EPIDEMIOLOGICAL STATISTICS–II

Don McNeil

Macquarie University, Sydney, New South Wales, Australia,
dmcneil@efs.mq.edu.au

Abstract. Epidemiological studies use statistical methods to understand the factors that cause, reduce, and prevent diseases in human populations. These studies estimate the parameters in a target population based on representative samples, and can be qualitative or quantitative, descriptive or comparative, and observational or experimental. Parameters of interest include levels of disease and associations between risk factors and disease outcomes. Concept maps are useful for showing causal paths between risk factors, disease outcomes, and intervening variables (confounders) that could distort the association. The statistical methods can be classified by the type and role of the data, and are subject to both inaccuracy or lack of statistical power (due to sampling variability) and bias (due to faulty measurement, differential selection of subjects, or confounding). Matching of subjects can improve accuracy. Mantel–Haenszel methods can reduce confounding bias. Meta-analysis is used to combine results from different studies. Statistical models include linear, logistic, and Poisson regression, and the proportional hazards model for handling censored data.

Keywords and Phrases. biostatistics, clinical trials, cohort analysis, demography, epidemics, Framingham, longitudinal data analysis, logistic regression, Mantel–Haenszel statistic, matched pairs, medical diagnosis, medicine, Poisson regression, proportional hazards model, Cox's;

Prospective Studies, retrospective studies, relative risk, selection bias, survival analysis.

Blind Entry.

AMS Subject Classification. **AMS Subject Classification. Primary: 60T42, 78J12; Secondary: 44R44, 43B43, 01A01**

Introduction

Epidemiology is the branch of medicine concerned with understanding the factors that cause, reduce, and prevent diseases by studying associations between disease outcomes and their suspected determinants in human populations. It involves taking measurements from groups of subjects and making inferences about relevant characteristics of a wider population typifying the subjects. Since statistics is the science that is primarily concerned with making inferences about population parameters using sampled measurements, statistical methods [38] provide the tools for epidemiological research. (*See* MEDICINE, STATISTICS IN).

An early example of epidemiological statistics in practice was a study by Louis [21], who investigated the effect of the entrenched medical practice of bloodletting on pneumonia patients in the 1830s, and found evidence that delaying this treatment reduced mortality. Another epidemiological pioneer was Snow [36] who found that the cholera death rate in London in 1854 was five times greater among residents drinking water from a particular supplier, thus identifying contaminated water as a risk factor for this disease.

While Louis [20] had already described many of the basic principles underlying experimental research in epidemiology, it was not until the middle of the twentieth century that a major clinical trial* was undertaken, when the British Medical Research Council sponsored Hill's investigation [12] of the effect of streptomycin treatment for tuberculosis. In contrast, in the last half-century, as new and more virulent diseases like AIDS have decimated populations, epidemiological research methods have become widely used [3, 4, 8, 11, 14, 15, 19, 25, 32, 35, 37–39].

Data roles

Each study attempts to answer a research question of interest, involving a target population, using a specified set of research methods. In epidemiological studies, the individual subject is typically the observational unit from which data are collected. Each subject supplies a set of measurements comprising one or more variables. The role of a variable is its definition as an outcome or a possible determinant of that outcome. An *outcome* is a measure of the subject's health status at a particular period of time, whereas a *determinant* is a possible cause of the outcome through its action at an earlier period of time.

Determinants include genetic factors affecting predisposition to disease, such as haemophilia, which may increase the risk of infection through a contaminated blood transfusion, and demographic factors, notably age, gender, occupation, and marital status. They also include environmental and occupational exposures such as contaminated water, asbestos dust, and excess fat and cholesterol in the diet. Behavioral determinants include tobacco and excess alcohol consumption, overexposure to the sun, unsafe sexual practice, and drug addiction. Determinants include preventative measures for health risk reduction, such as a vaccine for combating an infectious disease, screening for breast or prostate cancer, an exercise promotion campaign, or a treatment aimed at curing or alleviating a disease or preventing further deterioration in a person's health. In this general definition, determinants include medications such as aspirin, hypertensive and hormonal drugs, and treatments such as radiotherapy for cancer and surgery for heart disease.

Concept maps

It is useful to have a concept map or causal diagram [9; 25, p. 41] to illustrate graphically the roles of variables in a study. Figure 1 shows two examples.

The distinction between a determinant and an outcome is not always clear-cut. Inadequate prenatal care (measured by the number of visits to a clinic) during the first trimester of pregnancy (T1) is usually associated with an increase in the risk of perinatal mortality. However, reduced prenatal care during the third trimester of pregnancy (T3) is usually associated with reduced perinatal mortality. The reason

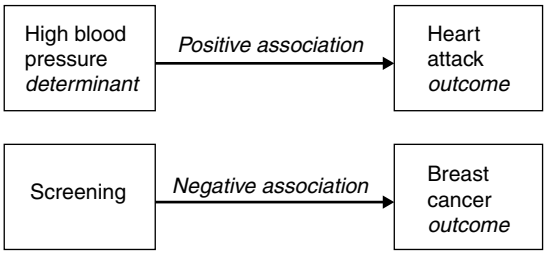

Figure 1 Concept map examples

for this apparent anomaly is that any pregnancy complication is likely to give rise to additional prenatal care, so that prenatal care in the first trimester is a determinant, but prenatal care during the third trimester is an outcome. The term *intervening variable* is used to describe a variable on the causal path between a determinant and an outcome. Figure 2 shows a concept map for this example. In this graph, the bold arrows show the true causal relations, while the other arrows show the associations that would be observed in the absence of consideration of the intervening variable.

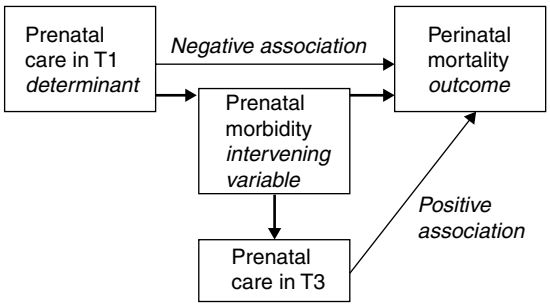

Figure 2 Concept map with an intervening variable

Typically, a study will have just one outcome variable of interest. If there is more than one outcome variable, these outcomes may be considered separately, and usually the analysis will focus on the one of primary interest. However, often there will be several determinants of interest, even though the research question will focus on the association between a particular determinant and the outcome under consideration.

Consider a study in which the research question is the extent to which calcium deficiency is a risk factor for hip fractures in an elderly population. Osteoporosis is

known to increase the likelihood of a hip fracture in this population, and calcium deficiency is also known to cause osteoporosis, as shown in Fig 3.

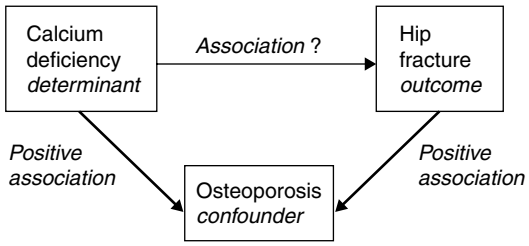

Figure 3 Concept map with a determinant and a confounder

A confounding variable (or confounder) is a determinant that affects the association between a determinant and an outcome. In this case, the association between the determinant (calcium deficiency) and the outcome (hip fracture) is of primary interest, but the intervening variable (osteoporosis) could interfere with the measurement of this association in a study. Dealing with confounding* is an important part of epidemiological statistics.

Data types

Variables can also be classified according to their data type. Data types are typically nominal* (having two or more distinct categories with no natural order), ordinal* (having three or more ordered categories), or interval (measured on a scaled range, also called continuous). Epidemiological outcomes are commonly of binary type, corresponding to the presence or absence of a disease. Binary outcomes are convenient from the medical diagnostic point of view: if a subject is diagnosed as having the disease in question, a specific treatment might be justified, but not otherwise.

Nominal variables with three or more categories often arise in epidemiological studies. Examples of such determinants include a subject's occupation, their country of birth, and marital status. Nominal outcomes include disease diagnosis (type of cancer, such as breast, ovarian, lung, colon, or stomach) and type of lung cancer cell involved (such as small, large, squamous or adeno).

Ordinal variables are useful when classification into just two categories is relatively uninformative. For infants with diarrhea, three categories of disease status (mild, moderate, and severe) are often used, giving four categories in all when disease absence is included. Similarly, a patient's outcome in a cancer trial could be classified as complete response, partial response, no change, or progressive disease. Ordinal determinants commonly measured in epidemiological studies include age group, duration of exposure to an occupational risk factor, and the status of some behavioral variable of interest such as exercise, snoring, smoking, or drug-taking.

Interval outcomes include CD4 count (typically ranging from 600 to 1200 in normal persons, a useful measure of health status for subjects with HIV), and the blood lead concentration for a child exposed to contaminated water or motor vehicle exhaust in a city. The duration of survival and a quality of life index are interval outcomes of major interest in cancer studies. Body mass index is a useful interval outcome in obesity studies. Many determinants are measured in this way: the duration of exposure to an environmental hazard, the dose level of a therapeutic drug, a disabled person's performance status, and the percentage of fat in a diet.

Disease measurement

A basic objective in epidemiological research is measuring the level of some disease in a population. If the outcome is simply disease presence or absence, then the *prevalence* is of interest: this is simply the proportion of individuals affected with the disease. For interval outcomes, a statistical summary such as the mean or the median could be used to represent the status of the population. Typical examples include *(i)* the mean birth weight of newborn babies at hospitals in an urban population, and *(ii)* the median survival time for patients in an organ transplant program.

A problem with measuring prevalence in a population is that it is likely to reflect what has happened in the past rather than the current situation. In HIV research, the extent to which new cases are occurring is of primary interest, for this measures the effectiveness of preventative programs. Thus the *incidence* of a disease, defined as the proportion of new cases that occur per unit time among persons free of the disease in the population, is another important epidemiological statistic.

Measuring associations

Associations between possible determinants and disease outcomes are usually of greater interest than disease levels. For example, knowing whether taking oral contraceptives affects a woman's risk of developing breast cancer is important, and many studies have addressed this question. Where an association has been established, it is often of interest to quantify it more accurately. For example, what is the relative risk* that a heavy smoker will develop lung cancer (compared to a nonsmoker)? If both the outcome of interest and the determinant are measured on an interval scale, their strength of association can be expressed as a correlation* coefficient. If the outcome is measured on an interval scale and the determinant is categorical, the association can be expressed as a function of the differences in the means of the outcomes associated with the levels of the determinant.

When both variables are categorical, it is more convenient to express the association as a matrix of conditional probabilities, where a typical element in this matrix is the probability of a specified outcome, given a specified level for the determinant. Equivalently, the conditional probabilities can be expressed as probability ratios (or relative risks). In the simplest situation, in which both the determinant and outcome are binary, this association is expressed as an odds ratio* or as a risk difference, as follows.

Denoting the probabilities (or risks) of the ("adverse") outcome for two individuals in two different levels of the determinant by p_1 and p_2, respectively, the *relative risk* is just the ratio p_1/p_2. The *odds* associated with a probability p is defined as $o = p/(1-p)$, so the *odds ratio* ω, say, comparing the two individuals is

$$\omega = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \tag{1}$$

(*See* ODDS-RATIO ESTIMATORS; RELATIVE RISK.) The *risk difference*, defined as $p_1 - p_2$, is also widely used, particularly in the public health literature, to measure the risk attributable to the exposure. Similarly, the *attributable risk* proportion is defined as $1 - p_2/p_1$ [32, pp. 203–211].

Where an association is suspected but has not yet been established, the research question is framed in terms of a null hypothesis, which states that there is no association between a specified determinant and a particular disease outcome in the

population of interest. On the question of a possible association between oral contraceptive use and breast cancer outcome, the null hypothesis would state that there is no association between these two variables in the target population.

The methods for testing a null hypothesis are essentially statistical (*See* HYPOTHESIS TESTING; P-VALUES). It is important to realize that a study might not provide a conclusive answer to the question, due to the limited size of the sample. Consequently the result of a study is not certain, but probable.

Types of studies

Most epidemiological studies require data collection*, but this is not essential. A study could be purely deductive, based on logic, starting with known facts or assumptions and arriving at a conclusion. Knowledge of appropriate theory can also reduce the need for data collection.

If a study involves data collection, it is said to be *inductive*, because the data in the sample are used to induce the characteristics of the target population. Inductive studies based on data collection are also called *empirical studies*, and are classified as either quantitative or qualitative.

A quantitative study involves structured data collection. In this case, the same characteristics are measured on all subjects in the study, using a protocol or set of guidelines that are specified in advance. The method of data collection in a quantitative study is often a questionnaire, or an instrument that automatically records the measurements from the sampled subjects. In contrast, a qualitative study is relatively unstructured. It may involve open-ended questioning of subjects, and may be opportunistic in the sense that the answer given to one question determines the next question. A qualitative study may precede a more formal quantitative study, with the aim of getting sufficient information to know what measurements to record in a quantitative study. For example, an investigator wishing to compare a new program for improving the reproductive health of mothers in an aboriginal community in outback Australia would spend some time in the community getting to know the people before embarking on the fully fledged study.

A quantitative study could be purely descriptive, or comparative. Descriptive studies simply aim to measure levels or prevalences, whereas comparative studies measure or test the existence of associations between determinants and outcomes.

Studies are also classified as experimental or observational. An experimental study is one in which the investigator has some control over a determinant. To investigate a possible association between beer consumption and stomach tumor incidence, for example, the investigator could take a sample of laboratory mice and divide them into two groups, force one group to drink beer and deprive the other, and observe each group to compare the incidences of tumors over a period of time.

Observational studies* do not involve any control by the investigator over the determinant, and are thus in a sense less rigorous than experimental studies. However, the latter studies require interventions that might be costly, and are not always feasible or ethical, particularly if human subjects are involved.

Clinical trials

Experimental studies in epidemiology investigate treatments, such as therapies for cancer or heart disease patients, or interventions, such as screening and health promotion studies. These studies are classified by various factors including the type of subjects and the size and extent of the study. The study is called a *clinical trial* if the subjects are hospital or doctors' patients, whereas the study is a *field trial* if it involves subjects in the community at large. Clinical trials usually involve patients who have some disease or condition, with the objective of investigating and comparing treatments for this condition. In field trials, the subjects usually are disease-free at the time of selection, and the trials aim to compare strategies for prevention.

Clinical trials are classified according to phases of development of new treatments. A Phase I trial evaluates the safety of a proposed new treatment, whereas a Phase II trial attempts to discover whether a treatment has any benefit for a specific outcome. A Phase III trial is used to compare a promising new treatment with a control treatment, which could be no treatment at all. Since patients often react positively to the idea of a treatment (even if it is otherwise ineffective), a placebo, that is, a treatment that looks like a real treatment but contains no active ingredient, is often used instead of no treatment. Phase IV trials are similar to field trials in that they

involve monitoring of treatments in the community; conceptually they differ from field trials only in the sense that they are concerned with subjects with some health problem whereas field trials usually focus on prevention.

Experimental studies often involve randomized allocation of subjects to treatment and control groups, an idea proposed by R.A. Fisher* in 1923 for comparing treatments at the Rothamsted agricultural research station in Britain. The aim of randomization is to form treatment and control groups that are initially as similar as possible, so that any substantial difference in outcomes observed in these groups cannot be ascribed to factors other than the treatment effects. Randomization ensures that the comparison groups are balanced, not just with respect to known determinants of the outcome, but with respect to all possible risk factors.

An early large field trial was the 1954 study of the Salk vaccine for preventing poliomyelitis [24]. In this study, 400,000 schoolchildren were randomly allocated to receive the vaccine or a placebo, with the result that only 57 cases of polio occurred in the vaccinated group compared with 142 in the control group.

Cohort studies

A *cohort study* is similar to a clinical trial, in the sense that the subjects are again selected (though self-selected) according to their determinant status, but it is observational rather than experimental. Cohort studies often involve monitoring subjects over an extended period of time, and consequently they are useful for investigating multiple determinants of outcomes. In a classical example (*See* FRAMINGHAM: AN EVOLVING LOGITUDINAL STUDY), from 1948 onwards residents of Framingham in Massachusetts were continuously monitored with respect to many risk factors and disease outcomes.

If the data collection is *prospective**, as is often the case in cohort studies, a cohort study can be an expensive and time-consuming exercise. Breslow and Day [4] provided a detailed account of the use of cohort studies in cancer research.

Case-control studies

A *case-control study* (*see* RETROSPECTIVE STUDIES) is similar to a cohort study, but the subjects are selected according to their outcome status rather than the determinant [3, 30, 34].

Both cohort studies and case-control studies involve differential selection of subjects according to their exposure or disease status. In a cohort study, a group of disease-free subjects exposed to the determinant of interest is first selected, together with a comparable group of subjects not exposed to the determinant, and the subsequent outcome status of the two groups is then compared. In a case-control study, the variables are reversed: first, a group of subjects with the outcome and a comparable outcome-free group are selected, and then the levels of prior exposure in the two groups are compared.

Since the exposure must logically precede the outcome, a cohort study cannot look for outcomes that occur before the exposure, and a case-control study cannot look for exposures that occur after the outcome. Thus, cohort studies are often said to be *prospective* (*see* PROSPECTIVE STUDIES) and case-control studies *retrospective*.

When the disease is rare, a cohort study is inefficient, because a large number of subjects will be needed to obtain sufficiently many adverse outcomes to obtain a conclusive result. In this situation, a case-control study is more efficient because one of the two groups being compared contains only the subjects with the disease, and the control group can be restricted to a comparable number of subjects. By the same token, a case-control design is inefficient when the exposure to a risk factor is rare and the adverse outcome is relatively common.

Cross-sectional studies

In a *cross-sectional study*, there is no differential selection of subjects, either by the determinant or the outcome: one simply selects subjects from the target population, without taking into account the outcome or the determinant. Snow's investigation of risk factors for cholera [36] was a cross-sectional study: here the sample comprised all the residents of a particular area of London where a cholera epidemic had occurred

during July and August 1854, and the subjects were classified according to death from cholera (the outcome) and their source of drinking water.

Figure 4 gives a graphical representation summarizing the various types of epidemiological studies.

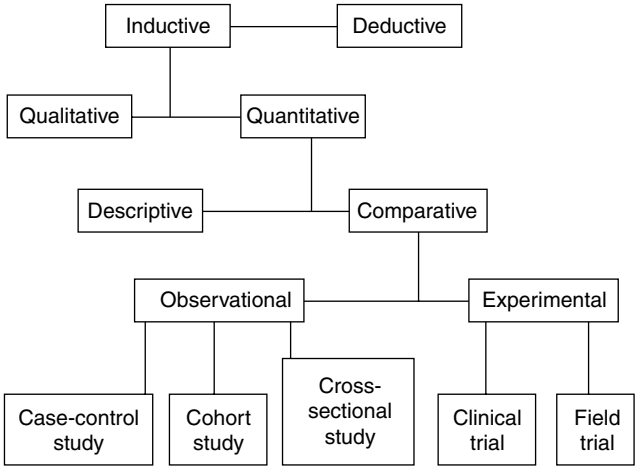

Figure 4 Classification of research studies

To summarize, inductive studies involve data collection, in contrast to purely deductive studies. Inductive studies can be qualitative or quantitative, depending on the extent to which similar data are collected from each subjects. Quantitative studies can be descriptive or comparative, and comparative studies can be observational or experimental. Observational studies can be of cohort, case-control, or cross-sectional type, and these are distinguished only by the method of selection of subjects. Since experimental studies involve allocation of subjects to treatment or exposure groups (ideally by randomization), which must occur prior to the occurrence of the outcome, they could be regarded as controlled cohort studies.

Bias

Two error factors reduce the credibility of a study: *(i)* systematic error, or bias, and *(ii)* chance error, or sampling variability.

Bias is a systematic distortion in a measured effect due to a deficiency in the study design. It may arise in three ways: *(i)* from poor measurement (information

bias), *(ii)* because the sample is unrepresentative of the target population (selection bias*), or *(iii)* from the differential effects of other determinants on the association (confounding).

There are many sources of measurement bias, such as faulty measuring instruments, different standards in different biochemical laboratories, errors made by clinicians in diagnosing diseases, bias by investigators consciously or unconsciously reporting more favorable results for treatments they believe in, biased reporting of symptoms by patients wishing to please their doctors, memory lapses by subjects in case-control studies when asked to recall past exposure to a risk factor (recall bias), lack of compliance by patients in clinical trials, and poor data quality management.

Some measurement biases can be reduced or eliminated by good study design. For example, blinding of investigators and subjects, so that they don't know which treatment a subject received until after the response has been evaluated, can reduce biased reporting in clinical trials. When the evaluators know the treatment allocation but the subjects do not, the study is said to be *single blind*. If neither patients nor evaluators know the treatment allocation, the trial is said to be *double blind*.

Selection bias has two levels. The first occurs when the sample does not represent the target population but still constitutes a representative sample from some restricted population that is a subset of the target population. In this case, the results obtained from the study might not be generalizable to the target population but are valid for a subpopulation. Such studies are said to have internal validity but lack external validity. Because their subjects usually constitute a select group that must satisfy tight eligibility criteria, randomized clinical trials only have internal validity.

The second, more serious, level is *differential selection bias*, which arises when the selection criteria for inclusion in a study vary with respect to a factor related to the outcome. Whenever a cross-sectional study has a low response rate, there is an opportunity for differential selection bias (response bias) because the nonresponders could provide different outcomes to the responders. Such bias often arises in studies in which questionnaires are mailed to sampled subjects.

Sackett [33] classified biases that can arise in case-control studies, which are particularly prone to differential selection bias, because it is difficult to get a control group that is representative of the noncases in the target population.

The third kind of bias in a study is called confounding bias, and it arises where the association between a determinant and an outcome is distorted by another determinant. Confounding arises whenever an outcome has two or more determinants that are themselves associated and one is omitted from consideration.

Confounding can make two unrelated variables appear to be related, and can also appear to remove or reverse a valid association.

Sampling Variability

Sampling variability arises because samples are finite, so even if there is no bias in a study, its conclusion is only probably true. Two statistical measures are associated with the sampling variability—a confidence interval* and a P-value*.

A 95% confidence interval is an interval surrounding an estimate of a population characteristic (such as a mean or a prevalence, or a relative risk or an odds ratio), which contains the population characteristic with probability 0.95. It is thus a measure of the precision with which a population parameter can be determined from a study: the narrower the confidence interval, the greater the precision. The chance that the population parameter will not be located within the given confidence interval is 0.05 (*see* CONFIDENCE INTERVALS AND REGIONS).

While 0.05 is the conventional statistical false positive error rate, wider confidence intervals, containing, say, 99% probability, are preferable when several parameters need to be estimated from the same study. Increasing this probability level can ensure that the overall risk of making an incorrect conclusion remains close to 0.05.

A 95% confidence interval for a population parameter may be given approximately and simply as the interval $(T - 1.96 \times SE, T + 1.96 \times SE)$, where T is the estimate obtained from the study and SE is its standard error, defined as the (estimated) standard deviation of T. This is based on the assumption that the sampling distribution of the estimate T is approximately normal, which can be justified by statistical theory to be so if the sample size is large enough (*see* ASYMPTOTIC NORMALITY and LARGE-SAMPLE THEORY). In the case of the distribution of an odds ratio, the approximation to normality is substantially improved by replacing the odds ratio by its logarithm.

A substantial part of epidemiological statistics is thus concerned with obtaining reliable formulas for the standard errors of the estimates of population parameters that arise in particular situations. If there is no bias, the standard error of an estimate generally decreases in proportion to the inverse square root of the sample size. Symbolically, this result may be expressed as

$$SE = \frac{c}{\sqrt{n}} \tag{2}$$

where c is a constant and n is the sample size. In particular, this means that you need to quadruple the sample size in order to reduce the width of the confidence interval by a factor of 2, and thus double the precision of the estimate. This result assumes that the observations are independent. In epidemiological studies, data are often clustered, either due to geographical proximity or because repeated measurements are taken on identical or similar subjects. Methods for detecting and adjusting for correlated outcomes have been developed [11, 29, 39].

A P-value is more complex than a confidence interval, in the sense that it involves a null hypothesis*, that is, a statement or claim that a target population parameter equals a specified null value.

When there is just one variable of interest, the parameter could be a mean or a prevalence. For a comparative study in which a possible association between two variables of interest may be present, the null hypothesis usually states that there is no association between the variables in the target population.

A P-value has the same conceptual basis as a confidence interval, namely, the idea of repeating the study many times under the same conditions, each time using an independent sample based on the same number of subjects. It is the probability, assuming that the null hypothesis is true, that another such study will give an estimate at least as distant from the null value as the one actually observed. The P-value is usually calculated as the area in the tails of a normal distribution or as the area in the right tail of a chi-squared distribution.

Although a small P-value provides evidence against a null hypothesis, a relatively large P-value needs to be supported by a large sample before it provides evidence in favor of a null hypothesis.

Matching

Both selection bias and confounding, as well as sampling variation, can be reduced by matching, a design technique that involves subdividing the treatment or exposure groups into smaller subgroups, or strata, so that the members of a stratum are homogeneous with respect to specified determinants, such as age or occupational status, which are not themselves of interest in the study. An important special case is the matched pairs* design, in which each stratum comprises just two subjects. In some situations it is feasible for each stratum to consist of just one individual, and the corresponding studies are called *crossover studies* (*see* CROSSOVER TRIALS), in which the subjects act as their own controls. In a crossover study, the subjects are first divided into two or more treatment groups as in a Phase III or IV trial, and after a specified period of time, each subject is given an alternative treatment [1, 5, 35].

While matching is often used in epidemiological studies to improve precision or reduce bias, it can be unnecessary and even counterproductive. If a covariate is not an independent risk factor for an outcome of interest, matching on it is simply a waste of effort. If a covariate is not an independent risk factor for the outcome but is associated with the exposure factor, matching on it is both wasteful and inefficient, since this will create uninformative matched sets whose members have the same exposure. When a covariate is associated with both the risk factor and the outcome, matching on it can introduce bias. The term *overmatching* is used to describe such situations. Consequently, matching is most effective when the covariate on which the matching is done is a risk factor not associated with the determinant of interest.

In epidemiological studies, it is very common to match with respect to age. The advantage of this strategy is that age is often a strong risk factor, but one that is not of primary interest in its own right. Effective matching will then improve the efficiency of the study by reducing the variation in the outcome due to the matched covariate, and can also reduce bias that might arise from the confounding effect of the covariate.

Statistical Significance versus Effect Importance

In statistical methodology, the term *significance* is defined by the P-value: the smaller the P-value, the more significant the result. But a statistically significant result does not necessarily equate to a worthwhile effect. A study could easily fail to detect a worthwhile effect. Alternatively, a result could be statistically significant, but of no practical importance.

Confidence intervals can illustrate and elucidate these apparent anomalies. Figure 5 shows 95% confidence intervals for five different studies involving the estimation of a population parameter. The population parameter has null value 0, and a value of δ or more is considered worthwhile or important. For example, the parameter might be the effect of a new but expensive drug in reducing the risk that an HIV-infected person will die from AIDS within five years. In this case, it might be reasonable to choose δ to be 0.1, on the grounds that any lesser benefit would not outweigh the high cost of the new drug.

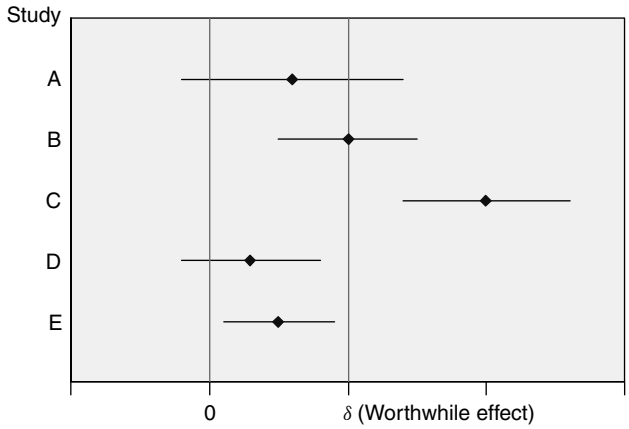

Figure 5 Examples contrasting statistical significance and effect importance

In Study A, the 95% confidence interval includes the null value, so the result is not statistically significant. However, the effect could be important, because the confidence interval also includes the value δ. In this case, the study is quite inconclusive. A larger study would need to be undertaken.

In Study B, the result is statistically significant because the 95% confidence interval does not contain the null value, but the effect might not be important. So the

study is still not completely conclusive, and a larger sample is needed to establish the importance of the effect.

In Study C, the result is statistically significant and the effect is important. So in this case a conclusive result has been obtained.
In Study D, the result is not statistically significant and the effect is not important. Despite the absence of statistical significance, the study is conclusive.

Study E illustrates the final possibility, in which the result is statistically significant but the effect is not important.
Ideally, a study should be large enough to detect a worthwhile effect, but not so large that it can detect an unimportant effect.

Basic Statistical Methods

For comparative studies, the statistical method is determined by the data types of the determinant X and the outcome Y. If both Y and X are binary, the data comprise a two-by-two contingency table of counts (*see* TWO-BY-TWO TABLES), giving two proportions to be compared, and a single odds ratio measures the strength of the association. If the variables have more than two categories, the data can be summarized using multiple proportions and odds ratios*. Logistic regression* is appropriate when the outcome is categorical and the determinant is of interval type. For interval outcome data, means rather than proportions are used to summarize the data, and differences of means rather than odds ratios measure the associations of interest. Correlation and regression* methods (*see* LINEAR REGRESSION) are most convenient for handling data in which both the outcome and the determinant are continuous.

As described below, the methods become more complex when a third variable exists.

Sample Size Determination

An important problem faced by the epidemiological statistician is that of determining the appropriate sample size in a study (*see* SAMPLE SIZE DETERMINATION). This choice depends on the precision needed in estimating the parameter of primary interest in the study. It is also desirable to compute the statistical power, which is

the probability of making a correct decision when rejecting the null hypothesis of interest.

Assuming that the sample size is sufficiently large for asymptotic normality properties to apply, the *precision* [22] with which a population parameter can be estimated is defined as the half-width of the confidence interval, that is,

$$d = z_{\alpha/2}, SE, \tag{3}$$

where $z_{\alpha/2}$ is the critical value for the standardized normal distribution corresponding to a two-tailed area α. To determine the sample size required to achieve a specified precision, a formula for the relevant standard error is thus needed.

The Power of a Study

When a major objective of a study is to test a new treatment, rejecting the null hypothesis could lead to a change in health policy. This can happen when a clinical trial comparing a promising new therapy with the standard treatment finds in favor of the new therapy. While there is always a chance of rejecting the null hypothesis when it is true, and this risk (α, the Type I error) is conventionally taken to be 5%, the probability of failing to detect a worthwhile benefit (β, the Type II error) must also be taken into account.

•Q1 The power* is $(1 - \beta)$ of a study •is as the probability that a worthwhile benefit will be detected. This probability depends on various factors, including the size of the worthwhile benefit, the variability in the data, and the sample size. The power of a study can be expressed in terms of these parameters:

Suppose that a worthwhile benefit δ exists and that the estimate d of this benefit based on the data is approximately normally distributed with mean δ and standard deviation SE, so that $Z = (d - \delta)/SE$ has a standardized normal distribution. The null hypothesis that δ is 0 is rejected whenever d exceeds $z_{\alpha/2}SE$ in magnitude. When $\delta = 0$, the null hypothesis is true, and this probability is α, the Type I error. But if δ is substantially greater than 0 (greater than SE, say), the null hypothesis is rejected when $|Z + \delta/SE|$ exceeds $z_{\alpha/2}$, which happens when $Z > z_{\alpha/2} - \delta/SE$ or $Z < -z_{\alpha/2} - \delta/SE$. Since $\delta > SE$, the probability of the second alternative is negligible for reasonable values of α, so the approximate probability of rejecting

the null hypothesis is the tail area corresponding to the critical value $z_{\alpha/2} - \delta/SE$, that is, $1 - \Phi\left(z_{\alpha/2} - \delta/SE\right)$, where $\Phi(z)$ is the cumulative standardized normal distribution function. This probability is also the power of the study, $1 - \beta$, which equals $\Phi(z_\beta) = 1 - \Phi(-z_\beta)$, so that

$$\delta = \left(z_{\alpha/2} + z_\beta\right) SE. \tag{4}$$

This formula assumes that the variability of the estimated benefit d does not depend on δ, and becomes more complicated when this assumption is relaxed. It is quite similar to the precision formula (3), and thus could be used to determine the sample size needed to achieve a specified power. Equation 4 can be inverted to determine the power of a study, given the sample size:

$$1 - \beta = \Phi\left(\delta/SE - z_{\alpha/2}\right). \tag{5}$$

Adjusting for a Stratification Variable

A large body of statistical methods used in epidemiology deals with categorical data* involving three factors. The first two factors comprise a determinant X and an outcome Y, and the third is a stratification variable Z. If X, Y, and Z have levels r, c, and s respectively, the data may be represented as an $r \times c \times s$ contingency table. The associations of X and Y with Z may distort the association between X and Y. When X and Y are both binary, a method due to Cochran [7] is used to test for an association between X and Y after adjusting for Z. This method uses a chi-squared test with one degree of freedom, and assumes that the association is the same for each level of Z. Birch [2] extended the test to general $r \times c \times s$ tables.

When r and s are both equal to 2, the association can be expressed in terms of a single odds ratio. Assuming that this odds ratio is the same in each stratum of Z, Mantel & Haenszel [23] gave a robust estimator of the odds ratio (*see* MANTEL–HAENSZEL STATISTIC), and Robins et al [31] derived the variance of its logarithm.

The common odds-ratio assumption can be tested using a chi-squared test with $s-1$ degrees of freedom [3, p. 142]. This test requires that the counts in the individual strata be reasonably large.

Clayton [6] showed that where both X and Y are ordinal, the association could be summarized in terms of a single odds ratio, defined in terms of the aggregated counts

for cells with values up to and including each specified value; it is assumed that all these odds ratios are identical in the target population. A chi-squared test comparing the observed counts with expected counts could test this homogeneity assumption.

Meta-Analysis

If a study is too small, it is unlikely to give a conclusive result. When planning clinical trials, it is conventional to require a power substantially greater than 0.5. But if the research question is important, any properly conducted study gives useful information. The evidence accrued from many small studies can be combined using a method called *meta-analysis*, which first arose in the social science literature [13] and is now widely used in medical research [26, 27, 37].

Ideally, each contributing study investigates the same treatment or risk factor and the same outcome for similar subjects, and each study will have the same type and quality of design. But selecting the studies to include is difficult because studies, even when they have the same research objective, tend to vary substantially in quality from place to place and at different times. Biases can arise if not all relevant studies are included. In practice, inconclusive or uninteresting studies often remain unpublished, giving rise to publication bias [10]. Undertaking a meta-analysis requires a professional team of scientists.

In general, it is always possible to do a meta-analysis whenever the estimate of the effect and its standard error is available for each contributing study. If y_k is the estimated effect and SE_k its standard error based on the kth study, the combined estimate and its standard error are given by

$$\widehat{y} = \frac{\sum w_k y_k}{\sum w_k},\ SE\left(\widehat{y}\right) = \frac{1}{\sqrt{\sum w_k}}, \tag{6}$$

where $w_k = \left(1/SE_k\right)^2$.

Survival Analysis

Survival analysis* (*see also* LONGITUDINAL DATA ANALYSIS) is a major area of epidemiological statistics concerned with measuring the risk of occurrence of an outcome event as a function of time. It thus focuses on the duration of time elapsed

from when a subject enters a study until the event occurs, and uses the survival curve to describe its distribution. Survival analysis is also concerned with the comparison of survival curves for different combinations of risk factors, and uses statistical methods to facilitate this comparison [14, 16, 18].

In general, survival analysis allows for the proper treatment of incomplete data due to subjects dropping into or out of the study, giving rise to censored (more precisely, right-censored) data. Survival data may be censored because *(i)* the subject withdraws from the study for any reason before experiencing the event (''loss to follow-up"), or *(ii)* an intervening event occurs prohibiting further observation on the subject, or *(iii)* the subject does not experience the event before the study ends (or before an analysis of the results is required).

When the event of interest occurs, the survival time is conventionally called a *failure time* (even though the event might be a "success", like recovery from some disease).

The *survival curve* is the proportion of subjects surviving beyond a given duration of time t. For a large population in which the survival times range continuously over an interval, this curve will be a smooth function of t that decreases from a maximum value of 1 when t is 0. In practice, the survival curve estimated from a sample of data is a step function that decreases only at the failure times [17].

A useful summary of survival that can be estimated directly from a survival curve is the *median survival time*. This is the survival time exceeded by 50% of the subjects, and is obtained by finding where the survival curve has the value 0.5.

Survival curves compare risks for different groups of subjects by showing the survival curves for the different groups on the same axes. Thus if one curve (for Group A, say) is entirely above another (Group B), then the subjects in Group A have better survival prospects than those in Group B. However, if the two curves cross, the situation is more complicated: it means that the relative risk of failure depends on how long a subject survives.

The logrank test [28] provides a P-value for testing the null hypothesis that two or more survival functions are identical. This is a special case of Birch's extension of the Mantel–Haenszel–Cochran test, which requires that the population odds ratios in

the different strata are the same. In survival analysis, this homogeneity assumption is called the *proportional hazards* assumption.

The proportional hazards model*, described in the next section, provides another method for analyzing survival data. It has the advantage that it can handle both continuous and nominal determinants simultaneously. However, in common with other statistical models, it makes additional assumptions about the associations and should be regarded as an accompaniment rather than an alternative to the methods described in this section.

Logistic Regression

Logistic regression* provides a method for modeling the association between a nominal outcome and multiple determinants. It is similar in many ways to linear regression*. For m determinants, it takes the form

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{j=1}^{m} \beta_j x_j. \tag{7}$$

The only other statistical assumption is that the outcomes are mutually independent. The model is known as *simple logistic regression* [19].

Logistic regression provides a further statistic, the *deviance**, which can assess the statistical significance of a set of determinants in the model. The deviance is defined as $-2\ln L$, where L is the likelihood associated with the data for the fitted parameters. Two logistic regression models are fitted to the data, one containing all the determinants of interest, and the other containing all the determinants except for those being assessed. Asymptotically, the difference between the values of the two deviances has a chi-squared distribution, with the number of degrees of freedom equal to the number of parameters in the determinants being assessed.

The logistic regression model described by Equation 6 can be extended to situations in which the outcome variable is nominal with more than two categories. If these outcome categories are coded as $0, 1, 2, \ldots, c$ and p_k is the probability that an outcome has the value k, the model takes the form, for $0 \leqslant k \leqslant c$,

$$p_k = \frac{\exp\left(\alpha_k + \sum_{j=1}^{m} \beta_{jk} x_j\right)}{1 + \sum_{k=1}^{c} \exp\left(\alpha_k + \sum_{j=1}^{m} \beta_{jk} x_j\right)}, \tag{8}$$

and is known as *polytomous logistic regression** [14].

For ordinal outcomes with more than two levels, the logistic model takes a different form. The outcome categories are again coded as $0, 1, 2, \ldots, c$ but p_k is now the probability that an outcome has value *at least* k. Thus, for $0 < k \leqslant c$, these probabilities are given by

$$\ln\left(\frac{p_k}{1 - p_k}\right) = \alpha_k + \sum_{j=1}^{m} \beta_j x_j. \tag{9}$$

This model incorporates an assumption of "proportional odds", meaning that for each determinant, the odds ratios are the same at each cut-point k.

Poisson Regression

The Poisson distribution* takes nonnegative integer values with probabilities

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{10}$$

The parameter λ is positive and is the mean of the distribution. This distribution plays an important role in epidemiological research, particularly in studies in which different subjects have different durations of exposure to a risk factor.

Suppose that the probability that a subject experiences an adverse event in a short interval of time δt is constant and equal to $\mu \delta t$. If the outcomes in different intervals are independent, the number of adverse events in a period T containing $N = T/\delta t$ short intervals has a binomial distribution ranging from 0 to N, with probabilities

$$p_k = \frac{N!}{k!(N-k)!} (\mu \delta t)^k (1 - \mu \delta t)^{N-k}.$$

Taking the limit as $\delta t \to 0$, we obtain the Poisson distribution with parameter $\lambda = \mu T$.

The Poisson regression* model now arises by expressing the incidence rate μ as the exponential of a linear function of determinants, that is,

$$\lambda = \mu T = \exp\left(\alpha + \sum_{j=1}^{m} \beta_j x_j\right) T. \tag{11}$$

If the determinants are nominal, Poisson regression is a limiting case of logistic regression, in which the probability of the adverse outcome is infinitesimally small and the number of subjects is correspondingly large. In logistic regression, each subject constitutes a separate experimental unit, whereas in Poisson regression an experimental unit corresponds to a short period of observation on each subject.

In the Poisson model, outcomes are interchangeable in the sense that a group of subjects has the same combination of risk factors, but only the total number of adverse events is important.

Proportional Hazards Regression

The Poisson regression model given by Equations 10 and 11 specifies the logarithm of the incidence density as a linear function of determinants. This incidence rate μ is the probability that the outcome event of interest occurs in a small interval of time $(t, t + \delta t)$ divided by the length of the interval (δt). Equation 11 can thus be written

$$\mu = \exp\left(\alpha + \sum_{j=1}^{m} \beta_j x_j\right).$$

In the context of survival analysis, this incidence rate is the *hazard function*, and is allowed to depend on the survival time t, simply by allowing the constant parameter α to be a function of t. This model, due to Cox [8], takes the form

$$h(t) = h_0(t) \exp\left(\sum_{j=1}^{m} \beta_j x_j\right), \tag{12}$$

where $h_0(t)$ is an arbitrary function of the survival time t, called the *baseline hazard function*. It is called the *proportional hazards model**, because the relative risk of an event for two subjects depends only on their determinants, and not on their duration of survival.

Note that if failures were known to occur at the times $t_1, t_2, \ldots, t_q$, $h_0(t)$ could be expressed as a step function with changes at these failure times, and Equation 12 would take the form

$$h(t) = \exp\left(\sum_{i=1}^{q} \alpha_i w_i\right) \exp\left(\sum_{j=1}^{m} \beta_j x_j\right),$$

where each w_i is an indicator variable taking the value 1 for a failure occurring at time t_i and 0 otherwise. It follows that the proportional hazards model is equivalent to the special case of a Poisson regression model in which an additional stratification variable corresponding to the failure time is included.

Correlated Outcomes

All of the statistical methods described in the preceding sections assume that the outcomes are independent. In practice, data collected in epidemiological studies are often correlated due to clustering. Spatial clustering occurs when subjects are sampled from villages or families sharing particular attributes, or when repeated measurements are taken on the same subjects. Clustering also occurs in time due to seasonal effects.

Standard errors of proportions based on correlated data can be corrected using variance inflation factors [29], which specify by how much the sample size of a cluster needs to be increased to compensate for the clustering. In the simplest case, when the correlations between binary outcomes in clusters of size m have equal correlation ρ, the variance inflation factor is $1 + (m - 1)\,\rho$, and the standard error of the log odds ratio is increased by the square root of this factor. This formula shows that even small correlations between outcomes can have a substantial effect in large clusters. For example, if $\rho = 0.1$ and $m = 31$, the standard error is doubled, and the sample size would need to be quadrupled to compensate for the clustering.

Zeger & Liang [39] invented a general and robust method for handling correlated outcomes, which is now widely used in epidemiological statistics. This method, known as generalized estimating equations* (GEE), can be applied to all the models described in this section, and has generated a great deal of research both in statistics and epidemiology [11].

References

[1] Armitage, P. and Hills, M. (1982). The two-period cross-over trial. *The Statistician*, **31**, 119–131.

[2] Birch, M.W. (1965). The detection of partial association II: the general case. *J. R. Stat. Soc. B*, **27**, 111–124.

[3] Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research: Volume I - The Analysis of Case-Control Studies*. IARC, Lyon.

[4] Breslow, N.E. and Day, N.E. (1987). *Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies*. IARC, Lyon.

[5] Brown, B.W. Jr. (1980). The crossover experiment for clinical trials. *Biometrics*, **36**, 69–70.

[6] Clayton, D.G. (1974). Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika*, **61**, 525–531.

[7] Cochran, W.G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*, **10**, 417–451.

[8] Cox, D.R. (1972). Regression models and life tables (with discussion). *J. R. Stat. Soc. B*, **34**, 187–220.

[9] Cox, D.R. and Wermuth, N. (1996). *Multivariate Dependencies—Models, Analysis and Interpretation*. Chapman & Hall. London.

[10] Easterbrook, P.J., Berlin, J.A, Gopalan, R., and Matthews, D.R. (1991). Publication bias in clinical research. *The Lancet*, **387**(8746), 867–872.

[11] Hardin, J.W. and Hilbe, J.M. (2003). *Generalized Estimating Equations*. Chapman & Hall/CRC•.

•Q2

[12] Hill, A.B. (1951). The clinical trial. *Br. Med. Bull.*, **7**, 278–287.

[13] Hodges, L.V. and Olkin, I. (1885). *Statistical Methods for Meta-analysis*. Academic Press, New York.

[14] Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

[15] Hosmer, D.W. and Lemeshow, S. (1999). *Applied Survival Analysis—Regression Modelling of Time to Event Data*. Wiley, New York.

[16] Kalbfleisch, J. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley. New York.

[17] Kaplin, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–481.

[18] Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis: Techniques for censored and Truncated Data*. Springer-Verlag, New York.

[19] Kleinbaum, D.G. and Klein, M. (2002). *Logistic Regression: A Self-Learning Text*, 2nd ed. Springer-Verlag, New York.

[20] Louis, P.C.A. (1834). *An Essay on Clinical Instruction*. Translated by Martin P. Highley, London•, pp. 26–27.

•Q3

[21] Louis, P.C.A. (1836). *Researches on the Effects of Bloodletting in some Inflammary Diseases, and on the Influence of Tartarized Antimony and Vesication in Pneumonius*. Translated by C.G. Putnum, Hilliard Gray, Boston, Mass.

[22] Lwanga, S.K. and Lemeshow, S. (1991). *Sample Size Determination in Health Studies. A Practical Manual*. WHO, Geneva.

[23] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.*, **22**, 719–748.

[24] Meier, P. (1972). "The Biggest Public Health Experiment Ever. The 1954 Field Trial Of The Salk Poliomyelitis Vaccine". In *Statistics. A Guide to the Unknown*, J.M. Tanur, ed. Holden Day, San Francisco, Calif., pp. 2–13.

[25] Newman, S.C. (2001). *Biostatistical Methods in Epidemiology*. Wiley, New York.

[26] Normand, S.L. (1999). Meta-analysis. formulating, evaluating, combining, and reporting. *Stat. Med.*, **18**(3), 321–359.

[27] Peto, R. (1987). Why do we need systematic overviews of randomized trials? *Stat. Med.*, **6**, 233–240.

[28] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariance test procedures (with discussion). *J. R. Stat. Soc. A*, **135**, 185–206.

[29] Rao, J.N.K. and Scott, A.J. (1992). A simple method for the analysis of clustered binary data. *Biometrics*, **48**, 577–585.

[30] Robertson, B., Fairley, C.K., Black, J., and Sinclair, M. (2003). "Case-Control Studies". In *Drinking Water and Infectious Disease. Establishing the Links*, P.R. Hunter, M. Waite, and E. Ronchi, eds. CRC Press, pp. 175–182.

[31] Robins, J., Breslow, N., and Greenland S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42**, 311–323.

[32] Sahai, H. and Khurshid. A. (1996). *Statistics in Epidemiology: Methods, Techniques, and Applications*. CRC Press, Boca Raton, Fla.

[33] Sackett, D.L. (1979). Bias in analytic research. *J. Chronic Dis.*, **32**, 51–63.

[34] Schlesselman, J. (1982). *Case-Control Studies*. Oxford University Press, New York.

[35] Senn, S. (1990). *Cross-over Trials in Clinical Research*. Wiley, New York.

[36] Snow, J. (1855). *On the mode of communication of cholera*. Churchill, London. (Reprinted in Snow on cholera: a reprint of two papers. Hafner, New York, 1965).

[37] Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., and Song, F. (2000). *Methods for Meta-Analysis in Medical Research*. John Wiley & Sons.

[38] Woodward, M. (1999). *Epidemiology: Study Design and Data Analysis*. Chapman and Hall/CRC, Boca Raton, Fla.

[39] Zeger, S. and Liang, K. (1966). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

(BIOSTATISTICS, CLASSICAL
CLINICAL TRIALS
COHORT ANALYSIS
DEMOGRAPHY
EPIDEMICS
FRAMINGHAM: AN EVOLVING LONGITUDINAL STUDY
LOGISTIC REGRESSION
LONGITUDINAL DATA ANALYSIS
MANTEL-HAENSZEL STATISTIC
MATCHED PAIRS
MEDICAL DIAGNOSIS, STATISTICS IN
MEDICINE, STATISTICS IN
POISSON REGRESSION
PROPORTIONAL HAZARDS MODEL, COX'S
PROSPECTIVE STUDIES
RETROSPECTIVE STUDIES
RELATIVE RISK
SELECTION BIAS
SURVIVAL ANALYSIS)