this condition depends on the situation. For example, let us consider our two data examples in Sects. 2.2.1 and 2.1.2. Note that in Sect. 2.1.1, when considering the case of the continuous data summarized also in the previous paragraph, the confidence interval for the fixed effects is narrower than the confidence interval for the random effects. Also the ES of \(-0.154\) for the random model is farther from the null value of 0 than is \(-0.143\) for the fixed effects model. The \( p \)-value for the fixed effects is 0.114 and for the random model is 0.350. Here, the \( p \)-value of the random model is not smaller than the fixed effects model or more supportive of causation despite the larger estimated ES with the random model. Also in Sect. 2.1.2 for the OR example, the ES for the fixed effects is 1.223 with a \( p \)-value=0.001 and the ES for the random effects is 1.187 which is closer to the null value of 1 and a \( p \)-value=0.028. All this indicates is that the random effects model is not always more conservative than the fixed effects model when one not only focuses on the confidence intervals and \( p \)-values, but also examines the magnitude of the ESs [18].

Usually, the presence of heterogeneity among studies motivates employing the random effects design. While the random effects approach is a good mathematical way of accounting for the variation across studies, the researcher should still be aware of and investigate the potential sources of heterogeneity. It may not always be possible to determine or explain the heterogeneity, but every attempt should be made to examine it, explain it, and determine if proceeding with the meta-analysis is appropriate. An assumption is that the characteristics of the studies that generated the results allow those results to be combined [18]. Some may question that assumption and after examining those characteristics decide that it may not be reasonable scientifically or clinically to combine those results. Hence, rather than being mutually exclusive, examining fixed effects, random effects, and tests of heterogeneity are all a part of the total reporting of a sound meta-analysis.

Another possible way of examining the data accounting for the randomness among the studies is by conducting a Bayesian meta-analysis.

2.1.5 Bayesian Meta-Analysis

There are times when one has a prior belief about the performance of therapies before the trial or study begins. Thus it is assumed that their performance on an average is random. This is also the case before proceeding with a systematic review to gather studies for a meta-analysis. Just as meta-analysis is the technique used to combine the information from similar trials into a formal summary, Bayesian analysis takes it one step further and combines prior information or prior beliefs with the meta-analysis. These beliefs are called the prior distribution. The term “Bayesian” comes from Bayes Theorem which is a simple mathematical formula for computing conditional probabilities. Thomas Bayes was an 18th century English clergyman [4].

For example, in our data in Table 2.2 from Sect. 2.1.2, one may believe before starting the study that the gains in odds of response on A is 20% better than that on B and that this is a random component with a reasonable probability distribution. The evidence from the trial or data from the literature is described as the sampling distribution of results. Combining the prior belief and the evidence in the data gives the posterior distribution of beliefs. This gives an estimate of the actual OR or what is termed the posterior OR. Differences from the analysis done in Sect. 2.1.2 or classical analyses include the incorporation of prior beliefs and the absence of \( p \)-values. The posterior estimate of the effect or OR is via a credibility interval which is analogous to a
classical point estimate of the OR and its associated confidence interval, but has a direct interpretation in terms of belief. As many people interpret a confidence interval as the region in which the effect probably lies, they are essentially acting as Bayesians.

Table 2.4 and Fig. 2.6 give the results of the Bayesian application to the data in Table 2.2. The prior belief incorporated into the analysis was that the prior odds of A in favor of B is random with a probability distribution and ranged from 0.9 to 1.26. We omit the details of the underlying distribution as it is not necessary for us to proceed with the overview of the procedure. Combined with the data in Table 2.2 from the columns titled, \( O_A, N_A, O_B, N_B \), we derive the posterior OR in Table 2.4 for each of the 6 studies. That is, the OR column of Table 2.2 updates the prior information to yield the “Posterior OR” column in Table 2.4. The columns titled Lower (2.5%) and Upper (97.5%) for studies 1–6 are the lower and upper limits respectively for the 95% posterior credible regions of the posterior OR from each of the studies. Note that in Table 2.4, studies 3, 4, and 5 have a lower limit below 1 indicating possibly no difference in the two treatments. Studies 1, 2, and 6 have a lower limit above the value 1 indicating superiority of A to B. The last row of Table 2.4 is the overall Bayesian meta-analytic posterior OR with the 95% credible regions. Note that the posterior OR = 1.195 with a lower limit equal to 0.972 or very close to one. This indicates possible evidence from the combined six studies for no difference between the two treatments. However, note Fig. 2.6, which is the posterior density of the OR from the combined studies. We note that the bulk of the density is above 1, favoring treatment A. The feature of the Bayesian approach is that now the parameters of interest, such as the ESs, are random and have a distribution which allow us to determine the credibility of certain values of that parameter with some probability. No \( p \)-value is generated confining us to a region of 0.05 or less. The interpretation is based on the credible interval or region and the posterior density distribution.

Also in Fig. 2.6, the label odd R stands for posterior OR from the meta-analysis and the number 10,000 is a feature of the software we used indicating the number of high speed iterations used to converge to our solutions [20]. We used much more than actually needed for this particular application. We note that the Bayesian computations can be computer intensive, but the availability of the software to perform these calculations makes it routine in most cases [20].

Another feature of the Bayesian analysis is that one can derive a predictive distribution. That is, if a new study were to be done comparing A and B, then based on our results, we can predict what the OR would be and it’s credible region. For our data, the predictive value of the OR is 1.227 with a credible region of (0.762, 1.761) indicating some evidence in favor of no treatment effect.

The Bayesian analysis is very attractive in that one can incorporate prior information into the meta-analysis and update this information with the new sample data from studies gathered. Also, should more articles addressing a similar question be published in the future, one can then use the posterior distribution from the present meta-analysis as the new prior distribution for the next meta-analysis. For example from Table 2.4, the new prior density region for the OR for the next meta-analysis could be in the range of 0.972–1.41. There are no statistical multiplicity issues involved in terms of \( p \)-value adjustment for repeated significance testing since there are no \( p \)-values to consider. The major criticism of the Bayesian approach is that one
has to specify a prior distribution and that it may be too subjective. Such need not be the case as one can always use previous data from previous studies or the literature to empirically construct a prior distribution.

2.2 Discussion

There are many advantages to doing a meta-analysis. They include, but are not limited to the following:
1. One is able to detect small but clinically relevant effects of an intervention.
2. Considering the heterogeneous nature of subjects across studies, the more diverse groups of subjects across studies hopefully permits a more global solution to a question or hypothesis. The conclusion has the potential advantage of being based on most, if not all, existing evidence from relevant studies.
3. One can avoid the time and expense of doing a larger confirmatory study if the evidence from many studies is available, properly handled statistically, and the meta-analysis provides a sufficiently conclusive result.
4. Although most meta-analyses are done from published results, the more powerful results are garnered from actual subject data. However, for many studies the data are too old, too expensive to retrieve, or cannot be retrieved. A meta-analysis of aggregate data may be all that is possible.
5. The statistical tools are such that most sources of bias and heterogeneity can be statistically examined. Sensitivity of statistically significant results can be measured.
6. The statistical techniques to analyze results from meta-analyses are generally not new and many standard techniques apply.
7. The resources to help accomplish a systematic review of studies to be included in the meta-analysis are available.
8. The statistical software to perform meta-analyses are available, but should not be used without appropriate understanding of the underlying methods and limitations.

All of these latter issues are addressed statistically and certainly do not preclude one performing meta-analyses. One should certainly be aware of these issues. The attempt has been made here to demonstrate most of the strengths and limitations of meta-analyses by way of definition and application.
generalization and application of results. Typically, the statistical tools at hand are certainly adequate for addressing these issues. There are good online sources which provide guidelines for conducting a meta-analysis. These include CONSORT, QUORUM, and MOOSE. These can be entered as key words to locate the resource. The Cochrane Collaborative mentioned above in section 1.0 provides excellent guidelines as well. A good source for arranging onsite training for nonstatistically oriented investigators with an overview for understanding and conducting meta-analyses can be found by contacting QuantitativeAppl@aol.com. However, one should keep in mind that meta-analyses should neither be a replacement for well-designed large-scale randomized studies [2] nor a justification for conducting small underpowered studies. It is a tool which, when properly utilized, helps one to arrive at a reasonable and defensible decision from the scientific information already presented.

References

Evidence-Based Practice: Toward Optimizing Clinical Outcomes
Chiappelli, F.; Caldeira Brant, X.M.; Neagos, N.; Oluwadara, O.O.; Ramchandani, M.H. (Eds.)
2010, XIV, 251 p. 50 illus., 30 illus. in color., Softcover
ISBN: 978-3-642-05024-4