CHAPTER THIRTEEN

Key Concepts

meta-analysis, effect size
selection bias:
  reporting bias

publishing bias
retrieval bias
Guides to a Critical Evaluation of Published Reports

In reading a report published in the literature, one often begins by reading the abstract or summary. While this step is important in that it quickly indicates whether the article is really of interest, its role must be kept in perspective. You must not yield to the temptation of accepting conclusions from the summary without appraising the merit and validity of the study itself. You must read the article critically before accepting its conclusions as being relevant to your research. Experience will improve your ability to evaluate research reported in the literature, but that ability will be best utilized if you approach your reading with a definite plan in mind. A wide variety of procedures are used to conduct, analyze, and report research findings, and so it is impossible to give a single set of hard and fast rules for evaluating all such reports. We have nevertheless compiled a few guidelines that you should find helpful to keep in mind as you read the literature.

THE RESEARCH HYPOTHESIS

A first step in reviewing any article is to identify the research hypothesis. Why was this research performed? Does it have relevance to you? Is there any practical or scientific merit to it? If not, there is no need to read any further.

VARIABLES STUDIED

When you have identified the research hypothesis, and before you read the report in detail, ask yourself what variables would shed light on the research hypothesis. Next, identify the variables included in the report. Which are the response variables? Which are the predictor variables? Are these variables relevant to the research hypothesis? Compare your list of variables with those included in the report. Were important – possibly confounding – variables overlooked? Was adequate information collected on all relevant concomitant variables? Age, race, and gender are three
important variables in many human studies. Ethnic origin, geographic location, and socioeconomic status are other variables that may be important.

The methods used to obtain the data are another important consideration. Were state-of-the-art clinical or laboratory techniques used? Do these methods produce precise measurements? The scale of measurement has an impact on the choice of statistical hypotheses to be tested and statistical methods for testing these hypotheses. Hence the scale of measurement used for recording each variable should be identified; some scales may be nominal, some ordinal, and some interval.

**THE STUDY DESIGN**

It is imperative to identify the study design because the appropriate methods of analysis are determined by it. Is the study experimental or observational? What are the study units and how were they selected? Are they a random sample, or were they chosen merely because it was convenient to study these particular units? What was the target population? How does it compare with the study population? If, as is so often the case in experimental studies, a convenient (rather than a random) sample of study units was used, it should be described in sufficient detail for you to have some feel for how general any of the conclusions are. If the study units are patients with some disease at a particular hospital, for example, you must decide how representative these patients are. Examine the inclusion and exclusion criteria for entry into the study. Could the fact that these are patients at that particular hospital somehow make them unrepresentative? For example, does that hospital specialize in the more severe cases? Do geographic location or climatic conditions have an effect? In other words, can you reasonably use the results of this study to guide you in your own work.

After the study units were selected, how was the study conducted? At what point was randomization used? Was an adequate control group included? Was double-blinding used to control observer bias?

**SAMPLE SIZE**

The greater the number of study units in an investigation, the more confident we can be of the results. Findings based on one or two subjects cannot be expected to be typical of large populations. The sample size must be large enough for the results to be reliably generalized. Statistical tests used to declare results significant or not significant depend upon the sample size. The larger the sample size, the more powerful the test and the more sensitive the test to population differences. Hence, if the sample size is enormous, even trivial population differences could be declared significant. In a small sample, on the other hand, a large difference may be nonsignificant. Especially in the case of a study that reports no significant
differences, it is necessary to determine how large a difference could reasonably have been detected by the sample size used. A good study that reports negative results will also quote the power of the statistical tests used. There are several websites that enable the calculation of appropriate sample sizes, e.g.

http://www.protocol-online.org/cgi-bin/prot/view_cache.cgi?ID=3864
http://www.sph.umich.edu/csg/abecasis/cats/
http://pngu.mgh.harvard.edu/~purcell/gpc/

**COMPLETENESS OF THE DATA**

Clinical studies invariably suffer from the problem that some patients have missing data for one reason or another. If an unusually large number of study units, say more than 20%, have a significant amount of incomplete data, then the credibility of the results should be questioned. Be careful of reports in which a portion of the data have been discarded as outliers because of gross errors or other unusual circumstances. If an investigator discards those data that do not support the research hypothesis, the scientific objectivity is lost. A good study will include all the data available in the various analyses performed.

**APPROPRIATE DESCRIPTIVE STATISTICS**

An overview of the findings of a study can often be gleaned by scanning the tables and graphs. Be watchful, however, that the tables or graphs are not misleading. Distinguish between the sample standard deviation and the standard error of an estimated parameter. If it is not reported, calculate the coefficient of variation for some of the important variables. If the relative variability is large (e.g., if the coefficient of variation is greater than 30%), then important population differences may be obscured by the ‘noise’ in the data. Be sure you understand what the numbers in the tables represent and exactly what is graphed. It sometimes helps (especially in a poorly written report) to reconcile numbers in the tables and graphs with numbers in the text.

**APPROPRIATE STATISTICAL METHODS FOR INFERENCES**

The names of statistical tests used in the analysis, such as Student’s $t$-test, paired $t$-test, analysis of variance, multiple regression analysis, and so forth, should be stated in the methods section of the report. Be wary of reports that state $p$-values without indicating the specific method used to analyze the data. In each case,
identify the specific null hypothesis that is being tested. Try to determine if the statistical methods used are appropriate for the study design, scale of measurement, etc. If the method of analysis is unfamiliar to you, consult a statistician. Each method requires certain assumptions of the data, such as independent samples, random samples, normal distributions, or homogeneous variances. Gross violations of these assumptions may bias the analysis. The report of a careful analysis will justify the use of each statistical test used, should there be any doubt. Remember that a \( p \)-value represents the chance that a difference in the sample data is the result of random variation, when in fact there is no difference in the populations from which the samples came. A \( p \)-value of 0.05 tells us there is a 5% chance that the observed difference (or a more extreme difference) could arise by chance if the null hypothesis is true. Suppose three independent statistical tests are carried out. If in each case the null hypothesis is true, the possibility that at least one of the tests results in significance at the 5% level is about 0.14 (\( = 1 - (1 - 0.05)^3 \)). If the three tests are not independent, then the probability is somewhere between 0.05 and 0.14. If \( c \) comparisons are made, independent or not, if we want to correct a quoted nominal \( p \)-value, \( p^* \), for the fact that \( c \) statistical tests have been performed, this can be done conservatively by multiplying it by \( c \); we can be sure that the appropriate \( p \)-value is less than \( cp^* \).

**LOGIC OF THE CONCLUSIONS**

Above all, remember that there is no substitute for evaluating the logic of the conclusions. A report that concludes it is safer to drive at high speeds because relatively few deaths from automobile accidents occur at speeds in excess of 100 miles per hour is clearly absurd. The frequency of deaths occurring at speeds in excess of 100 miles per hour needs to be related to the number of cars driven over 100 miles per hour, and this compared with some similar fraction for cars driven at slower speeds. Equally absurd conclusions, however, can be found in many research reports because improper – or no – comparisons are made. Always be on your guard against this type of fallacy; it is more common than you may suspect.

**META-ANALYSIS**

Finally, you should be aware of a special type of report, which is more and more commonly published in the health-care literature, called a *meta-analysis*. This refers to the statistical analysis of information from a series of studies carried out for the same general purpose by different investigators or, in some instances, by the same investigators at different times. The aim of a meta-analysis is to synthesize the
findings from different sources into an overall interpretation to guide practitioners. Thus, a meta-analysis attempts to combine all the available information on a given topic by pooling the results from separate studies.

From a purely statistical point of view, the most obvious benefit of a meta-analysis is that it effectively increases the sample size and, therefore, the power of the analysis to detect important group differences that may go undetected in small individual studies. The increased sample size will also lead to more precise estimates, and, because each study will have been performed under slightly different conditions or in somewhat different populations, a meta-analysis can give evidence for the generalizability of a particular result. However, the results of a series of investigations of a particular topic may differ, and we shall see that this may present either an opportunity or some difficulty in the overall interpretation. The combined analysis should include a critical evaluation of the design, analysis, and interpretation of the individual studies that are summarized. The subjects in the different studies may be heterogeneous, treatments may vary in dosage and compliance, and experimental skills and techniques may differ.

A meta-analysis begins with an effort to identify and obtain information on all studies relevant to the topic being investigated. The extent to which this can be done depends on the methods we use to search for published and unpublished studies and on our ability to identify clearly the focus of each candidate study. Three types of selection bias may occur. Reporting bias occurs when investigators fail to report results – for example, because they did not lead to statistical significance. Publishing bias occurs when journal editors decline to publish studies, either in whole or in part, because of lack of statistical significance. Retrieval bias occurs when the investigator conducting the meta-analysis fails to retrieve all the studies relevant to the topic of interest.

Any meta-analysis is necessarily limited by what is available in the studies retrieved. Sometimes individual investigators may be willing to share additional aspects of data when published results are lacking in detail. Studies may differ in design, quality, outcome measure, or population studied. They may vary from blinded, randomized, controlled trials to trials that do not use blinding, randomization, or controls. Criteria for including studies in a meta-analysis should be well defined before the search for candidate studies begins.

Studies done at different times or by different investigators are often taken to be statistically independent. However, although working separately, investigators of a specific topic may have similar backgrounds and prior beliefs, communicate frequently with each other, and modify later studies on the basis of earlier outcomes. Thus a meta-analysis should investigate time, location, and investigator effects on the outcome measure.

Before the results of studies are combined, there must be some assessment of the homogeneity of the results. It is expected that the results of different studies
will vary somewhat, but for investigations that are essentially similar, heterogeneity in the results must decrease our confidence in any conclusions. An appropriate meta-analysis will then require special statistical techniques to reflect this heterogeneity, by quoting a larger standard error for any estimates of how well a treatment performs. Alternatively, it may be that variation in outcomes across studies is the result of differences in treatment protocols; in this situation an opportunity may exist to determine the best treatment regimen.

Once it has been decided to combine the results of several studies to obtain a single overall result, it is necessary to choose whether to weight each study the same or differently. Some attempts have been made to weight studies by their relative quality, but this is subjective and very difficult to quantify reliably. Another approach is to conduct separate analyses for groups of studies of similar quality. Often, the best approach is to weight the studies according to sample size, with the larger studies weighted more heavily than the smaller ones. A good meta-analysis will use more than one weighting scheme, to be sure that the resulting tests and estimates are reasonably robust to which particular weighting scheme is chosen.

We have already seen a simple method of combining studies when we wish to perform an overall test of a null hypothesis; this is the method of combining $p$-values discussed in Chapter 9. Other methods are available and are discussed in the additional readings suggested at the end of this chapter. Analysis of variance techniques are often used to obtain pooled estimates, and these are similarly discussed in the additional reading. The important thing to remember is that if there is heterogeneity among studies, as indicated by a large interaction between treatments and the different studies, then this source of variation must be included in the standard error when computing an overall confidence interval for the treatment effect.

Finally, a concept often used in meta-analysis is that of effect size. This is defined for each candidate study as the estimated difference in mean outcome between those treated and the controls, divided by the estimated control standard deviation. This is often used to combine the effects of studies that measure different outcomes, such as might result from the use of different measuring instruments. Those studies using instruments that have larger measurement variability thus have less impact than those studies using more precise instruments.

**SUMMARY**

1. Do not accept conclusions solely on the basis of the abstract or summary of a published report.

2. Once you have determined the research hypothesis, determine whether all important relevant variables were studied, and whether they were studied appropriately.
3. Identify the study design and the study population to determine the relevance of the results.

4. Take account of the sample size and the completeness of the data, especially if the study reports that no significant differences were found.

5. Be sure you understand what the numbers in the text and tables represent, and what is graphed.

6. Try to determine if the statistical methods used are appropriate. If multiple significance tests are performed, multiply each \( p \)-value by the number of tests performed to obtain an upper bound for the overall significance level.

7. When reading a meta-analysis, note what steps were taken to reduce selection bias – reporting bias, publishing bias, and retrieval bias.

8. A meta-analysis must assess the homogeneity of the studies it includes. Criteria for inclusion should have been defined prior to beginning the search for candidate studies.

9. A meta-analysis can result in increased power, precision, and generalizability. But it should be demonstrated that the results are robust and do not depend critically on the particular weighting scheme used.

10. The effect size of a study – treatment average minus control average, divided by control standard deviation – is used to pool the results of studies that differ in measurement variability.

11. Whenever you read a report, whether a single study or a meta-analysis, ask yourself if the conclusions make sense.

FURTHER READING

Haines, S.J. (1981) Six statistical suggestions for surgeons. Neurosurgery 9: 414–417. (Some basic principles are given for interpreting statistical analyses in the medical literature. Catchy subheadings make this article enjoyable to read.)

Glantz, S.A. (1980) Biostatistics: How to detect, correct and prevent errors in the medical literature. Circulation 61: 1–7. (This article was referred to in Chapter 1; now is the time to read it.)


**PROBLEMS**

1. Read articles in your own area of interest critically, following the guidelines of this chapter.
2. Read and critique the design and analysis aspects of the studies in the following two brief papers:
