Chapter 9
Interpreting research results

9.1 Introduction

Researchers should describe their results clearly, and in a way that other researchers can compare them with their own results. They should also analyse the results, using appropriate statistical methods to try to determine the probability that they may have been chance findings, and may not be replicable in larger studies. But this is not enough. Results need to be interpreted in an objective and critical way, before assessing their implications and before drawing conclusions. Interpretation of research results is not just a concern for researchers. Health professionals reading or hearing research results should be able themselves to interpret them correctly, and to assess their implications for their work. Policymakers should also be aware of the possible pitfalls in interpreting research results and should be cautious in drawing conclusions for policy decisions.

9.2 Interpreting descriptive statistics

The mean or average is only meaningful if the data fall into a normal distribution curve, that is, they are evenly distributed around the mean. The mean or average, by itself, has a limited value. There is an anecdote about a man having one foot on ice and the other in boiling water; statistically speaking, on average, he is pretty comfortable. The range of the data, and their distribution (expressed in the standard deviation) must be known. It is sometimes more important to know the number or percentage of subjects or values that are abnormal than to know the mean.

Descriptive statistics cannot be used to define disease. The average should not be taken to indicate the “normal”. The standard deviation should not be used as a definition of “normal” range. To allow a cut-off point in a statistical distribution to define a disease is wrong. This is particularly important in laboratory data, where the range of normal is often based on measurements in a large number of healthy people. The standard deviation is based on the values in 95% of the apparently normal healthy people. Outlying values are considered abnormal though they do not indicate disease. With the modern tendency of using a large battery of laboratory tests for each patient, the likelihood of so-called abnormal values becomes high. For example, when 5% of each of 20 biochemical determinations in healthy people are routinely classified as deviant, the likelihood that
any non-diseased individual will have all 20 determinations reported as normal will be only 36% (Gehlbach, 1993). Graphs may distort the visual impression of relationships, if the scale on the $x$ and $y$ axes is put in different ways. An association or correlation does not mean causation. An association or correlation needs explanation. Because of the importance of this question, it will be dealt with in detail in another section in this chapter.

### 9.3 Interpreting “statistical significance”

Albert Einstein said, “Not everything that can be counted counts, and not everything that counts can be counted.” A statistically significant finding simply means that it is probably caused by something other than chance. Significant does not mean important.

To allow proper interpretation, exact $P$ values should be provided, as well as the statistical test used. The term “orphaned” $P$ values is used to describe $P$ values presented without indication of the statistical test used.

Statistical tests need to be kept in proper perspective. The size of the $P$ value should not be taken as an indication of the importance of the result. The importance of the result depends on the result itself and its implication. Results may be statistically significant but of little or no importance. Attaching a fancy $P$ value to trivial observations does little to enhance their importance. A statistically significant or even a highly significant difference does not necessarily mean a clinically important finding. A difference is a difference only if it makes a difference.

Differences may not be statistically significant but may still be important. The differences may be real but, because of the small size of the sample, they are not statistically significant. A $P$ value in the non-significant range tells you that either there is no difference or that the number of subjects is not large enough to show the difference. As discussed in Chapter 8, the study may not have had the power to show an effect of that size.

### 9.4 Bias

All studies are potentially subject to bias (literally defined as systematic deviation from the truth). Bias is a systematic error (in contrast to a random error due to chance). The effect of bias is called “like is no longer compared with like”. Bias has a direction. It either increases or decreases the estimate, but cannot do both. This is in contrast to chance findings that can have any effect on the estimate.

If the study sample is not representative of the population, the inference we make from the result may be misleading. Analytical statistics will be of no help if the sample
is not representative. Analytical statistics cannot correct our mistakes in designing the study. Every attempt should be made in the design of the study to ensure that the sample is representative. Bias cannot be addressed or corrected by statistics. The main protection is to think of the possibility of the bias and design it out. Using sophisticated computer programs does not guarantee the validity of the study. In computer jargon, they say “garbage in, garbage out”. If you feed the computer with the wrong information, you will get a wrong outcome. If the possibility of bias cannot be avoided completely in the planning of the study, the investigators must point this out when they present the findings of the study. Bias can occur when groups being compared differ systematically in a way that is related to the outcome. Main types of bias can occur at two levels: at the level of selection of subjects (selection bias) and at the level of collecting the information (information or measurement bias).

Selection bias is a systematic difference between subjects selected for a study and those who are not selected. Loss to follow-up can cause a selection bias. Attrition is the term used for reduction in the number of subjects who remain in a study. Attrition bias occurs when the subjects who drop out of a study are systematically different from those who complete the study. For example, those who develop complications or side-effects may be more likely to drop out of the study. Response bias is a specific type of selection bias in which respondents differ systematically from non-respondents to a questionnaire.

Measurement or information bias occurs when the methods of measurement or obtaining information are consistently dissimilar in different groups of patients. One type is recall bias. It is encountered, for example, when people with a certain condition are more likely to remember exposure to the variable under study than people without the condition. An example is if a study tries to compare the frequency of past oral contraceptive use among women admitted to hospital because of thrombophlebitis and a group of women admitted for other reasons. It is entirely possible that women with thrombophlebitis, if aware of the reported association between oestrogens and thrombotic events, might report use of oral contraceptives more completely than women without thrombophlebitis. A special type of information bias is surveillance or diagnostic suspicion bias encountered when patients with a risk factor may be tested for the outcome more frequently and carefully than those without the risk factor. For example, women using hormonal contraceptives may be screened more frequently and more carefully for neoplasia of the uterine cervix, than other women, leading to the diagnosis of more cases.
9.5 Confounding

In simple terms, confounders are all of the “other things” that could explain the result of the research. A careful investigator should look for all possible explanations of the results, before making a conclusion. In good scientific thinking, one should not try to assume one interpretation of the results, when other interpretations are also possible.

For example, a study may find that the risk of lung cancer is more in manual workers. A good investigator will not assume that manual work predisposes to lung cancer, before looking for other possible explanations. The result may, for example, be due to a fact that manual workers are more likely to smoke and it is smoking, not manual work, which is associated with lung cancer.

Another example is when an association is reported between herpesvirus infection and cervical cancer. Both herpesvirus and a number of other infectious agents, which may possibly cause cervical cancer, are transmitted by sexual contact. There is strong evidence that human papilloma virus infection leads to cervical cancer. It could be that the higher prevalence of herpesvirus infection in women with cervical cancer is only a consequence of greater sexual activity and so is indirectly related to a true cause, which is also transmitted sexually.

There are three ways to deal with confounding: to think of it in planning and designing the study; to measure/record the presence of the confounder during implementation of the study; and to allow for it in the analysis.

The case mix or patient mix, which refers to baseline differences among research subjects, can be a confounding factor. Matching is an important technique for creating a control group by pairing subjects, based on one or more confounding factors. An alternative is to use the control-table method, in which stratification is done afterwards. Rather than arranging subjects by groups as the study is designed, results are calculated within specified subdivisions. When more than a few confounding variables are present, the statistical technique of multivariate analysis is used.

As an example of analysis for confounding factors we may look at a study of the relationship between the working status of mothers and the duration of breastfeading. The study may show that women who are employed full-time are less likely to breastfeed for a long duration than women who are employed part-time and women who are not employed. However, the level of education of the mother may be a confounding variable, since it can affect the outcome (duration of breastfeeding) and it may correlate with the working status. Before blaming work for the shorter duration of breastfeading, there is a need to consider the confounding factor of education. Stratification may be used. A cross-tabulation table may be constructed for mothers at different educational levels, for example those who had no schooling, less than 5 years of schooling, 5–9 years and 10 years or more. For each table, we look at duration of breastfeeding in mothers who
are employed full-time, employed part-time and not employed. An alternative way of considering this confounding factor is matching at the design and implementation phase. For each employed mother with less than 5 years of schooling, we would choose a non-employed mother with a similar educational level.

Crude rates are the terms used when results have not been adjusted for confounding factors. Adjusted rates are the terms used when results have undergone statistical transformation to permit fair comparison between groups differing in some characteristic that may affect risk of disease.

9.6 Making the case for causation

The association of two variables does not necessarily mean causation and should not be interpreted as a causal relationship. Historically, scientists had to struggle with this issue, in the early days when microbiologists began to discover and report on the association of certain microorganisms with some disease conditions. In 1882, Koch stipulated that for an infectious agent to be considered the cause of a disease, the following criteria must be established:

• the organism must be present in every case of the disease;
• the organism must be isolated and grown in pure culture;
• the organism must cause a specific disease when inoculated into an animal; and
• the organism must then be recovered from the animal and identified.

A false appearance of association can occur through three mechanisms: chance, bias, or confounding. But, even after excluding, to the best of our effort, the possibilities of chance, bias and confounding variables, other criteria are needed to turn an association into causation. Sir Austin Bradford Hill proposed a set of features that should be sought when deciding whether a relationship is causal or just an association (Hill, 1965). They are still valid and are referred to as the Bradford Hill criteria:

• strength of the association
• consistency of the observed evidence
• specificity of the relationship
• temporality of the relationship
• biological gradient of the dose–response
• biological plausibility
• coherence of the evidence
• experimental confirmation
• reasoning by analogy.
• Strength of the association: A strong association between a purported cause and effect, as expressed for example by a large relative or absolute risk, is better evidence for a causal relationship than a weak association.

• Consistency of the observed evidence: When several studies conducted at different times in different settings and with different kinds of patients all come to the same conclusion, evidence for a causal relationship is strengthened.

• Specificity of the relationship: Specificity (one cause, one effect) is more often found for acute diseases. But for other diseases, there are often many causes for the same effect, and many effects may arise from the same cause. An example is the association of smoking and lung cancer. Smoking causes other diseases and lung cancer has other causes. Strong specificity is evidence for cause, but the absence of specificity is only weak evidence against a cause-and-effect relationship.

• Temporality of the relationship: Causes should obviously precede effects. This self-evident principle may, however, be overlooked when interpreting cross-sectional or case-control studies, in which both the cause and effect are measured at the same point in time.

• Biological gradient of the dose–response: A dose–response relationship is present when varying amounts of the purported cause are related to varying amounts of the effect. An example of a dose–response curve is when lung cancer rates are plotted against number of cigarettes smoked.

• Biological plausibility: When the assertion of cause and effect is consistent with our knowledge of the mechanisms of disease, as they are currently understood, plausibility is often given considerable weight when assessing causation. It is important, however, to remember that what is considered biologically plausible depends on the state of medical knowledge at the time.

• Coherence of the evidence: A factor is also more likely to be a cause of disease if its removal results in a decreased risk of disease. For example, if people give up smoking does this decrease the likelihood of lung cancer?

• Experimental confirmation: A causal association is more likely if supported by experimentation in animals.

• Reasoning by analogy: The argument of analogy for a cause-and-effect relationship is strengthened if there are examples of well established causes and effects that are analogous to the ones in question.
9.7 Interpreting end points to measure the outcome

The use of one end point may ignore the possible effect on other variables that may have a clinical impact. For example, one study reported a 44% reduction in heart attack for physicians who took low-dose aspirin. But there was a trend toward increased haemorrhagic stroke among treated subjects and an undiminished overall death rate from cardiovascular causes. The aspirin appears to be performing the role of platelet inhibition quite well, but selecting only one end point would have masked its possible other effects (Steering Committee of the Physicians’ Health Study Research Group, 1989).

It is also important to consider when an end point or outcome occurs in relation to the intervention. Outcomes that occur long after subjects stop taking a drug, or before benefits of an intervention can logically be expected, can complicate the interpretation.

Surrogate end points are sometimes used to determine an outcome, and caution should be exercised in extrapolating from the result. A surrogate end point can be defined as a variable that is relatively easily measured and that predicts a rare or distant outcome, but which is not itself a direct measure of either harm or clinical benefit. The two main advantages of the use of end points are that they can considerably reduce the sample size, duration, and therefore cost of studies. They can also allow treatments to be assessed in situations when the use of primary outcomes would be excessively invasive or unethical. For a surrogate end point to be a good measure of the outcome it must have a good positive predictor value, and a good negative predictor value, i.e. it should be both sensitive and specific. It is not enough for the end point to be biologically plausible to draw clinical conclusions. The end point should also be amenable to quality control monitoring. Examples of the use of surrogate end points include the use of lipid profile as a surrogate for the development of cardiovascular disease, and the use of the CD4 cell count to predict survival in HIV infection.

9.8 Interpreting studies of risk factors

Studies of risk factors are very important in the prevention of disease. They often attract public and media attention. Unless interpreted properly, they can lead to misinformation. Studies of risk factors cannot be interpreted without proper understanding of the following concepts: basic risk, relative risk, confidence intervals, attributable risk, and balancing risks and benefits.

Basic risk

Basic risk statements express the likelihood that a particular event will occur within a particular population. For example, the US Women’s Health Initiative (WHI) Study reported in July 2002, that postmenopausal women using hormone replacement therapy
(HRT) had an incidence of colorectal cancer of 10 per 100,000 women per year. This basic risk, however, does not mean much unless we know how many women would have developed the same disease condition without having used HRT. Without this information, we cannot interpret the finding. It may indicate a risk, but it may also indicate a protective effect. In the placebo-controlled group followed up in the same study, the incidence was 16. This means that HRT was actually protective against colorectal cancer (Writing group for the women’s health initiative investigators, 2002).

**Relative risk**

Relative risk is the ratio of the incidence of outcome in the exposed group to the incidence of the outcome in the unexposed group.

The odds ratio, a term used in case-control studies, measures the odds of having the risk factor among people with the disease divided by the odds of having the risk factor among people without the disease.

**Confidence interval**

The statistical concept of confidence interval has been discussed in Chapter 8. A study that reports the relative risk or the odds ratio without reporting the confidence interval cannot be adequately interpreted. Confidence intervals should always be presented for the relative risk and odds ratio. The important feature to look for in assessing the confidence interval is whether the boundaries include unity. A relative risk or odds ratio of 1 means there is no association between the risk factor and the disease. A relative risk may be much higher than 1.0, but if the 95% confidence interval overlaps 1.0, it can be concluded that the increase in risk is not statistically significant, and could have been a chance finding.

**Attributable risk**

The importance of a risk factor cannot be interpreted on the basis of the magnitude of the relative risk, without relating it to the prevalence of the particular disease condition. The term attributable risk is an estimate to quantify the contribution, which the particular risk factor makes in producing the disease within a population. The following two examples illustrate the importance of calculating the attributable risk.

The relative risk of lung cancer due to smoking is much greater than is the relative risk of myocardial infarction among smokers. However, heart disease is much more common than lung cancer. So, even though the risk associated with heart disease and smoking is small, its importance to the general health is magnified by its relatively higher incidence.
Women taking oral contraceptive pills are more likely to have a fatal heart attack than women not taking the pill. However, women of reproductive age (pill users) have a low incidence of myocardial disease. This increased relative risk translates into an attributable risk of only very few deaths per 100 000 users per year.

**Balancing risks and benefits**

Decisions cannot be made on risks alone, if there are benefits as well. This applies, for example, to the case of oral contraceptives, which have health risks, but also health benefits which outweigh the potential risks.

### 9.9 Interpreting studies of diagnostic tests

The term diagnostic test is used broadly to describe the value of a symptom (or collection of symptoms), signs or special investigation in the diagnosis of a clinical condition or health situation. In order to assess the diagnostic worth of tests, they must be compared to the gold standard, i.e. the best test currently available. Otherwise, there is no way to make sure that the test is diagnosing the condition in question. Proclamations of highly statistically significant associations between the test and the diagnosis are not sufficient. It is essential to provide information on the extent to which diagnostic tests misclassify subjects, i.e. make a diagnosis of a disease when it is not present, or miss the diagnosis when a disease is present. For this, the concepts of sensitivity, specificity, predictive value and efficiency are used.

**Sensitivity**

Sensitivity is the ability of a test to single out people who have the disease. Low sensitivity will mean that there will be many false negatives.

**Specificity**

Specificity is the ability of a test to label people who do not have the disease as negative. Low specificity means there will be many false positives.

**Predictive value**

The predictive value of a test gives the frequency with which a positive test actually signifies disease. It is more appropriately labelled positive predictive value. (The negative predictive value is less often used).
Efficiency

Efficiency is an overall estimate of a test’s ability to classify patients correctly. It is estimated by adding the numbers of the two correct classifications (true positive and true negative) and dividing by the total number of patients assessed.

Balancing sensitivity and specificity

Diagnostic tests cannot be expected to be perfect. Increasing the sensitivity of a test often results in decreased specificity and vice versa. The value of a test cannot be made on the basis of sensitivity and specificity alone, or on the basis of overall efficiency. Much depends on the prevalence of the disease condition. The predictive value is at the mercy of prevalence. Even a small percentage of false positives can become magnified when a disease is rare.

Choices between sensitivity and specificity must be made. The decision is not statistical; it is clinical and economical. When the consequences of missing a disease are crucial, sensitivity is paramount. But if the burden of creating false positives outweighs the advantages of capturing all cases of a disease, increasing specificity should be the goal.

With the economic aspects of health care drawing increasing attention, costs are becoming a concern in the evaluation of screening procedures. The following example of mandatory premarital serological testing of HIV infection illustrates the point. A test used may have a sensitivity of 98%, and a specificity of 99%. This sounds like an impressive performance. But in a community where HIV prevalence is low, even the low rate of false positives will mean that a large number of people will be unnecessarily alarmed. It also means that a very large number of people will need to be tested in order to detect one single true positive case. An estimate will need to be made of the cost of diagnosing one single case.

Many diagnostic tests yield continuous results, for example serum levels of prostate specific antigen (PSA) as a screening test for prostate cancer. With such tests, a decision must be made as to what value will constitute a positive test, a value called the “cut-off point”. This decision requires trading an increase in specificity for a decrease in sensitivity, or vice versa. Receiver operator characteristic (ROC) curves are useful for visualizing and selecting the most appropriate cut-off point for screening tests. The terminology comes from its first use in the field of electronics. ROC curves are a graphic way of portraying the trade-offs involved between improving either a test’s sensitivity or its specificity when different cutoff values are selected. For each cut-off point, statisticians plot the sensitivity on the vertical axis, and the value that is 1 minus specificity (false positives) on the horizontal axis (Newman et al. 2001).
9.10 Interpreting studies that report the results of interventions

Results based on uncontrolled studies do not mean much. In many cases, it is possible that no treatment could have achieved comparable results.

Results of controlled studies based on comparison of a treatment with a placebo do not mean much, if there are other available treatments. Results should be based on comparator drugs currently available.

The preference of one drug over another should not be based on one aspect only of its performance. It should give equal consideration to the four elements of the acronym “STEP”: safety, tolerability, efficacy, price.

A study may indicate that a new drug has resulted in more improvement than the currently available and used drug. A statistical test shows that this difference between the two drugs is statistically significant. What the statistical test says is that the difference in the result is unlikely to have happened by chance, and that the probability of its being a chance finding is less than 5%. The $P$ value can tell us how remote the possibility that the difference can be found by chance. This, however, does not mean that the result is clinically significant. The $P$ value does not tell us anything about the magnitude of the difference between the two treatments, and how the point of estimate of the difference will be changed if other samples from the same population are studied. For this we need to know the confidence intervals for the difference between the performance of the two drugs. Confidence intervals will show the likely range of the magnitude of the difference in the performance of the two drugs.

Even if a drug is proven to be superior to another, the question still remains about what this means to the individual patient or individual clinician. One has to calculate the number needed to treat in order to achieve the therapeutic advantage of the new intervention. If for example, one intervention reduces the absolute risk of dying by say 4%, it means that the number needed to treat with the new intervention to avoid one death will be 25. The “number needed to treat” has important cost implications when the result of the intervention is interpreted.

9.11 Interpreting results of qualitative research

Qualitative research methods involve the interpretation of textual material derived from talk or observation.

In interpreting qualitative findings, the investigators should carefully look into their credibility, dependability, confirmability and transferability.
Credibility means interpreting the qualitative data in a way that offers explanations that are consistent with the data collected. Negative findings should be adequately presented and addressed, and alternative explanations considered. As in quantitative research, the investigators should look for confounding variables. For example, a study may reveal that homes sprayed for malaria control had a higher incidence of malaria when sprayed in the afternoon. It could be that sprayers used most of the spray in the morning so that the load to carry in the afternoon would be lighter. To ensure credibility of the interpretation, the investigator should act as the “devil’s advocate”, considering all potentially competing explanations of the results.

Possible sources for bias should be checked, for example observer bias or the influence of the researcher on the research situation. A researcher’s background and position will affect the process of qualitative research. The investigator always enters a field of research with certain opinions about what it is all about. In qualitative research, this potential bias cannot be eliminated, but it should be exposed in a process termed “reflexivity”. Reflexivity starts by identifying preconceptions brought into the project by the researcher, representing previous personal and professional experiences, pre-study beliefs and qualifications for exploration of the field. During all steps of the research process, the effect of the researcher should be assessed, and, later on, shared. Adequate accounts of these effects should be considered in the limitations and strengths of the study, and transferability of findings.

Dependability means that data can be replicated. The replication is not necessarily of the results, but of the process used to obtain the results. Other investigators should be able to replicate the study.

Confirmability means that other researchers can have access to the data and can do their own analysis. The concept of “audit trail” enables others, on the basis of the collected data, to review the analysis decisions and verify the interpretations.

Transferability means the use of the findings to make inferences to other populations. This may not be possible because qualitative research is often context-specific. Qualitative research emphasizes depth more than breadth, and insight rather than generalization. In such cases, however, there are lessons learnt that may help in understanding the situation in other populations.

References and additional sources of information

Byrne DW. *Publishing your medical research paper*. Baltimore, Lippincott Williams & Wilkins, 1998.


