In a previous article in this series, we looked at systematic reviews – their methods, uses, and limitations. Systematic reviews are frequently followed by the use of a mathematical or statistical technique to pool the data from individual studies included in the review, to obtain a single summary measure of effect. This technique, known as meta-analysis, is the focus of this article.

DEFINITION

The term “meta-analysis” was first defined in 1976 by Glass as “The analysis of analyses ….. The statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.” A more recent definition of meta-analysis is: “a statistical analysis which combines or integrates the results of several independent clinical trials considered by the analyst to be combinable”.

REQUIREMENTS FOR A META-ANALYSIS

The performance of a meta-analysis is contingent upon a good-quality systematic review. We discussed the steps involved in conducting a systematic review in a recent article. If the systematic review reveals that there are an adequate number of studies with shared characteristics, then the reviewers may consider performing a formal meta-analysis. The systematic review process ensures that the meta-analysis includes data from all the available studies that fulfill certain selection and/or quality criteria, on the question being studied, in an unbiased manner. If this is not done, then the results of the meta-analysis may be biased and hence unreliable.

UNDERSTANDING HETEROGENEITY

The term “heterogeneity” refers to variability between the results of different studies on a research question. On any question, results of the available studies are unlikely to be identical, and a certain amount of variation between these is expected just by chance. This variation, or statistical heterogeneity, depends on two factors. One of these is the sample size, with smaller-sized studies showing a greater variation. The second relates to variability of the outcome variable, e.g., the random variation being larger if the event
rate is very low or very high, and smaller if the event rate is close to 50% for dichotomous variables.

In addition, studies on a particular research question included in a systematic review often differ from each other somewhat in population characteristics – e.g., age and gender distribution, ethnicity, baseline severity of disease, and presence of comorbid conditions – these are examples of clinical heterogeneity. Furthermore, the studies may have methodological heterogeneity, i.e., variations in dose, frequency, or route of administration of a drug used, use of different drugs for a particular condition (e.g., studies assessing the efficacy of beta-blockers in treating portal hypertension may use a variety of beta-blockers such as propranolol or nadolol), blinding techniques (open, single-blind, or double-blind), tools or exact measures used to evaluate the outcome (e.g., blood pressure may be measured by intra-arterial or noninvasive techniques), time points used for outcome assessment, etc. Hence, the studies included in a meta-analysis would be expected to show a greater variability than may be expected by chance alone.

The presence, the degree, and the nature of heterogeneity in methodology and results of the available studies influence the decisions about whether the studies can be combined using a meta-analysis and about the statistical tools to use. The available studies must be sufficiently related for those to be pooled. This is not a statistical decision but needs a careful evaluation by experts in the subject area of the research question. If there is obvious clinical and/or methodological heterogeneity, it is appropriate to not proceed with the meta-analysis. Once a decision to proceed with meta-analysis is taken, a more formal assessment of heterogeneity is done.

TESTS FOR HETEROGENEITY

Inter-study heterogeneity can be formally assessed. This is most often done using one of the two tests, namely the Cochran’s Q test and the Higgins $I^2$ test.

The Cochran’s Q test (also known as the Chi-square test for heterogeneity or the Chi-square test for homogeneity) looks at whether the results of individual studies differ from the expected average effect, i.e., differ by more than what would have been expected by chance. If the Cochran’s Q test is positive (shows a statistically significant result or low P value), then the heterogeneity between the studies exceeds the random expectation. This test is limited by the fact that its result is heavily dependent on the number of studies; if this number is very small or very large, the test tends to under- or overestimate heterogeneity, respectively.

Furthermore, it does not provide a quantitative measure of the extent of heterogeneity.

The Higgins $I^2$ test provides a numeric value, known as the $I^2$ statistic, for the degree of heterogeneity between studies beyond what would be expected by chance. It can vary from 0 to 100% (or 0 to 1.00), with lower values indicating less marked heterogeneity. This value represents the proportion of total variation across studies that is due to heterogeneity rather than chance. Higgins also specified empiric cutoffs of 25%, 50%, and 75% to indicate low, moderate and high heterogeneity respectively.[4]

HETEROGENEITY AND CHOICE OF MODEL: RANDOM-EFFECTS VERSUS FIXED-EFFECTS

In a meta-analysis, results of individual studies are combined to produce a single overall result. This does not imply that the number of events and subjects can be simply aggregated across studies. Instead, a statistical method is used to obtain a summary measure. Further, since the included studies vary in sample size, event rate, and individual results, these cannot be given equal importance. Thus, each study is given a different weightage, depending on its characteristics, with a study with more stable value of the outcome measure being awarded greater weight.

Broadly, there are two types of methods to pool data in a meta-analysis – the fixed-effects and the random-effects models.

The fixed-effects model assumes that all the studies in a review have a single true effect, and that any variations between the results of these studies represent a random error, which is largely a reflection of their relative sample sizes and the observed event rates. Therefore, the fixed-effects model assigns greater weightage to larger studies and those with nearly equal rate of events and non-events, and less weightage to smaller studies and those with either too few or too many events. For the fixed-effects model, at least three different calculation techniques are available to assign weightage to studies – the inverse variance technique, the Peto odds ratio, and the Mantel-Haenszel method, and any of these may be used.

The random-effects model assumes that the studies included in a review are drawn from somewhat different populations of studies, with slightly different treatment effects. In this model, larger studies are given proportionately less weightage, whereas smaller studies are given proportionately larger weight than in the fixed-effects
model; thus, study weights are more similar under this model. The DerSimonian and Laird method is the most commonly used random-effects technique.

The confidence intervals of the summary effect obtained using the random-effects model are wider than those for the fixed-effects methods.

There is little consensus on which of the two models to use. Although it appears tempting to choose between these based on the measure of heterogeneity, this is not recommended, and the clinical and methodological heterogeneity also need to be taken into account, with the random-effects model being used if one believes that there is sufficient heterogeneity. Some authors recommend that one should always analyze the data and report results using both the models. Others believe that it is safer to always use the random-effects model.

**CHOICE OF EFFECT MEASURE**

The outcome data being compared and pooled can be of different types. The main types include dichotomous data, where each individual’s outcome is one of only two possible categorical responses (e.g., cured or not cured); continuous data, where each individual’s outcome is a numerical quantity (e.g., weight gain in kg); count and rate (e.g., number of events per unit time, such as number of episodes of diarrhea per year) data; or survival data (time until an event of interest occurs).

For dichotomous data, the effect of an intervention (i.e., the difference in outcomes between the treated and the control group) can be represented as risk ratio, odds ratio, or risk difference. For continuous data, the treatment effect is most often represented as mean difference (or the difference between the means in the treatment and the comparator group). An alternative is the use of standardized mean difference (calculated as the mean difference between the groups divided by the standard deviation of data for all participants). This is nothing but the mean difference expressed using the standard deviation as a unit. Its use allows one to pool studies which measure the same outcome using different scales (e.g., improvement in depression using different psychometric scales).

For rates and survival data, the effect measures used are rate ratio and hazard ratio, respectively.

![Figure 1: A typical forest plot showing its various components. At times, some components (e.g., the columns with original data from the included studies) are not included, or further details are added (Figure adapted from Aggarwal et al).](image)
THE FOREST PLOT

The results of a meta-analysis are depicted using a graphic known as a forest plot. A forest plot includes identifiers for individual studies included in the analysis, the results of each study in brief and the weightage given to each study, and provides a visual representation of the degree of heterogeneity between the results of individual studies.

As an example, Figure 1 shows the forest plot of a meta-analysis of studies comparing zinc supplementation with placebo for the prevention of diarrhea in healthy children.\(^\text{[1]}\) The first column lists unique study identifiers (usually the last name of the first author and the year of publication). The next four columns depict the results of individual studies – in this case, the number of diarrheal episodes and cumulative years of follow-up in each study arm, i.e., the intervention arm and the comparator. For other types of studies, the data reported may include means and standard deviations (for continuous outcomes) or the number of events and the number of participants (for binary outcomes). The results of each study are visually depicted in the forest plot, with the summary statistic (in this case, rate ratio) represented by a square and a horizontal line representing the confidence intervals for the summary statistic. The confidence intervals used most often are 95%, but another value (e.g., 99%) can be used depending on the author’s choice. The weightage given to each study is shown as a percentage of the total, and sometimes also as the relative sizes of the squares on the bar for each study.

The combined treatment effect, as determined by the meta-analysis, is shown using a diamond, with its center representing the overall summary statistic and the horizontal limbs (the horizontal spread of the diamond) depicting the confidence intervals. The model used to calculate the weightage (random-effects in this example) is mentioned in the header of the last column. This column also has numeric values for the summary statistics and their confidence intervals.

The results of tests of heterogeneity – the Chi-square test and the \( F \) test – are also reported. In this example, the Chi-square test was significant with a low \( P \) value, and the \( F \) value was 77%, both suggesting statistical heterogeneity and supporting the choice of the random-effects model.

The results of individual studies and of the meta-analysis are read in relation to a vertical line, which represents the “line of no effect,” i.e., a situation where the treatment does not lead to any change in the outcome measure. This line is drawn at the value of 1.0 in case of ratio measures (e.g., odds ratio, risk ratio, etc., as in this case) or at 0 in case of linear measures (e.g., mean difference or standardized mean difference). Any horizontal lines or diamonds which do not cross this line are deemed as statistically significant.

The above text provides basic information on a meta-analysis. The technique has several additional nuances, which we propose to deal with in the next article in this series.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

REFERENCES
1. Ranganathan P, Aggarwal R. Study designs: Part 7-Systematic reviews. Perspect Clin Res 2020;11:97-100.
2. Glass, Gene V. Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, Vol. 5, no. 10; 1976. p. 3-8. JSTOR, Available from: www.jstor.org/stable/1174772. [Last accessed on 2020 Aug 20].
3. Huque MF. Experiences with meta-analysis in NDA submissions. Proc Biopharm Section Am Stat Assoc 1988;2:28-33.
4. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539-58.
5. Ranganathan P, Aggarwal R, Pramesh CS. Common pitfalls in statistical analysis: Odds versus risk. Perspect Clin Res 2015;6:222-4.
6. Sedgwick P, Jockes K. Interpreting hazard ratios. BMJ 2015;351:h4631.
7. Aggarwal R, Sentz J, Miller MA. Role of zinc administration in prevention of childhood diarrhea and respiratory illnesses: A meta-analysis. Pediatrics 2007;119:1120-30.