Rates, Risks, Measures of Association and Impact

Jacques Benichou, Mari Palta

Introduction

2.1

A major aim of epidemiologic research is to measure disease occurrence in relation to various characteristics such as exposure to environmental, occupational, or lifestyle risk factors, genetic traits or other features. In this chapter, various measures will be considered that quantify disease occurrence, associations between disease occurrence and these characteristics as well as their consequences in terms both of disease risk and impact at the population level. As is common practice, the generic term exposure will be used throughout the chapter to denote such characteristics. Emphasis will be placed on measures based on occurrence of new disease cases, referred to as disease incidence. Measures based on disease prevalence, i.e., considering newly occurring and previously existing disease cases as a whole will be considered more briefly.

We will first define the basic measure of disease incidence, namely the incidence rate, from which other measures considered in this chapter can be derived. These other measures, namely measures of disease risk, measures of association between exposure and disease risk (e.g., relative risk), and measures of impact of exposure-disease associations (e.g., attributable risk) will be considered successively. Additional points will be made regarding standardized incidence rates and measures based on prevalence.

Incidence and Hazard Rates

2.2

Definition

2.2.1

The incidence rate of a given disease is the number of persons who develop the disease (number of incident cases) among subjects at risk of developing the disease in the source population over a defined period of time or age. Incidence rates are not interpretable as probabilities. While they have a lower bound of zero, they have no upper bound. Units of incidence rates are reciprocal of person-time, such as reciprocals of person-years or multiples of person-years (e.g., 100,000 person-years). For instance, if 10 cases develop from the follow-up of 20 subjects and for a total follow-up time of five years, the incidence rate is $10/100 = 0.1$ cases per person-year (assuming an instantaneous event with immediate recovery and all 20 subjects being at risk until the end of the observation period).

Usually, incidence rates are assessed over relatively short time periods compared with the time scale for disease development, e.g., intervals of five-years for chronic diseases with an extended period of susceptibility such as many cancers.

Synonyms for incidence rate are average incidence rate, force of morbidity, person-time rate, or incidence density (Miettinen 1976), the last term reflecting the interpretation of an incidence rate as the density of incident case occurrences in an accumulated amount of person-time (Morgenstern et al. 1980).

Mortality rates (overall or cause-specific) can be regarded as a special case of incidence rates, the outcome considered being death rather than disease occurrence.

Incidence rates can be regarded as estimates of a limiting theoretical quantity, namely the hazard rate, $h(t)$, also called the incidence intensity or force of morbidity. The hazard rate at time t, $h(t)$, is the instantaneous rate of developing the disease of interest in an arbitrarily short interval Δ around time t, provided the subject is still at risk at time t (i.e., has not fallen ill before time t). Technically, it has the following mathematical definition:

$$h(t) = \mathrm{limit}_{\Delta \downarrow 0} \Delta^{-1} \Pr(t \leq T < t + \Delta | t \leq T) \,, \tag{2.1}$$

where T is the time period for the development of the disease considered and Pr denotes probability. Indeed, for time intervals in which the hazard rate can be assumed constant, the incidence rate as defined above represents a valid estimate of the hazard rate. Thus, this result applies when piecewise constant hazards are assumed, which can be regarded as realistic in many applications, especially when reasonably short time intervals are used, and leads to convenient estimating procedures, e.g., based on the Poisson model.

Strictly speaking, incidence and hazard rates do not coincide. Hazard rates are formally defined as theoretical functions of time whereas incidence rates are defined directly as estimates and constitute valid estimates of hazard rates under certain assumptions (see above). For the sake of simplicity however, we will use the terms incidence rates and hazard rates as synonyms in the remainder of this chapter unless a clear distinction is needed.

2.2.2 Estimability and Basic Principles of Estimation

From the definitions above, it ensues that individual follow-up data are needed to obtain incidence rates or estimate hazard rates. Alternatively, in the absence of individual follow-up data, person-time at risk can be estimated as the time period width times the population size at midpoint. Such estimation makes the assumption that individuals who disappear from being at risk, either because they succumb, or because they move in or out, do so evenly across the time interval. Thus, population data such as registry data can be used to estimate incidence rates as long as an exhaustive census of incident cases can be obtained.

Among the main designs considered in Part I of this handbook, the cohort design (cf. Chap. I.5) is the ideal design to obtain incidence or hazard rates for various levels or profiles of exposure, i.e., exposure-specific incidence or hazard rates. This is because follow-up is available on subjects with various profiles of exposure. In many applications, obtaining exposure-specific incidence rates is not trivial however. Indeed, several exposures are often considered, some with several exposed levels and some continuous. Moreover, it may be necessary to account for confounders or effect-modifiers. Hence, estimation often requires modeling.

Methods of inference based on regression models are considered in detail in Part II of this handbook, particularly Chaps. II.3 and II.4.

Case-control data (cf. Chap. I.6) pose a more difficult problem than cohort data because case-control data alone are not sufficient to yield incidence or hazard rates. Indeed, they provide data on the distributions of exposure respectively in diseased subjects (cases) and non-diseased subjects (controls) for the disease under study, which can be used to estimate odds ratios (see Sect. 2.4.3) but are not sufficient to estimate exposure-specific incidence rates. However, it is possible to arrive at exposure-specific incidence rates from case-control data if case-control data are complemented by either follow-up or population data, which happens for nested or population-based case-control studies. In a nested case-control study, the cases and controls are selected from a follow-up study. In a population-based case-control study, they are selected from a specified population in which an effort is made to identify all incident cases diagnosed during a fixed time interval, usually in a grouped form (e.g., number of cases and number of subjects by age group). In both situations, full information on exposure is obtained only for cases and controls. Additionally, complementary information on composite incidence (i.e., counts of events and person-time) can be sought from the follow-up or population data. By combining this information with odds ratio estimates, exposure-specific incidence rates can be obtained. This has long been recognized (Cornfield 1951, 1956; MacMahon 1962; Miettinien 1974, 1976; Neutra and Drolette 1978) and is a consequence of the relation (Miettinen 1974; Gail et al. 1989):

$$h_0 = h^*(1 - \mathrm{AR}) \,, \tag{2.2}$$

where AR is the attributable risk in the population for all exposures considered, a quantity estimable from case-control data (see Sect. 2.5.1), h_0 is the baseline incidence rate, i.e., the incidence rate for subjects at the reference (unexposed) level of all exposures considered and h^* is the composite or average incidence rate in the population that includes unexposed subjects and subjects at various levels of all exposures (i.e., with various profiles of exposure). The composite incidence rate h^* can be estimated from the complementary follow-up or population data. Equation (2.2) simply states that the incidence rate for unexposed subjects is equal to the proportion of the average incidence rate in the population that is not associated with any of the exposures considered. Equation (2.2) can be specialized to various subgroups or strata defined by categories of age, sex or geographic location such as region or center, on which incidence rates are assumed constant. From the baseline rate h_0, incidence rates for all levels or profiles of exposure can be derived using odds ratio estimates, provided odds ratio estimates are reasonable estimates of incidence rate ratios as in the case of a rare disease (see Sect. 2.4). Consequently, exposure-specific incidence rates can be obtained from case-control data as long as they are complemented by follow-up or population data that can be used to estimate average incidence rates.

Example 1 . Exposure-specific incidence rates of breast cancer were obtained based on age as well as family history in first-degree relatives, reproductive history (i.e., age at menarche and age at first live birth), and history of benign disease from the Breast Cancer Detection and Demonstration Project (BCDDP). The BCDDP combined the prospective follow-up of 284,780 women over five years, and a nested case-control study (Gail et al. 1989) with about 3000 cases and 3000 controls. For each five-year age group from ages 35 to 79 years, composite incidence rates were obtained from the follow-up data. In age groups 40–44 and 45–49 years, 162 and 249 new cases of breast cancer developed from the follow-up of 79,526.4 and 88,660.7 person-years, yielding composite incidence rates of 203.7 and 280.8 per 10^5 person-years, respectively. For all women less than 50 years of age, the attributable risk for family history, reproductive history and history of benign breast disease was estimated at 0.4771 from the nested case-control data (see Sect. 2.5.1). By applying (2.2), baseline incidence rates for women at the reference level of all these factors were $203.7 \times (1 - 0.4771) = 106.5$ and $280.8 \times (1 - 0.4771) = 146.8$ per 10^5 person-years, respectively. For a nulliparous woman of age 40, with menarche at age 12, one previous biopsy for benign breast disease, and no history of breast-cancer in her first-degree relatives, the corresponding odds ratio was estimated at 2.89 from logistic regression analysis of the nested case-control data (see Sect. 2.4.6), yielding an exposure-specific incidence rate of $106.5 \times 2.89 = 307.8$ per 10^5 person-years. For a 45-year old woman with the same exposure profile, the corresponding exposure-specific incidence rate was $146.8 \times 2.89 = 424.3$ per 10^5 person-years. ♦

Finally, cross-sectional data cannot provide any assessment of incidence rates but instead will yield estimates of disease prevalence proportions as discussed in Sect. 2.6 of this chapter.

2.2.3 Relation with Other Measures

The reason why exposure-specific incidence or hazard rates are central quantities is that, once they are available, most other quantities described in this chapter can be obtained from them, namely measures of disease risk, measures of association between exposure and disease risk, and measures of exposure impact in terms of new disease burden at the population level. However, it should be noted that measures of impact as well as some measures of association (i.e., odds ratios) can be estimated from case-control data alone without relying on exposure-specific incidence rates (see Sects. 2.3 and 2.4). Moreover, cross-sectional data can yield estimates of measures of association and impact with respect to disease prevalence (see Sect. 2.6.2).

Measures of Disease Risk

2.3

Definition

2.3.1

Disease risk is defined as the probability that an individual who is initially disease-free will develop a given disease over a specified time or age interval (e.g., one year or lifetime). Of all incidence and risk measures, this measure is probably the one most familiar and interpretable to most consumers of health data.

If the interval starting at time a_1 and ending just before time a_2, i.e., $[a_1, a_2)$, is considered, disease risk can be written formally as:

$$\pi(a_1, a_2) = \int_{a_1}^{a_2} h(a)\{S(a)/S(a_1)\}\,\mathrm{d}a\,. \qquad (2.3)$$

In (2.3), $h(a)$ denotes the disease hazard at time or age a (see Sect. 2.2). The function $S(\cdot)$, with $(\cdot)$ an arbitrary argument, is the survival function, so that $S(a)$ denotes the probability of still being disease-free at time at age a, and $S(a)/S(a_1)$ denotes the conditional probability of staying disease-free up to time or age a for an individual who is free of disease at the beginning of the interval $[a_1, a_2)$. Equation (2.3) integrates over the interval $[a_1, a_2)$ the instantaneous incidence rate of developing disease at time or age a for subjects still at risk of developing the disease (i.e., subjects still disease-free). Because the survival function $S(\cdot)$ can be written as a function of disease hazard through:

$$S(a_2)/S(a_1) = \exp\left\{-\int_{a_1}^{a_2} h(a)\,\mathrm{d}a\right\}\,, \qquad (2.4)$$

disease risk is also a function of disease hazard.

By specializing the meaning of functions $h(\cdot)$ and $S(\cdot)$, various quantities can be obtained that measure disease risk in different contexts. First, the time scale on which these functions as well as disease risk are defined corresponds to two specific uses of risk. In most applications, the relevant time scale is age, since disease incidence is influenced by age in most applications. Note that by considering the age interval $[0, a_2)$, one obtains lifetime disease risk up to age a_2. However, in clinical epidemiology settings, risk refers to the occurrence of an event, such as relapse or death in subjects already presenting with the disease of interest. In this context, the relevant time scale becomes time from disease diagnosis or, possibly, time from some other disease-related event, such as a surgical resection of a tumor or occurrence of a first myocardial infarction.

Second, risk definition may account or not for individual exposure profiles. If no risk factors are considered to estimate disease hazard, the corresponding measure of disease risk defines the average or composite risk over the entire population that includes subjects with various exposure profiles. This measure, also called cumulative incidence (Miettinen 1976), may be of value at the population level.

However, the main usefulness of risk is in quantifying an individual's predicted probability of developing disease depending on the individual's exposure profile. Thus, estimates of exposure-specific disease hazard have to be available for such exposure-specific risk (also called individualized or absolute risk) to be estimated.

Third, the consideration of competing risks and the corresponding definition of the survival function $S(\cdot)$ yields two separate definitions of risk. Indeed, although risk is defined with respect to the occurrence of a given disease, subjects can die from other causes (i.e., competing risks), which obviously precludes disease occurrence. The first option is to define $S(a)$ as the theoretical probability of being disease-free at time or age a if other causes of death (competing risks) were eliminated yielding a measure of disease risk in a setting with no competing risks. This measure may not be of much practical value. Moreover, unless unverifiable assumptions regarding incidence of the disease of interest and deaths from other causes can be made, for instance assuming that they occur independently, the function $S(\cdot)$ will not be estimable. For these reasons, it is more feasible to define $S(a)$ as the probability that an individual will be alive and disease-free at age a as the second option, yielding a more practical definition of disease risk as the probability of developing disease in the presence of competing causes of death (see Sect. 2.3.5).

From the definition of disease risk above, it appears that disease risk depends on the incidence rate of disease in the population considered and can also be influenced by the strength of the relationship between exposures and disease if individual risk is considered. One consequence is that risk estimates may not be portable from one population to another, as incidence rates may vary widely among populations that are separated in time and location or even among subgroups of populations, possibly because of differing genetic patterns or differing exposure to unknown risk factors. Additionally, competing causes of death (competing risks) may also have different patterns among different populations, which might also influence values of disease risk.

2.3.2 Range

Disease risk is a probability and therefore lies between 0 and 1, and is dimensionless. A value of 0 while theoretically possible would correspond to very special cases such as a purely genetic disease for an individual not carrying the disease gene. A value of 1 would be even more unusual and might again correspond to a genetic disease with a penetrance of 1 for a gene carrier but, even in this case, the value should be less than 1 if competing risks are accounted for.

2.3.3 Synonyms

Beside the term "disease risk", "absolute risk" or "absolute cause-specific risk" have been used by several authors (Dupont 1989; Benichou and Gail 1990a, 1995; Benichou 2000a; Langholz and Borgan 1997). Alternative terms include "individualized risk" (Gail et al. 1989), "individual risk" (Spiegelman et al. 1994), "crude probability" (Chiang 1968), "crude incidence" (Korn and Dorey 1992), "cumulative incidence" (Gray 1988; Miettinen 1976), "cumulative incidence risk" (Miettinen 1974) and "absolute incidence risk" (Miettinen 1976).

The term "cumulative risk" refers to the quantity $\int_{a_1}^{a_2} h(a)\,\mathrm{d}a$ and approximates disease risk closely in the case where disease is rare.

The term "attack rate" defines the risk of developing a communicable disease during a local outbreak and for the duration of the epidemic or the time during which primary cases occur (MacMahon and Pugh 1970, Chap. 5; Rothman and Greenland 1998, Chap. 27).

The term "floating absolute risk", introduced by Easton et al. (1991), refers to a different concept from disease risk. It was derived to remedy the standard problem that measures of association such as ratios of rates, risks or odds are estimated in reference to a baseline group, which causes their estimates for different levels of exposure to be correlated and may lead to lack of precision if the baseline group is small. The authors proposed a procedure to obtain estimates unaffected by these problems and used the term "floating absolute risk" to indicate that standard errors were not estimated in reference to an arbitrary baseline group.

Interpretation and Usefulness 2.3.4

If exposure profiles are not taken into account, the resulting average risk has little usefulness in disease prediction. Average risk estimates may be useful only for diseases for which no risk factors have been identified. Otherwise, they only provide overall results such as "one in nine women will develop breast cancer at sometime during her life" (American Cancer Society 1992), which are of no direct use in quantifying the risk of women with given exposure profiles and no direct help in deciding on preventive treatment or surveillance measures.

Upon taking individual exposure profiles into account, resulting individual disease risk estimates become useful in providing an individual measure of the probability of disease occurrence, and can therefore be useful in counseling. They are well suited to predicting risk for an individual, unlike measures of association that quantify the increase in the probability of disease occurrence relative to subjects at the baseline level of exposure, but do not quantify that probability itself.

Individual risk has been used as a tool for individual counseling in breast cancer (Benichou et al. 1996; Gail and Benichou 1994; Hoskins et al. 1995). Indeed, a woman's decision to take a preventive treatment such as Tamoxifen (Fisher et al. 1998; Wu and Brown 2003) or even undergo prophylactic mastectomy (Hartman et al. 2001; Lynch et al. 2001) depends on her awareness of the medical options, on personal preferences, and on individual risk. A woman may have several risk factors, but if her individual risk of developing breast cancer over the next 10 years is small, she may be reassured and she may be well advised simply to embark on a program of surveillance. Conversely, she may be very concerned about her absolute risk over a longer time period, such as 30 years, and she may decide to use prophylactic medical treatment or even undergo prophylactic mastectomy if her absolute risk is very high.

Estimates of individual risk of breast cancer are available based on age, family history, reproductive history and history of benign disease (Gail et al. 1989; Costantino et al. 1999) and were originally derived from the BCDDP that combined a follow-up study and a nested case-control study (Gail et al. 1989). This example illustrates that not only exposures or risk factors per se (such as family history) may be used to obtain individual risk estimates but also markers of risk such as benign breast disease which are known to be associated with an increase in disease risk and may reflect some premalignant stage. In the same fashion, it has been suggested to improve existing individual risk estimates of breast cancer by incorporating mammographic density, a risk marker known to be associated with increased breast cancer risk (Benichou et al. 1997). In the cardiovascular field, individual risk estimates of developing myocardial infarction, developing coronary heart disease, dying from coronary heart disease, developing stroke, developing cardiovascular disease, and dying from cardiovascular disease were derived from the Framingham heart and Framingham offspring cohort studies. These estimates are based on age, sex, HDL, LDL and total cholesterol levels, smoking status, blood pressure and diabetes history (Anderson et al. 1991).

Individual risk is also useful in designing and interpreting trials of interventions to prevent the occurrence of a disease. At the design stage, disease risk may be used for sample size calculations because the sample sizes required for these studies depend importantly on the risk of developing the disease during the period of study and the expected distribution of exposure profiles in the study sample (Anderson et al. 1992). Disease risk has also been used to define eligibility criteria in such studies. For example, women were enrolled in a preventive trial to decide whether the drug Tamoxifen can reduce the risk of developing breast cancer (Fisher et al. 1998). Because Tamoxifen is a potentially toxic drug and because it was to be administered to a healthy population, it was decided to restrict eligibility to women with somewhat elevated absolute risks of breast cancer. All women over age 59 as well as younger women whose absolute risks were estimated to equal or exceed that of a typical 60-year old woman were eligible to participate (Fisher et al. 1998). Individual risk has been used to interpret results of this trial through a risk-benefit analysis in order to help define which women are more likely to benefit from using Tamoxifen. Women were identified, who had a decrease in breast cancer risk and other events such as hip fracture from using Tamoxifen surpassing the Tamoxifen-induced increase in other events such as endometrial cancer, pulmonary embolism or deep vein thrombosis (Gail et al. 1999).

Disease risk can also be important in decisions affecting public health. For example, in order to estimate the absolute reduction in lung cancer incidence that might result from measures to reduce exposure to radon, one could categorize a general population into subgroups based on age, sex, smoking status and current radon exposure levels and then estimate the absolute reduction in lung cancer incidence that would result from lowering radon levels in each subgroup (Benichou and Gail 1990a; Gail 1975). Such an analysis would complement estimation of population attributable risk or generalized impact fractions (see Sect. 2.5).

The concept of risk is also useful in clinical epidemiology as a measure of the individualized probability of an adverse event, such as a recurrence or death in diseased subjects. In that context, risk depends on factors that are predictive of recurrence or death, rather than on factors influencing the risk of incident disease, and the time-scale of interest is usually time from diagnosis or from surgery rather than age. It can serve as a useful tool to help define individual patient management and, for instance, the absolute risk of recurrence in the next three years might be an important element in deciding whether to prescribe an aggressive and potentially toxic treatment regimen (Benichou and Gail 1990a; Korn and Dorey 1992).

Properties 2.3.5

Two main points need to be emphasized. First, as is evident from its definition, disease risk can only be estimated and interpreted in reference to a specified age or time interval. One might be interested in short time spans (e.g., five years), or long time spans (e.g., 30 years). Of course, disease risk increases as the time span increases. Sometimes, the time span is variable such as in lifetime risk.

Disease risk can be influenced strongly by the intensity of competing risks (typically competing causes of death, see above). Disease risk varies inversely as a function of death rates from other causes.

Estimability 2.3.6

It follows from its definition that disease risk is estimable as long as hazard rates for the disease (or event) of interest are estimable. Therefore, disease risk is directly estimable from cohort data, but case-control data have to be complemented with follow-up or population data in order to obtain the necessary complementary information on incidence rates (see Sect. 2.2.2).

It has been argued above (see Sect. 2.3.1) that disease risk is a more useful measure when it takes into account competing risks, that is the possibility for an individual to die of an unrelated disease before developing the disease (or disease-related event) of interest. In this setting, disease risk is defined as the probability of disease occurrence *in the presence* of competing risks, which is more relevant for individual predictions and other applications discussed above than the underlying (or "net" or "latent") probability of disease occurrence *in the absence* of competing risks. Moreover, disease risk is identifiable without any unverifiable competing risk assumptions in this setting, such as the assumption that competing risks act independently of the cause of interest because, as Prentice et al. (1978) emphasize, all functions of the disease hazard rates are estimable. Death rates from other causes can be estimated either internally from the study data or from external sources such as vital statistics.

Example 1. *(continued)*
In order to obtain estimates of breast cancer risk in the presence of competing risks, Gail et al. (1989) used 1979 United States (US) mortality rates from year 1979 for all causes except breast cancer to estimate the competing risks with more precision than from the BCDDP follow-up data. In age groups 40–44 and 45–49 years, these death rates were 153.0 and 248.6 per 10^5 person-years, respectively, hence of the same order of magnitude as breast cancer incidence rates. In older age groups, these death rates were much higher than breast cancer incidence rates, thus strongly influencing breast cancer risk estimates for age intervals including these age groups. For instance, death rates from causes other than breast cancer were 1017.7 and 2419.8 per 10^5 person-years in age groups 65–69 and 70–74 years, respectively, whereas average incidence rates of breast cancer were 356.1 and 307.8 per 10^5 person-years in these age groups, respectively. ♦

2.3.7 Estimation from Cohort Studies

Estimation of disease risk rests on estimating disease incidence and hazard rates, a topic also addressed in Part II of this handbook. Several approaches have been worked out fully for disease risk estimation. A brief review of these approaches is given here starting with average risk estimates that do not take exposure profiles into account and continuing with exposure-specific estimates.

Estimates of Average Disease Risk

The density or exponential method (Miettinen 1976; Kleinbaum et al. 1982, Chap. 6; Rothman and Greenland 1998, Chap. 3) relies on subdividing the time or age scale in successive time or age intervals $I_1, \ldots, I_i, \ldots, I_I$ (e.g., one- or five-year intervals) on which the rate of disease incidence is assumed constant (i.e., piecewise constant). Disease risk over time or age interval $[a_1, a_2)$, that is the probability for an individual to experience disease occurrence over interval $[a_1, a_2)$ is taken as one minus the probability of staying disease-free through the successive intervals included in $[a_1, a_2)$. Assuming that disease is rare on each of the successive intervals considered, disease risk can be estimated as:

$$\widehat{\pi}(a_1, a_2) = 1 - \exp\left(-\sum_i \widehat{h}_i \Delta_i\right) . \qquad (2.5)$$

The sum is taken over all intervals included in $[a_1, a_2)$. Notation Δ_i denotes the width of interval i, whereas $\widehat{h}_i$ denotes the incidence rate in interval i, obtained as the ratio of the number of incident cases over the person-time accumulated during follow-up in that interval.

While (2.5) is simple to apply, its validity depends on several assumptions. The assumption that disease incidence is constant over each time or age interval considered makes it a parametric approach. However, if intervals are small enough, this will not amount to a strong assumption. Moreover, it relies on the assumption that disease incidence is small on each interval. If this is not the case, a more complicated formula will be needed. Finally, this approach ignores competing risks.

Benichou and Gail (1990a) generalized this approach by lifting the condition on small incidence on each interval and allowing competing risks to be taken into account. They derived a generalized expression for the estimate of disease risk over time or age interval $[a_1, a_2)$ as:

$$\widehat{\pi}(a_1, a_2) = \sum_i \frac{\widehat{h}_{1i}}{\widehat{h}_{1i} + \widehat{h}_{2i}} \left[1 - \exp\left\{-\left(\widehat{h}_{1i} + \widehat{h}_{2i}\right)\Delta_i\right\}\right] A(i) \ , \tag{2.6}$$

$$\text{with} \quad A(i) = \prod_{j<i} \exp\left\{-\left(\widehat{h}_{1j} + \widehat{h}_{2j}\right)\Delta_j\right\} \ .$$

In (2.6), the sum is taken over all intervals included in $[a_1, a_2), \Delta_i$ denotes the width of interval i, $\widehat{h}_{1i}$ denotes the disease incidence rate in interval i, $\widehat{h}_{2i}$ the death rate from other causes in interval i, and the product in $A(i)$ is taken over time intervals in $[a_1, a_2)$ from the first one to the one just preceding interval i. Death rates can be obtained in a similar fashion as disease incidence rates. It should be noted that disease risk can be estimated for a much longer duration than the actual follow-up of individuals in the study if age is the time scale (open cohort) provided there is no secular trend in disease incidence.

Variance estimates were derived by Benichou and Gail (1990a). Moreover, based on simulations of a closed cohort, they found that resulting confidence intervals have satisfactory coverage, especially with the log transformation, and observed little or no bias on risk estimates with a sufficient number of intervals even when disease incidence varied sharply with time.

The actuarial method or life table method (Cutler and Ederer 1958; Elveback 1958; Fleiss et al. 1976; Kleinbaum et al. 1982, Chap. 6; Rothman and Greenland 1998, Chap. 3) shares similarities with the density method, although it was derived from a less parametric viewpoint. As with the density method, time is split into intervals. In each time interval i, the probability for an individual who is disease-free at the beginning of the interval to stay disease-free throughout the interval is estimated. Disease risk is obtained as one minus the estimated probability of staying disease-free throughout the successive time intervals included in $[a_1, a_2)$ as:

$$\widehat{\pi}(a_1, a_2) = 1 - \prod_i \frac{(n_i - w_i/2 - d_i)}{(n_i - w_i/2)} \ , \tag{2.7}$$

where the product is taken over all intervals included in $[a_1, a_2)$, n_i denotes the number of disease-free subjects at the beginning of interval i, d_i the number of incident cases occurring in interval i, and w_i the number of subjects either lost to follow-up or dying from other causes (competing risks) in interval i. The actuarial approach is most appropriate when grouped data are available and the actual follow-up of each individual in each interval is not known. The person-years of follow-up for subjects lost to follow-up or affected with competing risks in interval i is not used directly but, if one assumes that the mean withdrawal time occurs at the midpoint of the interval, then the denominator in each product term of (2.7) can be regarded as the effective number of persons at risk of developing the disease in the corresponding interval. Namely, it represents the number of disease-free persons that would be expected to produce d_i incident cases if all persons could be followed for the entire interval (Elandt-Johnson 1977; Kleinbaum et al. 1982, Chap. 6; Littell 1952). The actuarial method can be regarded as a refinement of the simple cumulative method (Kleinbaum et al. 1982, Chap. 6) that ignores quantity w_i and simply estimates disease risk as the number of individuals who contract the disease, divided by the total number in the cohort, or exposure subgroup of interest. The actuarial method is preferable to this direct method because, in practice, it is rare that a large enough cohort can be followed over a long enough time to reliably estimate the risk of disease by this simple method. Moreover, the simple cumulative method cannot handle the case when subjects are followed for varying lengths of time, which often occurs because subjects can be enrolled at different times whereas the follow-up ends at the same time for all subjects.

As shown by several authors (Cutler and Ederer 1958; Fleiss et al. 1976), the actuarial method results in biased estimates of risk even in the unlikely and most favorable event (in terms of bias) of all withdrawals occurring at the interval midpoints. Alternative approaches based on different choices of the quantity to subtract from n_i (i.e., choices different from $w_i/2$) are not subject to less bias, however (Elandt-Johnson 1977). The problem can be best handled by using narrow intervals but this is done at the expense of a larger random error (i.e., less precise estimates of risk).

Compared to the density method ((2.5) and (2.6)), the actuarial method has the advantage of not requiring knowledge of individual follow-up times in each interval but only knowledge of the number at risk at the beginning of the interval and the number of withdrawals. The density method could be used however without knowledge of follow-up time by assigning a follow-up time of half the interval width to subjects who are lost to follow-up, develop disease or die from other causes, in an analogous fashion as with the actuarial method (Benichou and Gail 1990a). The actuarial method requires neither the assumption of constant incidence rate nor rarity of disease incidence on all time intervals. However, bias is less of a problem with the density than the actuarial method and the density method applies naturally to open cohorts and extends easily to risk estimates that take exposure profiles into account (see below).

When individual follow-up times are all known, a fully nonparametric risk estimate can be obtained in the spirit of the Kaplan–Meier estimate of survival (Kaplan and Meier 1958; see also Chap. II.4 of this handbook). Disease risk is estimated through summation on all distinct times in $[a_1, a_2)$ at which new disease cases occur (Aalen and Johansen 1978; Kay and Schumacher 1983; Gray 1988; Matthews 1988; Keiding and Andersen 1989; Benichou and Gail 1990a; Korn and Dorey 1992). Corresponding variance estimates were derived (Aalen 1978; Aalen and Johansen 1978; Keiding and Andersen 1989; Benichou and Gail 1990a; Korn and Dorey 1992) from which confidence intervals can be obtained, based on the log transformation as suggested by Benichou and Gail (1990a) and Keiding and Andersen (1989), or based on the approach of Dorey and Korn (1987).

Upon comparing the generalized density method (see (2.6)) and the nonparametric method, Benichou and Gail (1990a) showed that the loss of efficiency of the nonparametric method is small compared to the density method. Moreover, the nonparametric method yields little bias in risk estimates as well nearly nominal coverage for confidence intervals of risk with the log transformation. Nominal coverage refers to the theoretical probability of a confidence interval to cover the true parameter and may be assessed using simulations (i.e., a 95% confidence interval will be said to have nominal coverage if it does include the true parameter value in 95% of the cases). Hence, properties of the generalized density and nonparametric methods agree closely. However, the generalized density method has the advantage of simplicity of computation and is better suited to open cohorts.

Estimates of Exposure-specific Disease Risk

In order to obtain risk estimates that depend on exposure profiles, the cohort could be subdivided into subcohorts based on exposure levels and the methods above applied to these subcohorts. However, this approach would be impractical because it would yield risk estimates with very low precision. In order to remedy this problem, a natural approach to incorporate exposures is to model incidence rates through regression models.

Benichou and Gail (1990a) proposed a direct extension of the generalized density method (2.6). This extension is based on assuming that the disease hazard rate on each time or age interval i is the product of a constant baseline hazard rate for subjects at the reference level of exposure in interval i and a function of the various exposures. The corresponding parameters, i.e., baseline hazard rates and hazard ratio parameters for exposure can be jointly estimated by maximizing the piecewise exponential likelihood, which is equivalent to the usual Poisson likelihood for the analysis of cohort data (Holford 1980; Laird and Oliver 1981). Corresponding variance estimates are available (Benichou and Gail 1990a). In simulations, risk estimates appeared subject to little bias, variance estimates were also little biased and coverage of confidence intervals was nearly nominal, except for the exposure profiles with very few subjects (Benichou and Gail 1990a). Other parametric approaches were considered to obtain risk estimates of cardiovascular events from the Framingham studies (Anderson et al. 1991). Semi-parametric estimators of risk were also derived (Benichou and Gail 1990a). In contrast with the previous approach where a piecewise exponential or Poisson distribution is assumed, the baseline disease hazard rate is expressed as an unspecified function of time or age rather than a constant, which corresponds to the semi-parametric Cox regression model (Cox 1972). Risk estimates are obtained as functions of the partial likelihood estimates (Cox 1975) of hazard ratio parameters and related Nelson-Aalen estimates of cumulative baseline hazards (Borgan 1998). From results in Tsiatis (1981) and Andersen and Gill (1982) on the joint distribution of these parameter estimates, Benichou and Gail (1990a) derived an asymptotic variance estimator.

Regression based methods appear well suited for estimating exposure-specific disease risk and are therefore useful for the purpose of individual prediction. Compared to the semi-parametric approach, the generalized density method appears easier to implement while providing a good compromise between bias and precision.

2.3.8 Estimation from Population-based or Nested Case-Control Studies

As discussed above, whereas disease risk is directly estimable from cohort data, case-control data have to be complemented with follow-up or population data in order to obtain the necessary information on incidence rates. If such complementary data are available, exposure-specific incidence rates and exposure-specific disease risk can be estimated. All approaches proposed in the literature rely on regression methods.

The Hybrid Approach

This approach relies on the assumption of piecewise constant incidence rates and on (2.2) to obtain baseline incidence rates in strata defined by factors such as age, sex, race or geographic area (see Sect. 2.2.2). Odds ratio estimates are then combined with baseline incidence rates to arrive at exposure-specific incidence rates (see Sect. 2.2.2). Applying (2.6) to these rates and death rates from competing causes, disease risk estimates can be obtained for desired time intervals. This approach has been used in practice to obtain individual risks of breast cancer by Gail et al. (1989) (see Example 1 below). Resulting disease risk estimates can be termed estimates of individual breast cancer risk since they depend on age and individual exposure profile (216 profiles were considered overall). The approach can be seen as a multivariate extension of earlier work by Miettinen (1974). It has been termed a hybrid approach (Benichou 2000a) since it relies on two models, namely the piecewise exponential model that underlies the density method (i.e., constant incidence by age group) and the logistic model used to obtain odds ratio estimates from the nested case-control data (see Sect. 2.4.6). It can be applied to population-based case-control data with no individual follow-up of subjects in a similar manner as to nested case-control data, as discussed and illustrated for bladder cancer by Benichou and Wacholder (1994) (see Example 2 below).

Variance estimators for risk estimates are complex since exposure-specific incidence rate estimates involve odds ratio parameters obtained through logistic regression from the case-control data and counts of incident cases from the follow-up or population data. Estimators of variances and covariances of age- and exposure-specific incidence rates that take into account all sources of variability have been fully worked out for various sampling schemes regarding control selection in the general case (Benichou and Gail 1990a) and specifically to account for the special features of the BCDDP data (Benichou and Gail 1995). Simulations tailored to the BCDDP data showed a small upward bias in risk estimates due to the small upward bias incurred by using odds ratios to estimate hazard ratios when the rare-disease assumption appeared questionable. Variance estimates had very little bias and yielded confidence intervals with near nominal coverage. Coverage was improved with the logit transformation.

Example 1. *(continued)*

Applying (2.6) to exposure-specific incidence rates of breast cancer estimated from the BCDDP data (see Sect. 2.2.2) and death rates from other causes estimated from US mortality data (see Sect. 2.3.6), risk estimates of breast cancer can be obtained. For instance, the 10-year risk of developing breast cancer between ages 40 and 50 years for a woman initially free of breast cancer at age 40 years and with the exposure profile considered in Sect. 2.2.2 (i.e., nulliparous woman with menarche at age 12 years, one previous biopsy for benign breast disease, and no history of breast cancer in her first-degree relatives) is obtained as a sum of two terms. The first term $\widehat{\pi}_1$, corresponding to age interval 40–44, is obtained from (2.6) as:

$$\begin{aligned}\widehat{\pi}_1 &= \frac{307.8\times10^{-5}}{307.8\times10^{-5}+153.0\times10^{-5}}\left[1-\exp\left\{-5\left(307.8\times10^{-5}+153.0\times10^{-5}\right)\right\}\right]\\ &= 0.0152\,.\end{aligned}$$

The second term $\widehat{\pi}_2$, corresponding to age interval 45–49, is obtained from (2.6) as the product of the probability of developing breast cancer in age interval 45–49 times the probability of having stayed free of breast cancer and not died from other causes in age interval 40–44:

$$\begin{aligned}\widehat{\pi}_2 = &\frac{424.3\times10^{-5}}{424.3\times10^{-5}+248.6\times10^{-5}}\left[1-\exp\left\{-5\left(424.3\times10^{-5}+248.6\times10^{-5}\right)\right\}\right]\\ &\times\exp\left\{-5\left(307.8\times10^{-5}+153.0\times10^{-5}\right)\right\} = 0.0204\,.\end{aligned}$$

Thus, the 10-year risk of developing breast cancer is obtained as the sum 0.0152 + 0.0204 = 0.0356, or 3.6%. The corresponding 95% confidence interval based on taking all sources of variability into account can be estimated as 3.0% to 4.2% through computations described in Benichou and Gail (1995). Breast cancer risk estimates can be obtained for all age intervals in the range 20–80 years and all 216 exposure profiles including the profile considered above. This whole approach to individual breast cancer risk estimation is known as the "Gail model" and has enjoyed widespread use in individual counseling, designing and interpreting prevention trials. Practical implementation has been greatly facilitated by the development of graphs (Benichou et al. 1996) as well as a computer program (Benichou 1993a) and its modified version that is available on the US National Cancer Institute web site at http://bcra.nci.nih.gov/brc/. ♦

Example 2 . In the year 1978, incident cases of bladder cancer were identified through 10 cancer registries in the United States. For instance, 32 incident cases were identified among white males aged 45–64 years whose population numbered 97,420 individuals. Assuming that this population remained constant throughout the year 1978, these data yielded an average incidence rate of 32.8 per 10^5 person-years. The National Bladder Cancer Study was a population-based case-control study conducted at the ten cancer registries. Incident cases aged 21–84 years were selected from the registries. Controls aged 21–84 years were selected from telephone sampling or Health Care Financing Administration rosters and frequency-matched to cases on geographic area, age and sex. Based on case-control data from two states (Utah and New Jersey) and one large city (Atlanta), odds ratios were estimated for smoking status (never smoker, ex-smoker, current light smoker, current heavy smoker) and occupational exposure to carcinogens (yes, no) using logistic regression (see Sect. 2.4.6). Moreover, the attributable risk for smoking and occupational exposure was estimated for white males in each of the nine strata resulting from the three areas and three age groups (i.e., 21–44, 45–64 and 65+ years) (see Sect. 2.5.1). Among white males aged 45–64 years in Utah, it was estimated at 54.0%, yielding a baseline incidence rate of $32.8 \times (1 - 0.540) = 15.1$ per 10^5 person-years. The odds ratios for current heavy smokers (≥ 20 cigarettes per day) and occupational exposure were estimated at 2.9 and 1.6. Hence, among white males aged 45–64 years in Utah, exposure-specific incidence rates were estimated at $15.1 \times 1.6 = 24.1$ per 10^5 person-years for never smokers with a history of occupational exposure, and $15.1 \times 2.9 \times 1.6 = 69.8$ per 10^5 person-years for current heavy smokers with a history of occupational exposure assuming a multiplicative effect of smoking and occupational exposure (and allowing for rounding error). From these exposure-specific incidence rates, estimates of the risk of bladder cancer over specified age intervals could be derived, using (2.6). ♦

Other Parametric Approaches

A pseudo-likelihood approach also relying on the assumption on piecewise constant incidence (i.e., piecewise exponential model) has been proposed as an alternative to the hybrid approach (Benichou and Wacholder 1994). In each stratum separately, observed distributions of exposure in the cases and controls are applied to counts of incident cases and person-time to obtain respective expected numbers of incident cases and of person-time per stratum and exposure level. Then, baseline incidence rates and hazard ratios are jointly estimated from these expected quantities under a piecewise exponential model. Joint estimation proceeds from maximizing the likelihood corresponding to this model. Since this likelihood includes expected rather than observed counts, it is termed a pseudo-likelihood. Thus, the procedure includes two steps. In the first step, expected numbers of incident cases and person-time per exposure and stratum are calculated. Then, the parameters of interest (i.e., stratum-specific baseline incidence rates and hazard ratios) are estimated from these expected counts through maximizing a pseudo-likelihood. This approach is easy to implement, as was illustrated on population-based case-control data of bladder cancer.

Example 2. *(continued)*
Among white males aged 45–64 years and in all other strata separately, observed proportions of cases (respectively controls) with given joint level of smoking and occupational exposure among the eight (four times two) joint levels considered were applied to counts of incident cases (respectively person-time) to obtain expected counts by stratum and joint exposure level. Namely, the products of the counts by the observed proportions were formed. Using these expected counts, a pseudo-likelihood based on the piecewise exponential model was maximized yielding estimates of relative hazards and stratum-specific baseline incidence rates. For instance, the baseline incidence rate for white males aged 45–64 years in Utah was estimated at 13.7 per 10^5 person-years and relative hazards for current heavy smoking and occupational exposure were estimated at 2.9 and 1.5, respectively. Hence, among white males aged 45–64 years in Utah, exposure-specific incidence rates were estimated at $13.7 \times 1.5 = 20.6$ per 10^5 person-years for never smokers with a history of occupational exposure, and $13.7 \times 2.9 \times 1.5 = 61.9$ per 10^5 person-years for current heavy smokers with a history of occupational exposure still assuming a multiplicative effect of smoking and occupational exposure (and allowing for rounding error). ♦

A full likelihood approach has also been proposed based on the piecewise exponential model (Benichou and Wacholder 1994). All parameters (i.e., baseline rates, hazard ratios and conditional probabilities for the distribution of exposure in the cases and controls) are estimated jointly through maximizing a likelihood involving all parameters. This approach may prove intractable in practice except in simple situations with few exposure levels considered. A full likelihood approach based on the logistic model (Greenland 1981) appears much easier to implement. Baseline incidence rates are obtained by simply adding to the stratum parameter estimates from the logistic model a term corresponding to the logarithm of the ratio of sampling fractions among cases and controls in the stratum (Greenland 1981; Prentice and Pyke 1979; also similar to discussion of (2.8) in Sect. 2.4.6).

Example 2. *(continued)*
Although it required the estimation of 60 additional parameters relative to the pseudo-likelihood approach, the full likelihood approach based on the piecewise exponential model could be implemented. The 60 additional parameters described the conditional probabilities of exposure (smoking and occupational exposure) in the cases and controls for all nine strata. For instance, the baseline incidence rate for white males aged 45–64 years in Utah was estimated at 13.9 per 10^5 person-years and relative hazards for current heavy smoking and occupational exposure were estimated at 2.9 and 1.6, respectively. Hence, among white

males aged 45–64 years in Utah, exposure-specific incidence rates were estimated at $13.9 \times 1.6 = 22.2$ per 10^5 person-years for never smokers with a history of occupational exposure, and $13.9 \times 2.9 \times 1.6 = 64.1$ per 10^5 person-years for current heavy smokers with a history of occupational exposure still assuming a multiplicative effect of smoking and occupational exposure (and allowing for rounding error). ♦

Upon comparing the pseudo-likelihood, full likelihood and hybrid approach on population-based case-control data of bladder cancer, Benichou and Wacholder (1994) noted that the hybrid approach seemed to be less efficient for incidence rate estimation than the other two approaches, which were themselves equally efficient. They discussed other advantages of the pseudo-likelihood and full likelihood approaches. Namely, these approaches allow direct estimation of hazard ratios rather than odds ratios. Furthermore, the pseudo-likelihood approach and the full likelihood approach (in its version relying on the piecewise exponential model) can be applied to more general regression models, e.g., models with an additive form using hazard rate difference parameters rather than hazard ratio parameters (see Sects. 2.4.4 and 2.4.6). Finally, all three approaches require that cases and controls be selected completely at random and that incident cases or at least a known proportion of them (i.e., known sampling fraction) be fully identified.

Semi-parametric Approach

In nested case-control studies, controls are usually individually matched to cases on time. Namely, for each case, one (or several) control(s) is (are) selected among subjects with the same age and length of follow-up in the cohort as the case (Breslow et al. 1983; Liddell et al 1977; Mantel 1973; see also Chap. I.7 of this handbook). The three parametric approaches described above do not apply readily to this context of individual time matching of controls to cases. Langholz and Borgan (1997) developed a semi-parametric approach to handle this case. Their approach can be regarded as an extension of the semi-parametric approach for cohort studies described above (see Sect. 2.3.7). Incidence rates are expressed as the product of baseline incidence rates of an unspecified form times a function of the covariates representing the hazard ratio (Cox 1972). Hazard ratio parameter estimates are obtained from maximizing the partial likelihood of the Cox model for nested case-control data (Oakes 1981; Prentice and Breslow 1978). Risk estimates are obtained by combining partial likelihood hazard ratio parameter estimates and corresponding cumulative hazard estimates.

A direct comparison of the semi-parametric approach with the parametric approaches presented above is not possible because the semi-parametric approach applies only to time-matched data, which the parametric approaches cannot handle. The semi-parametric approach requires observation of individual follow-up time of each subject in the original cohort in order to form the risk sets for each failure time and enable control selection. It is therefore potentially less widely applicable than the parametric approaches.

Final Notes and Additional References 2.3.9

General problems of definition of disease risk, interpretation and usefulness, properties, estimation and special problems have been reviewed in detail (Benichou 2000a). Special problems include accounting for continuous or time-dependent exposure, estimation of disease risk from two-stage case-control data, and validation procedures for disease risk estimates. Finally, an important challenge is to increase awareness of the proper interpretation and use of disease risk in practice and develop general software for easier implementation.

Measures of Association 2.4

Definitions and General Points 2.4.1

Measures of association have a long history and have been reviewed in many textbooks. They assess the strength of associations between one or several exposures and the risk of developing a given disease. Thus, they are useful in etiologic research to assess and quantify associations between potential risk (or protective) factors and disease risk. The question addressed is whether and to what degree a given exposure is associated with occurrence of the disease of interest. In fact, this is the primary question that most epidemiologic studies are trying to answer.

Depending on the available data, measures of association may be based on disease rates, disease risks, or even disease odds, i.e., $\pi/(1-\pi)$, with π denoting disease risk. They contrast rates, risks or odds for subjects with various levels of exposure, e.g., risks or rates of developing breast cancer for 40-year old women with or without a personal history of benign breast disease. They can be expressed in terms of ratios or differences of risks or rates among subjects exposed and non-exposed to given factors or among subjects with various levels of exposure.

Measures of association can be defined for categorical or continuous exposures. For categorical exposures, any two exposure levels can be contrasted using the measures of association defined below. However, it is convenient to define a reference level to which any exposure level can be contrasted. This choice is sometimes natural (e.g., non-smokers in assessing the association of smoking with disease occurrence) but can be more problematic if the exposure considered is of continuous nature, where a range of low exposures may be considered potentially inconsequential. The choice of a reference range is important for interpreting results. It should be wide enough for estimates of measures of association to be reasonably precise. However, it should not be so wide that it compromises meaningful interpretation of the results, which depend critically on the homogeneity of the reference level. For continuous exposures, measures of association can also be expressed per unit of exposure, e.g., for each additional gram of daily alcohol consumption. The reference level may then be a precise value such as no daily alcohol consumption or a range of values such as less than 10 grams of daily alcohol consumption.

2.4.2 Usefulness and Interpretation

When computing a measure of association, it is usually assumed that the relationship being captured has the potential to be causal, and efforts are taken to remove the impact of confounders from the quantity. Section 2.4.6 provides a summary of techniques for adjustment for confounders. Nonetheless, except for the special case of randomized studies, most investigators retain the word "association" rather than "effect" when describing the relationship between exposure and outcome to emphasize the possibility that unknown confounders may still influence the relationship.

Rothman and Greenland (Chap. I.4 of this handbook) take efforts to differentiate the concepts of effect and association, and adopt the framework of *counterfactuals*, popular in the field of economics (Wooldridge 2001), to define the term *effect size*. They then define "measure of association" as computed to compare two actual populations. Hence, the distinction is one of a true causal concept versus one that may be subject to the confounding of the true effect arising from the population mix of characteristics at hand. These definitions are more precise and serve as reminders of the true nature of causality. We will retain the less precise, but more common terminology where "measure of association" refers to either or both concepts. We also note that the discussion here is limited to measures of association with a *binary* (i.e. coded as 1 = present, 0 = absent) or event *count* (number of events) outcome. In many situations, classification into disease versus no disease is not clear-cut. For example, the definition of an abnormal lipid profile has undergone frequent change. In such cases, using measures based on continuous outcomes may be a better choice. We comment on relationships between measures of association for continuous and categorical outcomes in Sect. 2.4.6.

When choosing a measure of association, the primary goal is interpretability and familiarity to consumers of the information. Another guideline is that the measure of association should allow as simple a description of the association as possible. For example, it has been empirically observed that risk ratios are more likely than risk differences to remain constant across subpopulations with different risk levels (Breslow and Day 1980, Chap. 2), hence simplifying description of the association of the exposure with the outcome. Breslow and Day (1980, Chap. 2) also point out that ratios can be converted to differences by taking the logarithm of the risk or rate.

Definitions and properties of measures of association as well as relations among them are reviewed below for measures based on ratios and measures based on differences. Then, estimability of these measures from cohort and case-control designs and general points regarding estimation of these measures are considered, including an overview of techniques to adjust for confounders. More details regarding inference, namely estimating these measures and assessing the statistical significance of apparent associations, will be presented in Part II of this handbook.

The below Table 2.1 provides an overview of measures of association discussed in this chapter:

Table 2.1. Measures of association discussed in this chapter (GLM = generalized linear model; see Sect. 2.4.6)

Measure	Lower limit	Upper limit	Null value	Definition	Link function in GLM
Rate ratio (HR)	0	$+\infty$	1	$h_E/h_{\overline{E}}$	Log
Risk ratio (RR)	0	$+\infty$	1	$\pi_E/\pi_{\overline{E}}$	Log
Odds ratio (OR)	0	$+\infty$	1	$[\pi_E/(1-\pi_E)]/$ $[\pi_{\overline{E}}/(1-\pi_{\overline{E}})]$	Logit
Rate difference	$-\infty$	$+\infty$	0	$h_E - h_{\overline{E}}$	Identity
Risk difference	-1	$+1$	0	$\pi_E - \pi_{\overline{E}}$	Identity

Measures Based on Ratios 2.4.3

General Properties

Ratio based measures of association are particularly appropriate when the effect of the exposure is multiplicative, which means there is a similar percent increase or decrease associated with exposure in rate, risk or odds across exposure subgroups. As noted above, effects have often been observed to be multiplicative, leading to ratios providing a simple description of the association (e.g., see Breslow and Day 1980, Chap. 2). Ratio measures are dimensionless and range from zero to infinity, with one designating no association of the exposure with the outcome. When the outcome is death or disease, and the ratio has the rate, risk or odds of the outcome with the exposed group in the numerator, a value less than one indicates a protective effect of exposure. The exposure is then referred to as a protective factor. When the ratio in this set-up is greater than one, there is greater disease occurrence with exposure, and the exposure is then referred to as a risk factor.

It can be shown that numerically, the odds ratio falls the furthest from the null, and the risk ratio the closest, with the rate ratio in between. For example, from the below Table 2.2, based on a fictitious data from a cohort study for a disease that is not rare, we would obtain a risk ratio $\widehat{\text{RR}} = 0.3/0.1 = 3.00$ and an odds ratio $\widehat{\text{OR}} = [(30)(90)]/[(10)(70)] = 3.86$. If we assume a constant hazard, so that the risk for each group is $1 - \exp(-hT)$, with T being the follow-up time for each subject, we have the rate ratio $\widehat{\text{HR}} = \ln(1-0.3)/\ln(1-0.1) = 3.39$ (see Sects. 2.3.1 and 2.4.6). Hence $1 < \widehat{\text{RR}} < \widehat{\text{HR}} < \widehat{\text{OR}}$.

Table 2.2. Data from fictitious cohort study

	Exposed	Unexposed
Diseased	30	10
Non-diseased	70	90

The difference in magnitude between the above ratio measures is important to keep in mind when interpreting them for diseases or outcomes that are not rare. For rare outcomes the values of the three ratio measures tend to be close. Ratios become differences on the logarithmic scale, and estimation and inference often take place on the log scale, where zero indicates no association.

Rate Ratios

As the name implies, the rate ratio is the ratio between the rate of disease among those exposed and those not exposed or $h_E/h_{\overline{E}}$. Conceptually, the rate ratio is identical to a hazard ratio HR. The latter term tends to be used when time dependence of the rate is emphasized, as the hazard is a function that may depend on time. The situation of a constant rate ratio over time is referred to as *proportional hazards*. The proportional hazards assumption is often made in the analysis of rates (see below). Theoretically, the hazard ratio at a given time point is the limiting value of the rate ratio as the time interval around the point becomes very short, just as the hazard is the limiting quantity for incidence rate (see Sect. 2.2.1). The rate ratio has also been called the *Incidence Density Ratio* (Kleinbaum et al. 1982, Chap. 8). It may be noted that the rate ratio is attenuated by less than perfect specificity of the outcome criteria, but relatively unaffected by less than perfect sensitivity, especially when the rate is low, as long as the sensitivity is unaffected by exposure. In other words, if cases are equally missed in the exposed and unexposed groups, the rate ratio is relatively unaffected. However, if non-cases are considered cases, the ratio will be lower than if diagnostic criteria identified only true cases. Even in the fictitious example above with high incidence rates, 80% sensitivity leads to a slightly attenuated rate ratio of 3.29 from

$$\widehat{\mathrm{HR}} = \ln\left[1 - (0.80)(0.3)\right] / \ln[1 - (0.80)(0.1)] = 3.29$$

(as compared to the correct rate ratio of 3.39 from Table 2.2), while 80% specificity leads to a severely biased rate ratio of

$$\widehat{\mathrm{HR}} = \ln[0.80(1 - 0.3)]/\ln[0.80(1 - 0.1)] = 1.77\ .$$

Rate ratios are extremely useful because of the ease of estimating them in many contexts. They refer to population dynamics, and are not as easily interpretable on the individual level. It has been argued, however, that rate ratios make more sense than risk ratios (see below) when the period subjects are at risk is longer than the observation period (Kleinbaum et al. 1982, Chap. 8). Numerically, the rate ratio is further from the null than the risk ratio. When rates are low, the similarity of risk and rate leads to rate ratios being close to risk ratios, as discussed below. Some investigators tend to refer to rate ratios as relative risks, creating some confusion in terminology. Further considerations of how the rate ratio relates to other ratio based measures of association are offered by Rothman and Greenland (1998, p 50).

Risk Ratios

The risk ratio, relative risk or ratio of risks of disease among those exposed π_E and those not exposed $\pi_{\overline{E}}$, RR = $\pi_E/\pi_{\overline{E}}$, has been viewed as the gold standard among measures of association for many years. It is eminently interpretable on the individual level as a given-fold increase in risk of disease. Like other ratio-based measures, it tends to be more stable than the risk difference across population groups at widely different risk. However, similar to rate ratios and odds ratios (introduced in Sect. 2.4.3), the risk ratio can be viewed as misleading in the public eye when the risk among both the unexposed and the exposed is very low, yet many-fold increased by exposure. Another disadvantage of the risk ratio is its asymmetry with respect to the definition of an event, so that the risk ratio for not having an event, $(1-\pi_E)/(1-\pi_{\overline{E}})$, cannot be directly computed from the risk ratio for having an event. For example, knowing that the risk ratio for an event RR = 3.00, the scenario $\pi_E = 0.3$, $\pi_{\overline{E}} = 0.1$ results in $(1-\pi_E)/(1-\pi_{\overline{E}}) = 0.7/0.9 = 0.78$, while the scenario $\pi_E = 0.6$, $\pi_{\overline{E}} = 0.2$, which represents the same risk ratio of 3.00, results in $(1-\pi_E)/(1-\pi_{\overline{E}}) = 0.4/0.8 = 0.50$. The risk ratio depends on the length of the time interval considered because risk itself refers to a specific interval (see Sect. 2.3.1). In the literature, the term relative risk is often used to denote the rate ratio as well as the risk ratio, creating some confusion. Therefore, we will avoid the term "relative risk" in the following. Numerically the risk ratio is closer to the null than the rate ratio for the same data (see above).

Cornfield et al. (1959), in the smoking versus lung cancer debate, derived several theoretical properties of the risk ratio, which have further supported its use. In this debate, Cornfield, along with Doll and Hill, argued against strong opposition from R.A. Fisher and Joseph Berkson that the association was causal, and not likely due to unmeasured confounders, such as a genetic predisposition to both smoke and contract lung cancer. First of all, Cornfield et al. (1959) turned attenuation of the risk ratio due to lack of specificity of the outcome into an advantage, by noting that the ratio will become stronger as the disease subtype affected by the exposure is honed. Second, Cornfield et al. demonstrated that if a confounder is to explain the outcome with exposure risk ratio RR > 1, that confounder has to have risk ratio at least RR, and in addition the prevalence of the confounder must be at least RR times greater among the exposed than among the unexposed. Lin et al. (1998) presented more general formulas that confirm Cornfield et al.'s assertions under assumptions of no interaction between the confounder and exposure. These theoretical results have led investigators to reason that high risk ratios (say above 1.4; Siemiatycki et al. 1988) are not likely to be explained by uncontrolled confounding.

Odds Ratios

For several reasons, the odds ratio has emerged as the most popular measure of association. The odds ratio is the ratio of odds, OR = $[\pi_E/(1-\pi_E)]/[\pi_{\overline{E}}/(1-\pi_{\overline{E}})]$. Historically, the odds ratio was considered an approximation to the risk ratio obtainable from case-control studies. The reason for this is that the probabilities of being sampled into case and control groups cancel in the calculation of the odds ratio, as long as sampling is independent of exposure status. Furthermore, when π_E and $\pi_{\overline{E}}$ are small, the ratio $(1-\pi_{\overline{E}})/(1-\pi_E)$ has little influence on the odds ratio, making it approximately equal to the risk ratio $\pi_E/\pi_{\overline{E}}$. The assumption of small π_E and $\pi_{\overline{E}}$ is referred to as the *rare-disease assumption.* Kleinbaum et al. (1982) have pointed out that in a case-control study of a stable population with incident cases and controls being representative of non-cases, the odds ratio is the rate ratio. Numerically, the odds ratio is the furthest from the null of the three ratio measures considered here.

More recently, the odds ratio has gained status as an association measure in its own right, and is often applied in cohort studies and clinical trials, as well as in case-control studies. This is due to many desirable properties of the odds ratio. First of all, focusing on risk rather than odds may be a matter of convention rather than a preference based on fundamental principles, and using the same measure across settings has the advantage of consistency and makes comparisons and meta-analyses easy. In contrast to the risk ratio, the odds ratio is symmetric so that the odds ratio for disease is the inverse of the odds ratio for no disease. Furthermore, the odds ratio based on exposure probabilities equals the odds ratio based on disease probabilities, a fact that follows from Bayes' theorem (e.g., Cornfield 1951; Miettinen 1974; Neutra and Drolette 1978) or directly from consideration of how cases and controls are sampled. The disease and exposure odds ratios are sometimes referred to as *prospective* and *retrospective odds ratios*, respectively. Finally, odds ratios from both case-control and cohort studies are estimable by logistic regression, which has become the most popular approach to regression analysis with binary outcomes (see Sect. 2.4.6).

Some investigators feel that the risk ratio is more directly interpretable than the odds ratio, and have developed methods for converting odds ratios into risk ratios for situations when risks are not low (Zhang and Yu 1998).

2.4.4 Measures Based on Differences

General Properties

Difference based measures are appropriate when effects are additive (e.g., see Breslow and Day 1980, Chap. 2), which means that the exposure leads to a similar absolute increase or decrease in rate or risk across subgroups. The difference in odds is very rarely used, and not addressed here. As noted above, additive relationships are less common in practice, except on the logarithmic scale, when they are equivalent to ratio measures. However, difference measures may be more understandable to the public when the outcome is rare, and relate directly to measures of impact discussed below (see Sect. 2.5).

The numerical ranges of difference measures depend on their component parts. The rate difference ranges from minus to plus infinity, while the risk difference is bounded between minus and plus one. The situation of no association is reflected by a difference measure of zero. When the measure is formed as the rate or risk among the exposed minus that among the non-exposed, a positive value indicates that the exposure is a risk factor, while a negative value indicates that it is a protective factor. It can be shown that the risk difference falls numerically nearer to the null than does the rate difference. For example, Table 2.2 yields a risk difference of $0.30 - 0.10 = 0.20$, while the rate difference is $\ln(0.70) + \ln(0.90) = 0.25$. However, they will be close for rare outcomes. In contrast to ratio measures, difference measures are always attenuated by less than perfect sensitivity (i.e., missed cases), but the rate difference is unaffected by less than perfect specificity. The risk difference is also relatively unaffected when risk is low. In the fictitious example above, if the sensitivity of the test used to detect disease is 80%, the rate difference is $-\ln[1-(0.80)(0.3)] + \ln[1-(0.80)(0.1)] = 0.19$, but if the specificity is 80%, the rate difference remains at 0.25.

Rate Differences

The rate difference is defined as $h_E - h_{\overline{E}}$, and has been commonly employed to compare mortality rates and other demographic rates between countries, time periods and/or regions. In such comparisons, the two rates being compared are often *directly standardized* (see Sect. 2.6) to the age and sex distribution of a standard population chosen, e.g., as the population of a given country in a given census year.

For the special case of a dichotomous exposure, the rate difference, i.e., the difference between the incidence rates in the exposed and unexposed subjects has been termed “excess incidence” (Berkson 1958; MacMahon and Pugh 1970; Mausner and Bahn 1974), “excess risk” (Schlesselman 1982), “Berkson’s simple difference” (Walter 1976), “incidence density difference” (Miettinen 1976), or even “attributable risk” (Markush 1977; Schlesselman 1982), which may have caused some confusion.

Risk Differences

The risk difference $\pi_E - \pi_{\overline{E}}$ is parallel to the rate difference discussed above, and similar considerations apply. Due to the upper and lower limits of plus, minus one on risk, but not on rate, risk differences are more difficult to model than rate differences.

Estimability 2.4.5

Because exposure-specific incidence rates and risks can be obtained from cohort data, all measures of association considered (based on ratios or differences) can be obtained as well. This is also true of case-control data complemented by follow-up or population data (see Sects. 2.2 and 2.3). Case-control data alone allow estimation of odds ratios thanks to the identity between disease and exposure odds ratios (see Sect. 2.4.3) that extends to the logistic regression framework. Prentice and Pyke (1979) showed that the unconditional logistic model (see also Breslow and Day 1980, Chap. 6) applies to case-control data as long as the intercept is disregarded (see Sect. 2.4.6). Interestingly, time-matched case-control studies allow estimation of hazard rates (e.g., see Miettinen 1976; Greenland and Thomas 1982; Prentice and Breslow 1978).

2.4.6 Estimation

The most popular measures of association have a long history of methods for estimation and statistical inference. Some traditional approaches have the advantage of being applicable in small samples. Traditional methods adjust for confounders by *direct standardization* (see Sect. 2.6.1) of the rates or risks involved, prior to computation of the measure of association, or by *stratification*, where association measures are computed separately for subgroups and then combined. For measures based on the difference of rates or risks, direct standardization and stratification can be identical, if the same weights are chosen (Kahn and Sempos 1989). Generally, however, direct standardization uses predetermined weights chosen for external validity, while *optimal* or *efficient* weights are chosen with stratification. Efficient weights make the standard error of the combined estimator as small as possible. *Regression adjustment* is a form of stratification, which provides more flexibility, but most often relies on large sample size for inference.

In modern epidemiology, measures of association are most often estimated from regression analysis. Such methods tend to require large sample sizes, in particular when based on *generalized linear models* (often abbreviated *GLM*). In this context, the ratio, difference or other association measures arise from the regression coefficient of the exposure indicator, and different measures of association result depending on the transformation applied to the mean of the outcome variable. Note that the mean of an event count over a unit time interval is the rate, and the mean of a binary outcome is the risk. For example a model may use the logarithm of the rate ($\ln(h)$) or risk ($\ln(\pi)$) as the outcome to be able to estimate ratio measures of association.

The function applied to the rate or risk in a regression analysis is referred to as the *link function* in the framework of generalized linear models underlying such analyses (see McCullagh and Nelder (1989) and Palta (2003) for theory and practical application). For example, linear regression would regress the risk or rate directly on exposure without any transformation, which is referred to as using the *identity link*. When the exposure is the only predictor in such a model, all link functions fit equally well and simply represent different ways to characterize the association. However, when several exposures or confounders are involved, or if the exposure is measured as a continuous or ordinal variable, some link functions and not others may require interaction or non-linear terms to improve the fit. The considerations in choosing the link function parallel those for choosing a measure of association as multiplicative or additive and as computed from rates, risks or odds, discussed above (see Table 2.1).

Both traditional and regression estimation is briefly overviewed below, with more details provided in Chap. II.3 and Chap. II.4 of this handbook.

Estimation and Adjustment for Confounding of Rate Ratios

Estimation of the rate or hazard ratio between exposed and non-exposed individuals can be based on either event counts (overall or in subgroups and/or subintervals of time), or on the time to event for each individual, where the time for subjects without events are entered as time to end of follow-up, and are referred to as being *censored* (see Chap. II.4).

In the first case, estimation can proceed directly by forming ratios of interest, or by modeling the number of events on exposure by a generalized linear model. When ratios are formed directly as the ratio of the number of cases D_E divided by the person time at risk t_E, i.e. D_E/t_E, in those exposed and $D_{\overline{E}}/t_{\overline{E}}$ in those unexposed, the 95% confidence interval of the resulting rate ratio HR $= D_E/t_E/D_{\overline{E}}/t_{\overline{E}}$ is obtained as (Rothman and Greenland 1996)

$$\left[\exp\left(\ln(\widehat{\mathrm{HR}}) - 1.96(1/D_E + 1/D_{\overline{E}})^{1/2}\right), \exp\left(\ln(\widehat{\mathrm{HR}}) + 1.96(1/D_E + 1/D_{\overline{E}})^{1/2}\right)\right] .$$

In either case, it is often necessary to adjust for confounding factors, including age and sex. When rate ratios are formed directly, the rates are generally adjusted by direct standardization (see Sect. 2.6.1) or by use of the *standardized mortality (or morbidity or incidence) ratio SMR or SIR* (see Sect. 2.6.1). The SMR and SIR have found wide application in investigations of the potential health effects of occupational exposures.

A common regression approach to estimating rate ratios requires information on event count and person time at risk for each subgroup, time interval and exposure level of interest. To obtain rate ratios from the regression requires that the logarithm of the mean number of events be modeled. This is referred to in the generalized linear model framework as using a *log link* function. The resulting regression equation is

$$\ln(h_i) = -\ln(t_i) + \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots ,$$

where the subscript i indicates subject, $i = 1, \dots, n$, E_i is an indicator that equals 0 for the unexposed and 1 for the exposed. In this equation, β_0 is the logarithm of the rate per time unit for the unexposed with confounder values, $X_1, X_2, \dots = 0$. Care should be taken to center confounders so that this intercept is meaningful. The quantity $\ln(t_i)$ is referred to as the offset, and allows event counts over different size denominators to be used as the outcome variable. In the case when disease rates in a population are modeled, t_i are population sizes. The rate ratio for exposure adjusted for confounders $X_1, X_2, \dots$ is obtained as $\exp(\beta_E)$. Differences in rate ratios across levels of X can easily be accommodated by the inclusion of interaction terms in the model. Inferences on the rate ratio follow from the standard error of the estimate $\widehat{\beta}_E$ of β_E, which is approximately normally distributed with reasonable large sample sizes, so that a 95% confidence interval for the rate ratio is

$$\left[\exp\left(\widehat{\beta}_E - 1.96 se\left(\widehat{\beta}_E\right)\right), \exp\left(\widehat{\beta}_E + 1.96 se\left(\widehat{\beta}_E\right)\right)\right] .$$

The standard errors $se(\widehat{\beta}_E)$ can be obtained from maximum likelihood theory, assuming that the counts follow a *Poisson* or *negative binomial distribution.* The variance of the Poisson distribution equals the rate, while the negative binomial distribution allows for possible clustering of events leading to the variance being larger than the rate.

There are also several approaches available in most statistical software packages to adjust standard errors for so called *overdispersion*. Overdispersion refers to variability in rates being larger than expected from a Poisson count process. For example, events may cluster in time, or there may be unmeasured characteristics of the population influencing the rate, so that the overall count arises from a mixture of different rates. An example of overdispersion (Palta 2003) arises in overall cancer rates because different cancers predominate for different ages and genders. One of the approaches to adjusting for overdispersion, is to use a *robust* or *sandwich* estimator of the standard error of $\widehat{\beta}_E$ available in software packages, such as PROC GENMOD in SAS (1999) that fit *generalized estimating equations* (Liang and Zeger 1986).

When the data consist of times to event for individuals, the rate ratio, or hazard ratio can be estimated by techniques designed for *survival analysis* (e.g., see Hosmer and Lemeshow 1999 and Chap. II.4 of this handbook). Most parametrically specified survival distributions (i.e., distributions $S(t) = 1 - F(t)$, where F is the distribution of time to event) lead to hazard ratios $h_E(t)/h_{\overline{E}}(t)$ that vary over time. When the hazard ratio remains constant, this is referred to as proportional hazards. This property holds when the time to event follows the *exponential distribution*, so that the probability of avoiding an event up to time t is given by $S(t) = \exp(-ht)$ where h is a constant hazard, and for the *Weibull distribution* $S(t) = \exp[(-ht)^{\gamma}]$ as long as γ is the same for the exposed and non-exposed groups. Models are sometimes fit that assume that the exponential distribution holds over short intervals, i.e., piecewise constant hazard. In these models, the hazard ratio is constant across short intervals, but can be allowed to change over time. An exponential distribution for time to event leads to the Poisson distribution for number of events in a given time period.

In the situation of proportional hazards, estimation of the hazard ratio can proceed without specifying the actual survival distribution via the *Cox model*, where estimation is based on so called *partial likelihood* (Cox 1972). The reason this works is that the actual level of the hazard cancels out; similarly to how the offset becomes part of the intercept in the regression model given by (2.8) above.

Estimation and Adjustment for Confounding for Risk Ratios

In a cohort study with a fixed follow-up time, the risk ratio can be estimated in a straightforward manner. From a 2 × 2 table (see Table 2.3) with cells a, b, c, d, where a is the number diseased and exposed, b is the number diseased and unexposed, c the number non-diseased and exposed and d the number non-diseased and unexposed,the risk ratio is estimated by $\widehat{\mathrm{RR}} = \{a/(a + c)\}/\{b/(b + d)\}$.

Table 2.3. Notation for a generic 2 × 2 table from cohort or case-control study

	Exposed	Unexposed
Diseased	a	b
Non-diseased	c	d

Statistical inference can be based on the approximate standard error (Katz et al. 1978) which can be estimated as $se(\ln(\widehat{\mathrm{RR}})) = \{a/(a+c)+d/b(b+d)\}^{1/2}$. In cases where follow-up time is not fixed, the risk ratio can be calculated from the individual risks estimated from the rate or hazard function. However, this is rarely done, as investigators tend to prefer the rate or hazard ratio as the measure of association in such situations. The risk ratio can be estimated from case-control studies only when the ratio of sampling probabilities of cases and controls is known, or by using the odds ratio (see above) as an approximation.

Although standardization can be used either as direct standardization to adjust risks before forming ratios or as indirect standardization to compute the SMR (see Sect. 2.6.1) from risks in a reference population, it is often more appropriate to apply stratified analyses to adjust the risk ratio for confounders (see also Sect. 2.6.1). For example, a study of cancer risk in individuals exposed or not exposed to a risk factor may be stratified into age groups, or a study investigating outcomes in neonates may be stratified by birth weight. Stratum-specific risk ratio estimates can be calculated and then be combined for instance by the popular Mantel–Haenszel estimator that is known to have good properties. It is given by

$$\widehat{\mathrm{RR}}_{\mathrm{MH}} = \sum \left(a_i \left(b_i + d_i\right)/n_i\right) \Big/ \sum \left(b_i \left(a_i + c_i\right)/n_i\right) \ ,$$

where the sums are across strata and n_i is the number of subjects in stratum i. This estimator is stable in small samples, but has a larger standard error than the corresponding estimator from regression modeling. Formulas for the standard error are provided by Breslow and Day (1987) and by Rothman and Greenland (1998).

From regression analysis, the risk ratio can be obtained as $\exp(\beta_E)$ from fitting the binary or *binomial* (grouped binary events) outcome to the model:

$$\ln(\pi_i) = \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \dots \ .$$

This is a generalized linear model with error distribution reflecting each binary outcome being independent with variance $\pi_i(1 - \pi_i)$ and log link. Clearly, the log link is not ideal, as $\pi_i > 1$ can result from some exposure-confounder combinations. Nonetheless, this model tends to be reasonable with low risks. Maximum likelihood or generalized estimating equation fitting automatically provides large sample inference, with or without adjustment for deviations from the binomial error structure by robust standard errors. Deviations from binomial structure may result from clustering or correlation between events within subgroups, or from multiple events per person (e.g., cavities in teeth when teeth are individually counted).

Another option for the link function when modeling the risk by a generalized linear model is the so-called *complementary log-log link* resulting in the model:

$$\ln(-\ln(1 - \pi_i)) = \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \dots \ .$$

This model has the advantage of always estimating risks to be in the range 0 to 1. However, $\exp(\beta_E)$ is the rate ratio rather than the risk ratio.

Estimation and Adjustment for Confounding for Odds Ratios

In the traditional setting, the odds ratio in an unmatched case-control or cohort study is estimated from a 2 × 2 table (see Table 2.3) as $\widehat{\text{OR}} = ad/bc$. Inference can be based on exact methods, which historically were difficult to implement, but are now available in most statistical software packages, such as the SAS procedure PROC FREQ. With the exact approach, the confidence interval for the odds ratio is obtained from the *non-central hypergeometric distribution*. Over the years, many approximations to this interval have been developed, the most accurate of which is the *Cornfield approximation* (Cornfield 1956). Another, less accurate method is based on the approximate standard error of $\ln(\widehat{\text{OR}})$ known as the Woolf (1955) or *logit method*, where $se(\ln(\widehat{\text{OR}}))$ is calculated as $(1/a + 1/b + 1/c + 1/d)^{1/2}$. The logit method takes its name from being related to an approximation used for fitting logistic regression. Although the approximation has limited use for reporting final study results, it is useful to have an explicit approximation of the standard error for study planning purposes.

Stratified methods for estimating the odds ratio either build on taking a weighted average of the stratum specific log odds ratios, using the inverses of the logit method standard errors for each stratum as the weights, or using the Mantel–Haenszel stratified odds ratio estimator (Mantel and Haenszel 1959),

$$\widehat{\text{OR}}_{\text{MH}} = \left(\sum a_i c_i/n_i\right) / \left(\sum b_i d_i/n_i\right) ,$$

where the sums are across strata with tables as depicted in Table 2.3 for each stratum and n_i is the number of subjects in stratum i. This odds ratio estimator has been shown to have excellent properties even when strata are very small (Birch 1964; Breslow 1981; Breslow and Day 1980, Chaps. 4–5; Landis et al. 1978, 2000; Greenland 1987; Robins and Greenland 1989). The confidence interval for a stratified odds ratio can be obtained by exact methods or by the approximation of Miettinen (1976) where $se(\ln(\widehat{\text{OR}}))$ is calculated as $\ln(\widehat{\text{OR}}_{\text{MH}})/\chi_{\text{MH}}$. Here χ_{MH} is the square root of the *Mantel–Haenszel stratified chi-square test* used to test the null hypothesis that the odds ratio equals one (Mantel and Haenszel 1959). This test statistic is computed as

$$\chi^2_{\text{MH}} = \sum \left[a_i - (a_i + b_i)(a_i + c_i)/n_i\right]^2 / \\ \left[(a_i + b_i)(a_i + c_i)(d_i + b_i)(d_i + c_i)/\left(n_i^2(n_i - 1)\right)\right] .$$

The 95% confidence interval for the odds ratio is then given by

$$\left[\exp\left(\ln\left(\widehat{\text{OR}}\right) - 1.96\left(\ln\left(\widehat{\text{OR}}_{\text{MH}}\right)/\chi_{\text{MH}}\right)\right), \exp\left(\ln\left(\widehat{\text{OR}}\right) + 1.96\left(\ln\left(\widehat{\text{OR}}_{\text{MH}}\right)/\chi_{\text{MH}}\right)\right)\right] .$$

A special case of stratification occurs when data are pair matched, such that each case is matched to a control, e.g., based on kinship or neighborhood. In this case, the Mantel–Haenszel odds ratio estimator becomes m_{+-}/m_{-+}, where m_{+-} is the number of pairs (matched sets) where both the case and the control are exposed, and m_{-+} is the number of pairs where neither is exposed. Breslow and Day (1980, Chap. 7) provide additional formulas for the situation when several controls are matched to each case. Confidence intervals can again be obtained by exact formulas (Breslow and Day 1980, Chap. 7). It is well known that although matched studies are not technically confounded by the factors matched on because cases and controls are balanced on these, odds ratios based on the matched formula are larger than odds ratios not taking the matching into account. We discuss this phenomenon further in the logistic model framework below.

Increasingly, logistic regression is used for the estimation of odds ratios from clinical trials, cohort and case control studies. Logistic regression fits the equation:

$$\ln\left(\pi_i/(1-\pi_\iota)\right) = \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \ldots\,, \tag{2.8}$$

with $\boldsymbol{E}_i$ denoting the exposure status and $\boldsymbol{X}_{1i}$, $\boldsymbol{X}_{2i}, \ldots$ the confounder variables of individual i. For a cohort study β_0 is $\ln(\pi_i/(1-\pi_\iota))$ for an unexposed individual with all *confounders equal* to 0. For such a person, then, the risk of disease $\pi_i = \exp(\beta_0)/[1+\exp(\beta_0)]$. In a case-control study, the intercept in (2.11) is $\beta_0 = \beta_{0,\text{cohort}} + \ln(P_1/P_0)$, with P_1 and P_0 the probabilities for being sampled into the study for cases and controls, respectively. We see again that risk can be estimated from a case-control study only when the sampling scheme of cases and controls is known. The odds ratio for exposure, adjusted for confounders is $\exp(\beta_E)$.

In the generalized linear model framework, (2.8) is said to use the *logit link*, where the logit function is defined as $g(\pi) = \ln(\pi/(1-\pi))$. The logit link is the one that follows most naturally from the mathematical formulation of the binomial distribution (McCullagh and Nelder 1989), and is referred to as the *canonical link*, whereas the log is the canonical link for rates. Just as for other generalized linear models, maximum likelihood based and robust standard errors are available, with the latter taking into account clustering of events. It should be noted, however, that generalizations of logistic regression to the longitudinal or clustered setting by generalized estimating equations do not work for case-control studies (Neuhaus and Jewell 1990).

Matched data can be analyzed by *conditional logistic regression* that fits the model:

$$\ln(\pi_{ji}/(1-\pi_{ji})) = \beta_{0j} + \beta_E E_{ji} + \beta_1 X_{1ji} + \ldots \tag{2.9}$$

for individual i in the matched set j. Estimation of β_E and $\beta_1, \ldots$ is based on algorithms that compare individuals only within and not between matched sets. For example, for matched pairs, estimation is based on differences in exposure and confounders. These algorithms do not actually estimate the matched set specific intercepts β_{0j} that cancel out. All variables that do not vary within matched sets are automatically absorbed into β_{0j} although interactions of such variables with those that vary within set can be included in the model. For example, in SAS, PROC PHREG can be tricked into fitting this model (e.g., see Palta 2003). While the conditional logistic regression model is usually fit by large sample methods, such as maximum likelihood, exact procedures have also become available (e.g., Mehta et al. 2000). Again, taking matching into account in the analysis results in larger coefficients than those of the unmatched model (2.11). When all matching variables are explicit (such as age and sex) they can be directly entered as covariates in (2.11).

It is useful to know that, when an outcome is originally normally distributed, but dichotomized and analyzed by logistic regression, the resulting coefficients in the unconditional model (2.11) are approximately 1.7 times as large as the coefficients that would have resulted from ordinary regression of the original continuous outcome, "standardized" by being divided by its residual standard deviation. (Note that the word *standardized* here is used to denote a conversion to standard deviation units, rather than in the sense of direct standardization discussed in Sect. 2.6.1.) This result emerges from the relationship between the variances of the logistic and normal distributions (Johnson and Kotz 1970). While the logit link is related to the logistic distribution, another link function, the *probit* can be shown to arise directly when a continuous outcome from ordinary regression with normally distributed errors is dichotomized (Palta 2003). The probit link is defined as $g(\pi) = \Phi^{-1}(\pi)$ where Φ^{-1} is the inverse of the cumulative normal distribution. This link yields the same coefficients as the "standardized" ones from ordinary regression of the continuous outcome. Apart from this difference, the logit and probit provide a very similar fit. In both cases, of course, dichotomizing the outcome results in loss of information and thus in loss of statistical efficiency, which yields larger standard errors relative to the size of the regression coefficients.

The idea of logistic regression providing coefficients that are related by "standardization" to those that would arise from regression analysis of an underlying continuous variable (e.g., blood pressure being dichotomized into hypertension or not) also provides a framework for understanding the difference between a matched and an unmatched analysis. In an unmatched analysis, the coefficients are for the outcome "standardized" to the scale of the overall residual standard deviation across the population. This means that the original continuous regression coefficient is divided by that standard deviation. In a matched analysis, the coefficients are "standardized" to the residual standard deviation within each matched set. This happens by explicitly including a matched set specific intercept in the model (see (2.12)). Hence, the standard deviation within matched sets does not contain the variation arising from different matched sets having a different level of the outcome, and hence matched coefficients are larger (Palta et al. 1997; Palta and Lin 1999).

Estimation and Adjustment for Confounding for Rate Differences

Regression estimation of the rate difference with and without adjustment for confounders can be done in the generalized linear model framework by specifying the identity link function, resulting in linear regression of the rates with variance arising from the Poisson distribution. Overdispersion can be handled the same way as for ratios. However, unequal time intervals cannot be as easily accommodated with the identity link. Instead, weighted ordinary regression of observed rates can be employed, where inverse variance weights automatically account for the interval length (Breslow and Day 1987, Chap. 4).

2.5 Measures of Impact

Measures of impact are used to assess the contribution of one or several exposures to the occurrence of incident cases at the population level. Thus, they are useful in public health to weigh the impact of exposure on the burden of disease occurrence and assess potential prevention programs aimed at reducing or eliminating exposure in the population. They are sometimes referred to as measures of *potential* impact to convey the notion that the true impact at the population level may be different from that reflected by these measures except under very specific conditions (see Sect. 2.5.1). The most commonly used measure of impact is the attributable risk. This measure is presented in some detail below. Then, other measures are briefly described. Table 2.4 provides an overview of measures of impact discussed in this chapter.

2.5.1 Attributable Risk

Definition

The term "attributable risk" (AR) was initially introduced by Levin in 1953 (Levin 1953) as a measure to quantify the impact of smoking on lung cancer occurrence. Gradually, it has become a widely used measure to assess the consequences of an association between an exposure and a disease at the population level. It is defined as the following ratio:

$$\mathrm{AR} = \left\{\Pr(D) - \Pr\left(D|\overline{E}\right)\right\} / \Pr(D) \; . \qquad (2.10)$$

The numerator contrasts the probability of disease, $\Pr(D)$, in the population, which may have some exposed, E, and some unexposed, $\overline{E}$, individuals, with the hypothetical probability of disease in the same population but with all exposure eliminated $\Pr(D|\overline{E})$. Thus, it measures the additional probability of disease in the population that is associated with the presence of an exposure in the population, and AR measures the corresponding proportion. Probabilities in (2.10) will usually refer to disease risk although, depending on the context, they may be replaced with incidence rates.

Unlike measures of association (see Sect. 2.4), AR depends both on the strength of the association between exposure and disease and the prevalence of exposure in the population p_E. This can be seen for instance through rewriting AR from (2.10). Upon expressing $\Pr(D)$ as

$$\Pr(D|E)p_E + \Pr\left(D|\overline{E}\right) p_{\overline{E}} \quad \text{with} \quad p_{\overline{E}} = 1 - p_E \; ,$$

both in the numerator and the denominator, and noting that

$$\Pr(D|E) = \mathrm{RR} \times \Pr\left(D|\overline{E}\right) \; ,$$

the term $\Pr(D|\overline{E})$ cancels out and AR is obtained as (Cole and MacMahon 1971; Miettinen 1974):

$$\mathrm{AR} = \left\{p_E(\mathrm{RR} - 1)\right\} / \left\{1 + p_E(\mathrm{RR} - 1)\right\} \; , \qquad (2.11)$$

Table 2.4. Measures of impact discussed in this chapter

Measures	Range	Definition[a]	Usual interpretation(s)[b]
Attributable risk (AR)	$-\infty$ to 1 0 to 1 for risk factor	1) $\{\Pr(D) - \Pr(D\|\overline{E})\}/\Pr(D)$ 2) $\{p_E(\mathrm{RR}-1)\}/\{1 + p_E(\mathrm{RR}-1)\}$ 3) $p_{E\|D}(\mathrm{RR}-1)/\mathrm{RR}$	Proportion of disease cases in the population attributable to exposure. Proportion of disease cases in the population potentially preventable by eliminating exposure
Attributable risk among the exposed (AR_E)	$-\infty$ to 1 0 to 1 for risk factor	1) $\{\Pr(D\|E) - \Pr(D\|\overline{E})\}/\Pr(D\|E)$ 2) $(\mathrm{RR}-1)/\mathrm{RR}$	Proportion of disease cases among the exposed attributable to exposure
Sequential attributable risk	0 to 1 for risk factor	Contributions of a given exposure to the joint attributable risk to several exposures for a given order of exposures	Proportion of disease cases in the population potentially preventable by eliminating a given exposure when several exposures are removed in a given sequence
Partial attributable risk	0 to 1 for risk factor	Average contribution of a given exposure to the joint attributable risk to several exposures over all possible exposure orders	Average proportion of disease cases in the population potentially preventable by eliminating a given exposure when several exposures are removed in sequence over all possible orders of removal
Prevented (preventable) fraction (PF)	$-\infty$ to 1 0 to 1 for protective factor	1) $\{\Pr(D\|\overline{E}) - \Pr(D)\}/\Pr(D\|\overline{E})$ 2) $p_E(1-\mathrm{RR})$	Proportion of disease cases averted (“prevented fraction”) in relation to the presence of a protective exposure or intervention in the population. Proportion of cases that could be potentially averted (“preventable fraction”) if a protective exposure or intervention were introduced *de novo* in the population

Measures	Range	Definition[a]	Usual interpretation(s)[b]
Generalized impact fraction	$-\infty$ to 1 0 to 1 for risk factor and modified distribution with lowering of exposure	$\{\Pr(D) - \Pr^*(D)\}/\Pr(D)$	Proportion of disease cases potentially averted (fractional reduction of disease occurrence) from changing the current distribution of exposure in the population to some modified distribution
Person-years of life lost (PYLL)	≥ 0 for risk factor (person-years)	Difference between current life expectancy and life expectancy with exposure removed at the population level	Person-time of life lost at the population level attributable to exposure
Average potential years of life lost (PYLL)	≥ 0 for risk factor (years)	Average difference per exposed person between current life expectancy and life expectancy with exposure removed	Average loss of life expectancy per person attributable to exposure

[a] $\Pr(D)$, $\Pr(D|\overline{E})$, $\Pr(D|E)$ and $\Pr^*(D)$ denote probabilities of disease (disease risks), namely the overall probability of disease in the population, the probability of disease in the population with all exposure eliminated, the probability of disease among exposed individuals, and the overall probability of disease under a modified distribution of exposure, respectively. Alternatively, they may refer to disease rates depending on the context. The terms p_E and $p_{E|D}$ respectively refer to the overall exposure prevalence in the population and the exposure prevalence in the diseased individuals. The term RR refers to risk or rate ratios for exposed relative to unexposed individuals.

[b] Interpretations subject to conditions (see text)

a function of both the prevalence of exposure in the population, p_E, and the rate ratio or relative risk, RR.

An alternative formulation underscores this joint dependency in yet another manner. Again, upon expressing Pr(D) as

$$\Pr(D|E)p_E + \Pr\left(D|\overline{E}\right) p_{\overline{E}} \quad \text{with} \quad p_{\overline{E}} = 1 - p_E$$

and noting that

$$\Pr(D|E) = \text{RR} \times \Pr\left(D|\overline{E}\right) ,$$

the numerator in (2.10) can be rewritten as

$$p_E\Pr(D|E) - p_E\Pr(D|E)/\text{RR} .$$

From using Bayes' theorem to express $\Pr(D|E)$ as $\Pr(E|D)\Pr(D)/p_E$, it then becomes equal to

$$\Pr(D)p_{E|D}(1 - 1/\text{RR}) ,$$

after simple algebra. This yields (Miettinen 1974):

$$\text{AR} = p_{E|D}(\text{RR} - 1)/\text{RR} , \tag{2.12}$$

a function of the prevalence of exposure in diseased individuals, $p_{E|D}$, and the rate ratio or relative risk, RR.

A high relative risk can correspond to a low or high AR depending on the prevalence of exposure, which leads to widely different public health consequences. One implication is that, portability is not a usual property of AR, as the prevalence of exposure may vary widely among populations that are separated in time or location. This is in contrast with measures of association such as the relative risk or rate ratio which are more portable from one population to another, as the strength of the association between disease and exposure might vary little among populations (unless strong interactions with environmental or genetic factors are present). However, portability of RR can be questioned as well in the case of imperfect specificity of exposure assessment, since misclassification of non-exposed subjects as exposed will bias RR towards unity, which will affect differentially RR estimates in various populations depending on their exposure prevalence. This is not a problem with AR, which is not affected by imperfect specificity of exposure assessment.

Range

When the exposure considered is a risk factor (RR > 1), it follows from the above definition that AR lies between 0 and 1. Therefore, it is very often expressed as a percentage. AR increases both with the strength of the association between exposure and disease measured by RR, and with the prevalence of exposure in the population. A prevalence of 1 (or 100%) yields a value of AR equal to the attributable risk among the exposed, that is $(RR - 1)/RR$ (see Sect 2.5.2). AR approaches 1 for an infinitely high RR provided the exposure is present in the population (i.e., non-null prevalence of exposure).

AR takes a null value when either there is no association between exposure and disease ($RR = 1$) or there are no exposed subjects in the population. Negative AR values are obtained for a protective exposure ($RR < 1$). In this case, AR varies between 0 and $-\infty$, a scale on which AR lacks a meaningful interpretation. One solution is to reverse the coding of exposure (i.e., interchange exposed and unexposed categories) to go back to the situation of a positive AR, sometimes called the preventable fraction in this case (Benichou 2000c; Greenland 1987; Last 1983). Alternatively, one must consider a different parameter, namely the prevented fraction (see Sect. 2.5.4).

Synonyms

Some confusion in the terminology arises from the reported use of as many as 16 different terms in the literature to denote attributable risk (Gefeller 1990, 1995). However, a literature search by Uter and Pfahlberg (Uter and Pfahlberg 1999) found some consistency in terminology usage, with "attributable risk" and "population attributable risk" (MacMahon and Pugh 1970) the most commonly used terms by far followed by "etiologic fraction" (Miettinen 1974). Other popular terms include "attributable risk percentage" (Cole and MacMahon 1971), "fraction of etiology" (Miettinen 1974), and "attributable fraction" (Greenland and Robins 1988; Last 1983; Ouellet et al. 1979; Rothman and Greenland 1998, Chap. 4).

Moreover, additional confusion may originate in the use by some authors (MacMahon and Pugh 1970; Markush 1977; Schlesselman 1982) of the term "attributable risk" to denote a measure of association, the excess incidence, that is the difference between the incidence rates in exposed and unexposed subjects (see Sect. 2.4.4). Context will usually help the readers detect this less common use.

Interpretation and Usefulness

While measures of association such as the rate ratio and relative risk are used to establish an association in etiologic research, AR has a public health interpretation as a measure of the disease burden attributable or at least related to one or several exposures. Consequently, AR is used to assess the potential impact of prevention programs aimed at eliminating exposure from the population. It is often thought of as the fraction of disease that could be eliminated if exposure could be totally removed from the population.

However, this interpretation can be misleading because, for it to be strictly correct, the three following conditions have to be met (Walter 1976). First, estimation of AR has to be unbiased (see below). Second, exposure has to be causal rather than merely associated with the disease. Third, elimination of exposure has to be without any effect on the distribution of other risk factors. Indeed, as it might be difficult to alter the level of exposure to one factor independently of other risk factors, the resulting change in disease load might be different from the AR estimate.

For these reasons, various authors elect to use weaker definitions of AR, such as the proportion of disease that can be related or linked, rather than attributable, to exposure (Miettinen 1974).

A fundamental problem regarding causality has been discussed by Greenland and Robins (1988) and Robins and Greenland (1989) who considered the proportion of disease cases for which exposure played an etiologic role, i.e., cases for which exposure was a component cause of disease occurrence. They termed this quantity the etiologic fraction and argued that it was a more relevant measure of impact than AR. Rothman and Greenland (1998, Chap. 4) argued that AR and the etiologic fractions are different quantities using logical reasoning regarding causality and the fact that disease occurrence may require several component causal factors rather than one. The main problem with the etiologic fraction is that it is usually impossible to distinguish exposed cases for whom exposure played an etiologic role from those where exposure was irrelevant. As a consequence, estimating the etiologic fraction will typically require non-identifiable biologic assumptions about exposure actions and interactions to be estimable (Cox 1984, 1985; Robins and Greenland 1989; Seiler 1986). Thus, despite its limitations, AR remains a useful measure to assess the potential impact of exposure at the population level and can serve as a suitable guide in practice to assess and compare various prevention strategies.

Several authors have considered an interpretation of AR in terms of etiologic research. The argument is that if an AR estimate is available for several risk factors jointly, then its complement to 1, 1 − AR, must represent a gauge of the proportion of disease cases not explained by the risk factors used in estimating AR. Hence, 1 − AR would represent the proportion of cases attributable to other (possibly unknown) risk factors. For instance, it was estimated that the AR of breast cancer was 41% for late age at first birth, nulliparity, family history of breast cancer and higher socioeconomic status, which suggested that at least 59% of cases had to be attributable to other risk factors (Madigan et al. 1995). A similar type of reasoning was used in several well-known reports of estimated percentages of cancer death or incidence attributable to various established cancer risk factors (e.g., smoking, diet, occupational exposure to carcinogens …). Some of these reports conveyed the impression that little remained unexplained by factors other than the main established preventable risk factors and that cancer was a mostly preventable illness (Colditz et al. 1996, 1997; Doll and Peto 1981; Henderson et al. 1991; Ames et al. 1995). Such interpretation has to be taken with great care since ARs for different risk factors may add to more than 100% because multiple exposures are usually possible (e.g., smoking and occupational exposure to asbestos). Moreover, this interpretation can be refuted on the basis of logical arguments regarding the fact that disease occurrence may require more than one causal factor (see Rothman and Greenland 1998, Chap. 2). Furthermore, one can note that once a new risk factor is considered, the joint unexposed reference category changes from lack of exposure to all previously considered risk factors to lack of exposure to those risk factors *and* the new risk factor (Begg 2001). Because of this change in the reference category, the AR for the new risk factor may surpass the quantity 1 − AR for previously considered risk factors. Thus, while it is useful to know that only 41% of breast cancer cases can be attributed to four established risk factors in the above example, it is entirely conceivable that new risk factors of breast cancer may be elicited which yield an AR of more than 59% by themselves in the above example.

Properties

AR has two main properties. First, AR values greatly depend on the definition of the reference level for exposure (unexposed or baseline level). A more stringent definition of the reference level corresponds to a larger proportion of subjects exposed and, as one keeps depleting the reference category from subjects with higher levels of risk, AR values and estimates keep rising. This property has a major impact on AR estimates as was illustrated by Benichou (1991) and Wacholder et al. (1994). For instance, Benichou (1991) found that the AR estimate of esophageal cancer for an alcohol consumption greater or equal to 80 g/day (reference level of 0–79 g/day) was 38% in the Ille-et-Vilaine district of France, and increased dramatically to 70% for an alcohol consumption greater or equal to 40 g/day (i.e., using the more restrictive reference level 0–39 g|day) (see Example 3 below). This property plays a role whenever studying a continuous exposure with a continuous gradient of risk and when there is no obvious choice of threshold. Therefore, AR estimates must be reported with reference to a clearly defined baseline level in order to be validly interpreted.

Example 3 . A case-control study of esophageal cancer conducted in the Ille-et-Vilaine district of France included 200 cases and 775 controls selected by simple random sampling from electoral lists (Tuyns, Pequignot and Jensen 1977). The assessment of associations between alcohol consumption and smoking with esophageal cancer has been the focus of detailed illustration by Breslow and Day (1980) who presented various approaches to odds ratio estimation with or without adjustment for age. As in previous work (Benichou, 1991), four levels of alcohol consumption (0–39, 40–79, 80–119 and 120+ g/day) are considered here as well as three levels of smoking (0–9, 10–29, 30+ g/day) and three age groups (25–44, 45–54, 55+ years). There were 29, 75, 51 and 45 cases with respective alcohol consumptions of 0–39, 40–79, 80–119 and 120+ g/day. Corresponding numbers of controls were 386, 280, 87 and 22, respectively. The first reference level considered, 0–79 g/day, included 104 cases and 666 controls, leaving 96 cases and 109 controls in the exposed (i.e., 80+ g/day) category (see Table 2.5). The corresponding crude (unadjusted) odds ratio was estimated as $(96 \times 666)/(104 \times 109) = 5.6$ (see Sect. 2.4.6). Using methods described below, the crude AR estimate was 39.5% for alcohol consumption and the age- and smoking-adjusted AR estimates were close to 38%. The second reference level considered, 0–39 g/day, was more restrictive and included only 29 cases and 286 controls, leaving 171 cases and 489 controls in the exposed (i.e., 40+ g/day) category (see Table 2.5). The corresponding crude odds ratio was estimated as $(171 \times 386)/(29 \times 389) = 5.9$ (see Sect. 2.4.6). Using

Table 2.5. Numbers of cases and controls in the reference and exposed categories of daily alcohol consumption according to two definitions of the reference category – Data from a case-control study of esophageal cancer (from Tuyns, Pequignot and Jensen 1977)

More restrictive definition of reference category (0–39 g/day)

	Reference category (0–39 g/day)	Exposed category (40+ g/day)	Total
Cases	29	171	200
Controls	386	389	775
Total	315	660	975

Less restrictive definition of reference category (0–79 g/day)

	Reference category (0–79 g/day)	Exposed category (80+ g/day)	Total
Cases	104	96	200
Controls	666	109	775
Total	770	205	975

methods described below, the crude AR estimate was 70.9% and adjusted AR estimates were in the range 70% to 72%. The marked increase mainly resulted from the much higher proportion of subjects exposed with the more restrictive definition of the reference category (63% instead of 14% of exposed controls). ♦

The second main property is distributivity. If several exposed categories are considered instead of just one, then the sum of the category-specific ARs equals the overall AR calculated from combining those exposed categories into a single one, regardless of the number and the divisions of the original categories (Benichou 1991; Wacholder et al. 1994; Walter 1976), provided the reference category remains the same. This property applies strictly to crude AR estimates and to adjusted AR estimates calculated on the basis of a saturated model including all possible interactions (Benichou 1991). It applies approximately to adjusted estimates not based on a saturated model (Wacholder et al. 1994). Thus, if an overall AR estimate is the focus of interest, there is no need to break the exposed category into several mutually exclusive categories, even in the presence of a gradient of risk with increasing level of exposure. Of course, if the impact of a partial removal of exposure is the question of interest, retaining detailed information on the exposed categories will be necessary (Greenland 2001).

Example 3. *(continued)*

For the more restrictive definition of the reference category of daily alcohol consumption (0–39 g/day), the crude AR was estimated at 70.9%. The separate contributions of categories 40–79, 80-119 and 120+ g/day were 27.0%, 22.2% and 21.7%, summing to the same value 70.9%. Similarly, for the less restrictive definition of the reference category (0–79 g/day), the crude AR was estimated at 39.5% and the separate contributions of categories 80–119 g/day and 120+ g/day were 18.7% and 20.8%, summing to the same value 39.5%. ♦

Estimability and Basic Principles of Estimation

AR can be estimated from cohort studies since all quantities in (2.10), (2.11) and (2.12) are directly estimable from cohort studies. AR estimates can differ depending on whether rate ratios, risk ratios or odds ratios are used but will be numerically close for rare diseases. For case-control studies, exposure-specific incidence rates or risks are not available unless data are complemented with follow-up or population-based data (see Sect. 2.2.2). Thus, one has to rely on odds ratio estimates, use (2.11) and estimate p_E from the proportion exposed in the controls, making the rare-disease assumption also involved in estimating odds ratios rather than relative risks. For crude AR estimation, the estimate of the odds ratio is taken as ad/bc and that of p_E as $c/(c+d)$, where, as in Table 2.3, a, b, c and d respectively denote the numbers of exposed cases, unexposed cases, exposed controls and unexposed controls. Alternatively, one can use (2.12), in which the quantity $p_{E|D}$ can be directly estimated from the diseased individuals (cases) as $a/(a+b)$ and RR can be estimated from the odds ratio again as ad/bc. Using either equation, the resulting point estimate is given by $(ad-bc)/\{d(a+b)\}$.

Variance estimates of crude AR estimates are based on applying the delta-method (Rao 1965). For instance, an estimate of the variance for case-control data is given by the quantity

$$\operatorname{var}\left(\widehat{\mathrm{AR}}\right) = b(c+d)\{ad(c+d)+bc(a+b)\}/\{d^3(a+b)^3\}\ .$$

Various $(1-\alpha)\%$ confidence intervals for AR have been proposed that can be applied to all epidemiologic designs once point and variance estimates are obtained. They include standard confidence intervals for AR based on the untransformed AR point estimate, namely

$$\widehat{\mathrm{AR}} \pm z_{1-\alpha/2}se\left(\widehat{\mathrm{AR}}\right)\ ;$$

AR confidence intervals based on the log-transformed variable $\ln(1-\mathrm{AR})$, namely

$$1-\left(1-\widehat{\mathrm{AR}}\right)\left[\exp\left\{\pm z_{1-\alpha/2}se\left(\widehat{\mathrm{AR}}\right)/\left(1-\widehat{\mathrm{AR}}\right)\right\}\right] \quad \text{(Walter 1975)}\ ;$$

as well as confidence intervals based on the logit-transformed variable $\ln\{\mathrm{AR}/(1-\mathrm{AR})\}$, namely

$$\left\{1+\left\{\left(1-\widehat{\mathrm{AR}}\right)/\widehat{\mathrm{AR}}\right\}\left(\exp\left[\pm z_{1-\alpha/2}se\left(\widehat{\mathrm{AR}}\right)/\left\{\widehat{\mathrm{AR}}\left(1-\widehat{\mathrm{AR}}\right)\right\}\right]\right)\right\}^{-1}$$

(Leung and Kupper 1981).

In the previous formulae, $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$th percentile of the standard normal distribution, $\widehat{\mathrm{AR}}$ denotes the AR point estimate and $se(\widehat{\mathrm{AR}})$ its corresponding standard error estimate. Whittemore (1982) noted that the log-transformation yields a wider interval than the standard interval for AR > 0. Leung and Kupper (1981) showed that the interval based on the logit transform is narrower than the standard interval for values of AR strictly between 0.21 and 0.79, whereas the reverse holds outside this range for positive values of AR. While the coverage probabilities of these intervals have been studied in some specific situations and partial comparisons have been made, no general studies have been performed to determine their relative merits in terms of coverage probability.

Detailed reviews of estimability and basic estimation of AR for various epidemiologic designs can be found in Walter (1976) and Benichou (2000b, 2001) who provide explicit formulae for $\widehat{\mathrm{AR}}$ and $se(\widehat{\mathrm{AR}})$ for cohort and case-control designs.

Example 3. *(continued)*

For the more restrictive definition of the reference category of daily alcohol consumption (0–79 g/day), the crude AR estimate was obtained as:

$$(171 \times 386 - 29 \times 389)/(386 \times 200) = 0.709 ,$$

or 70.9%. Its variance was estimated as:

$$29 \times 775 \times (171 \times 386 \times 775 + 29 \times 389 \times 200)/\left(386^3 \times 200^3\right) = 0.00261 ,$$

yielding a standard error estimate of 0.051, or 5.1%. The corresponding 95% confidence intervals for AR are given by 60.9% to 80.9% (no transformation), 58.9% to 79.4% (log transformation), and 60.0% to 79.8% (logit transformation), very similar to each other in this example. ♦

Adjusted Estimation

As is the case for measures of association, unadjusted (or crude or marginal) AR estimates may be inconsistent (Miettinen 1974; Walter 1976, 1980, 1983). The precise conditions under which adjusted AR estimates that take into account the distribution and effect of other factors will differ from unadjusted AR estimates that fail to do so were worked out by Walter (1980). If E and X are two dichotomous factors taking levels 0 and 1, and if one is interested in estimating the AR for exposure E, then the following applies. The adjusted and unadjusted AR estimates coincide (i.e., the crude AR estimate is unbiased) if and only if (a) E and X are such that $\Pr(E = 0, X = 0)\Pr(E = 1, X = 1) = \Pr(E = 0, X = 1)\Pr(E = 1, X = 0)$, which amounts to the independence of their distributions, or (b) exposure to X alone does not increase disease risk, namely $\Pr(D|E = 0, X = 1) = \Pr(D|E = 0, X = 0)$. When considering one (or several) polychotomous factor(s) X forming J levels ($J > 2$), conditions (a) and (b) can be extended to a set of analogous sufficient conditions. Condition (a) translates into a set of $J(J-1)/2$ conditions for all pairs of levels j and j' of X, amounting to an independent distribution of E and all factors in X. Condition (b) translates into a set of $J-1$ conditions stating that in the absence of exposure to E, exposure to any of the other factors in X, alone or in combination, does not increase disease risk.

The extent of bias varies according to the severity of the departure from conditions (a) and (b) above. Although no systematic numerical study of the bias of unadjusted AR estimates has been performed, Walter (1980) provided a revealing example of a case-control study assessing the association between alcohol, smoking and oral cancer. In that study, severe positive bias was observed for crude AR estimates, with a very large difference between crude and adjusted AR estimates both for smoking (51.3% vs. 30.6%, a 20.7 difference in percentage points and 68% relative difference in AR estimates) and alcohol (52.2% vs. 37.0%, a 15.2% absolute difference and 48% relative difference). Thus, the prudent approach must be to adjust for factors that are suspected or known to act as confounders in a similar fashion as for estimating measures of associations.

Two simple adjusted estimation approaches discussed in the literature are inconsistent. The first approach was presented by Walter (1976) and is based on a factorization of the crude risk ratio into two components similar to those in Miettinen's earlier derivation (Miettinen 1972). In this approach, a crude AR estimate is obtained under the assumption of no association between exposure and disease (i.e., values of RR or the odds ratio are taken equal to 1 separately for each level of confounding). This term reflects the AR only due to confounding factors since it is obtained under the assumption that disease and exposure are not associated. By subtracting this term from the crude AR estimate that ignores confounding factors and thus reflects the impact of both exposure and confounding factors, what remains is an estimate of the AR for exposure adjusted for confounding (Walter 1976). The second approach is based on using (2.11) and plugging in a common adjusted RR estimate (odds ratio estimate in case-control studies), along with an estimate of p_E (Cole and MacMahon 1971; Morgenstern 1982). Both approaches, while intuitively appealing, were shown to be inconsistent (Ejigou 1979; Greenland and Morgenstern 1983; Morgenstern 1982) and, accordingly, very severe bias was exhibited in simulations of cross-sectional and cohort designs (Gefeller 1995).

By contrast, two adjusted approaches based on stratification yield valid estimates. The Mantel–Haenszel approach consists in plugging-in an estimate of the common adjusted RR (odds ratio in case-control studies) and an estimate of the prevalence of exposure in diseased individuals, $p_{E|D}$, in (2.12) in order to obtain an adjusted estimate of AR (Greenland 1984, 1987; Kuritz and Landis 1987, 1988a,b). In doing so, it is possible to adjust for one or more polychotomous factors forming J levels or strata. While several choices are available for a common adjusted RR or odds ratio estimator, a usual choice is to use a Mantel–Haenszel estimator of RR in cohort studies (Kleinbaum et al. 1982, Chaps. 9 and 17; Landis et al. 2000; Rothman and Greenland 1998, Chaps. 15–16; Tarone 1981) or odds ratio in case-control studies (Breslow and Day 1980, Chaps. 4–5; Kleinbaum et al. 1982, Chaps. 9, 17; Landis et al., 2000; Mantel and Haenszel 1959; Rothman and Greenland 1998, Chaps. 15–16) (see Sect. 2.4.6). For this reason, the term "Mantel–Haenszel approach" has been proposed to denote this approach to adjusted AR estimation (Benichou 1991). When there is no interaction between exposure and factors adjusted for, Mantel–Haenszel type estimators of RR or odds ratio have favorable properties, as they combine lack of (or very small) bias even for sparse data (e.g., individually matched case-control data) and good efficiency except in extreme circumstances (Birch 1964; Breslow 1981; Breslow and Day 1980, Chaps. 4–5; Landis et al. 1978; Landis et al., 2000). Moreover, variance estimators are consistent even for sparse data ("dually-consistent" variance estimators) (Greenland 1987; Robins and Greenland 1989). Simulation studies of cohort and case-control designs (Gefeller 1992; Greenland 1987; Kuritz and Landis 1988a,b) showed that adjusted AR estimates are little affected by small-sample bias when there is no interaction between exposure and adjustment factors, but can be misleading if such interaction is present.

Example 3. *(continued)*

In order to control for age and smoking, nine strata (joint categories) of smoking × age have to be considered. The Mantel–Haenszel odds ratio estimate can be calculated from quantities a_j, b_j, c_j and d_j that respectively denote the numbers of exposed cases, unexposed cases, exposed controls and unexposed controls in stratum j, using the methods in Sect. 2.4.6. With the more restrictive definition of the reference category for daily alcohol consumption, the Mantel–Haenszel odds ratio was estimated at 6.2, thus slightly higher than the crude odds ratio of 5.9. Combined with an observed proportion of exposed cases of $171/200 = 0.855$, this resulted in an adjusted AR estimate of $0.855 \times (6.2 - 1)/6.2 = 0.716$ or 71.6% using (2.12) (allowing for rounding error), slightly higher than the crude AR estimate of 70.9%. The corresponding estimate of the standard error was 5.1%. ♦

The weighted-sum approach also allows adjustment for one or more polychotomous factors forming J levels or strata. The AR is written as a weighted sum over all strata of stratum-specific ARs, i.e., $\sum_{j=1}^{J} w_j \mathrm{AR}_j$ (Walter 1976; Whittemore 1982, 1983). Using crude estimates of AR_j separately within each stratum j and setting weights w_j as proportions of diseased individuals (cases) yields an asymptotically unbiased estimator of AR, which can be seen to be a maximum-likelihood estimator (Whittemore 1982). This choice of weights defines the "case-load method". The weighted-sum approach does not require the assumption of a common relative risk or odds ratio. Instead, the relative risks or odds ratios are estimated separately for each adjustment level with no restrictions placed on them, corresponding to a fully saturated model for exposure and adjustment factors (i.e., a model with all interaction terms present). From these separate relative risk or odds ratio estimates, separate AR estimates are obtained for each level of adjustment. Thus, the weighted-sum approach not only accounts for confounding but also for interaction. Simulation studies of cohort and case-control designs (Gefeller 1992; Kuritz and Landis 1988a,b; Whittemore 1982) show that the weighted-sum approach can be affected by small sample bias, sometimes severely. It should be avoided when analyzing sparse data, and should not be used altogether for analyzing individually matched case-control data.

Example 3. *(continued)*
As with the Mantel–Haenszel approach, nine strata (joint categories) of smoking × age have to be considered in order to control for age and smoking. In each stratum separately, an AR estimate is calculated using the methods for crude AR estimation (see above). For instance, among heavy smokers (30+ g/day) in age group 65+ years, there were 15 exposed cases, five unexposed cases, four exposed controls, and six unexposed controls, yielding an odds ratio estimate of 4.5 and an AR estimate of 58.3%. The corresponding weight was 20/200 = 0.1, so that the contribution of this stratum to the overall adjusted AR was 5.8%. Summing the contributions of all nine strata yielded an adjusted AR estimate of 70.0%, thus lower than both the crude and Mantel–Haenszel adjusted AR estimates. The corresponding standard error estimate was 5.8%, higher than the standard error estimate from the Mantel–Haenszel approach because fewer assumptions were made. Namely, the odds ratio was not assumed common to all strata, so that nine separate odds ratios had to be estimated (one for each stratum) rather than a single common odds ratio from all strata. To circumvent the problem of empty cells, the standard error estimate was obtained after assigning the value 0.5 to all zero cells. ♦

A natural alternative to generalize these approaches is to use adjustment procedures based on regression models, in order to take advantage of their flexible and unified approach to efficient parameter estimation and hypothesis testing. Regression models allow one to take into account adjustment factors as well as interaction of exposures with some or all adjustment factors. This approach was first used by Walter (1976), Sturmans et al. (1977) and Fleiss (1979) followed by Deubner et al. (1980) and Greenland (1987). The full generality and flexibility of the regression approach was exploited by Bruzzi et al. (1985) who developed a general AR estimate based on rewriting AR as

$$1 - \sum_{j=1}^{J} \sum_{i=0}^{I} \varrho_{ij} \mathrm{RR}_{i|j}^{-1} \, .$$

Quantities ϱ_{ij} represent the proportion of diseased individuals with level i of exposure ($i = 0$ at the reference level, $i = 1, \ldots, I$ for exposed levels) and j of confounding and can be estimated from cohort or case-control data (or cross-sectional survey data) using the observed proportions. The quantity $\mathrm{RR}_{i|j}^{-1}$ represents the inverse of the rate ratio, risk ratio or odds ratio depending on the context, for level i of exposure at level j of confounding. It can be estimated from regression models (see Sect. 2.4.6), both for cohort and case-control data (as well as cross-sectional data), which allows confounding and interactions to be accounted for. Hence, this regression-based approach to AR estimation allows control for confounding and interaction and can be used for the main epidemiologic designs. Depending on the design, conditional or unconditional logistic, log-linear or Poisson models can be used. Variance estimators were developed based on an extension of the delta-method to implicitly related random variables in order to take into account the variability in estimates of terms ϱ_{ij} and $\text{RR}_{i|j}^{-1}$ as well as their correlations (Basu and Landis 1995; Benichou and Gail 1989, 1990b). This regression approach includes the crude and two stratification approaches as special cases and offers additional options (Benichou 1991). The unadjusted approach corresponds to models for $\text{RR}_{i|j}^{-1}$ with exposure only. The Mantel–Haenszel approach corresponds to models with exposure and confounding factors, but no interaction terms between exposure and adjustment factors. The weighted-sum approach corresponds to fully saturated models with all interaction terms between exposure and confounding factors. Intermediate models are possible, for instance models allowing for interaction between exposure and only one confounder, or models in which the main effects of some confounders are not modeled in a saturated way.

Example 3. *(continued)*
Still considering the more restrictive definition of the reference category for daily alcohol consumption, an unconditional logistic model (see Sect. 2.4.6) with two parameters, one general intercept and one parameter for elevated alcohol consumption, was fit, ignoring smoking and age. The resulting unadjusted odds ratio estimate was 5.9 as above. The formula above for $1 - \text{AR}$ reduced to a single sum with two terms ($i = 0, 1$) corresponding to unexposed and exposed categories, respectively. The resulting unadjusted AR estimate was 70.9% (standard error estimate of 5.1), identical to the crude AR estimate above. Adding eight terms for smoking and age in the logistic model increased the fit significantly ($p < 0.001$, likelihood ratio test) and yielded an adjusted odds ratio estimate of 6.3, slightly higher than the Mantel–Haenszel odds ratio estimate of 6.2 (see above). This resulted in an adjusted AR estimate of 71.9%, slightly higher than the corresponding Mantel–Haenszel AR estimate of 71.6%, and with a slightly lower standard error estimate of 5.0%. Adding two terms for interactions of smoking with alcohol consumption (thus allowing for different odds ratio estimates depending on smoking level) resulted in a decreased AR estimate of 70.3% (with a higher standard error estimate of 5.4% because of the additional parameters estimated). Adding six more terms allowed for all two-by-two interactions between alcohol consumption and joint age × smoking level and yielded a fully saturated model. Thus nine odds ratios for alcohol consumption were estimated (one for each stratum) as with the weighted-sum approach. This resulted in little change as regards AR, with an AR estimate of 70.0%, identical to the AR estimate with the weighted sum approach, which precisely corresponds to a fully saturated model. The corresponding standard error estimate was increased to 5.6% due to the estimation of additional parameters. ♦

A modification of Bruzzi et al.'s approach was developed by Greenland and Drescher (1993) in order to obtain full maximum likelihood estimates of AR. The modification consists in estimating the quantities ρ_{ij} from the regression model rather than simply relying on the observed proportions of cases. The two model-based approaches seem to differ very little numerically (Greenland and Drescher 1993). Greenland and Drescher's approach might be more efficient in small samples although no difference was observed in simulations of the case-control design even for samples of 100 cases and 100 controls (Greenland and Drescher 1993). It might be less robust to model misspecification, however, as it relies more heavily on the RR or odds ratio model used. Finally, it does not apply to the conditional logistic model, and if that model is to be used (notably, in case-control studies with individual matching), the original approach of Bruzzi et al. is the only possible choice.

Detailed reviews of adjusted AR estimation (Benichou 1991, 2001; Coughlin et al. 1994; Gefeller 1992) are available. Alternative methods to obtain estimates of variance and confidence intervals for AR have been developed either based on resampling techniques (Gefeller 1992; Greenland 1992; Kahn et al. 1998; Kooperberg and Petitti 1991; Llorca and Delgado-Rodriguez 2000; Uter and Pfahlberg 1999) or on quadratic equations (Lui 2001a,b, 2003).

Final Notes and Additional References

General problems of AR definition, interpretation and usefulness as well as properties have been reviewed in detail (Benichou 2000b; Gefeller 1992; Miettinen 1974; Rockhill et al. 1998a,b; Walter 1976). Special issues were reviewed by Benichou (2000b, 2001). They include estimation of AR for risk factors with multiple levels of exposure or with a continuous form, multiple risk factors, recurrent disease events, and disease classification with more than two categories. They also include assessing the consequences of exposure misclassification on AR estimates. Specific software for attributable risk estimation (Kahn et al. 1998; Mezzetti et al. 1996) as well as a simplified approach to confidence interval estimation (Daly 1998) have been developed to facilitate implementation of methods for attributable risk estimation. Finally, much remains to be done to promote proper use and interpretation of AR as illustrated in a recent literature review (Uter and Pfahlberg 2001).

Attributable Risk Among the Exposed 2.5.2

The attributable risk in the exposed (AR_E) or attributable fraction in the exposed is defined as the following ratio (Cole and MacMahon 1971; Levin 1953; MacMahon and Pugh 1970; Miettinen 1974):

$$\mathrm{AR}_E = \left\{\Pr(D|E) - \Pr\left(D|\overline{E}\right)\right\} / \Pr(D|E) \,, \tag{2.13}$$

where $\Pr(D|E)$ is the probability of disease in the exposed individuals (E) and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same subjects but with all exposure eliminated. Depending on the context, these probabilities will refer to disease risk or may be replaced with incidence rates (see Sect. 2.5.1). AR_E can be rewritten as:

$$\mathrm{AR}_E = (\mathrm{RR} - 1)/\mathrm{RR} \,, \tag{2.14}$$

where RR denotes the risk or rate ratio. Following Greenland and Robins (1988), Rothman and Greenland (1998, Chap. 4) proposed to use the terms "excess fraction" for the definition of AR_E based on risks or risk ratios and "rate fraction" for the definition of AR_E based on rates or rate ratios.

Like AR, AR_E lies between 0 and 1 when exposures considered are risk factors ($\mathrm{RR} > 1$) with a maximal limiting value of 1, is equal to zero in the absence of association between exposure and disease ($\mathrm{RR} = 1$), and is negative for protective exposures ($\mathrm{RR} < 1$).

As for AR, AR_E has an interpretation as a measure of the disease burden attributable or at least related to one or several exposures among the exposed subjects. Consequently, AR_E could be used to assess the potential impact of prevention programs aimed at eliminating exposure from the population. These interpretations are subject to the same limitations as corresponding interpretations for AR however (see Sect. 2.5.1). Moreover, AR_E does not have a clear public health interpretation because it does not depend on the exposure prevalence but only on the risk or rate ratio of which it is merely a one-to-one transformation. For the assessment of the relative impact of several exposures, AR_E will not be an appropriate measure since AR_E for different exposures refer to different groups of subjects in the population (i.e., subjects exposed to each given exposure).

AR_E being a one-to-one function of RR, issues of estimability and estimation for AR_E are similar to those for RR. They depend on whether rates or risks are considered. For case-control studies, odds ratios can be used. Greenland (1987) specifically derived adjusted point estimates and confidence intervals for AR_E based on the Mantel–Haenszel approach.

2.5.3 Sequential and Partial Attributable Risks

Upon considering multiple exposures, separate ARs can be estimated for each exposure as well as the overall AR for all exposures jointly. Except in very special circumstances worked out by Walter (1983) (i.e., lack of joint exposure or additive effects of exposures on disease risk or rate), the sum of separate AR estimates over all exposures considered will not equal the overall AR estimate.

Because this property is somewhat counter-intuitive and generates misinterpretations, three alternative approaches have been suggested, one based on considering variance decomposition methods (Begg et al. 1998) rather than estimating AR, one based on estimating assigned share or probability of causation of a given exposure with relevance in litigation procedures for individuals with multiple exposures (Cox 1984, 1985; Lagakos and Mosteller 1986; Seiler 1986; Seiler and Scott 1987; Benichou 1993b; McElduff et al. 2002), and one based on an extension of the concept of AR (Eide and Gefeller 1995; Land et al. 2001). This last approach relies on partitioning techniques (Gefeller et al. 1998; Land and Gefeller 1997) and keeps with the framework of AR estimation by introducing the sequential AR that generalizes the concept of AR. The principle is to define an order among the exposures considered. Then, the contribution of each exposure is assessed sequentially according to that order. The contribution of the first exposure considered is calculated as the standard AR for that exposure separately. The contribution of the second exposure is obtained as the difference between the joint AR estimate for the first two exposures and the separate AR estimate for the first exposure, the contribution of the third exposure is obtained as the difference between the joint AR estimates for the first three and first two exposures, etc …. Thus, a multidimensional vector consisting of contributions of each separate exposure is obtained.

These contributions are meaningful in terms of potential prevention programs that consider successive rather than simultaneous elimination of exposures from the population. Indeed, each step yields the additional contribution of the elimination of a given exposure once higher-ranked exposures are eliminated. At some point, additional contributions may become very small, indicating that there is not much point in considering extra steps. By construction, these contributions sum to the overall AR for all exposures jointly, which constitutes an appealing property. Of course, separate vectors of contributions are obtained for different orders. Meaningful orders depend on practical possibilities in implementing potential prevention programs in a given population. Average contributions can be calculated for each given exposure by calculating the mean of contributions corresponding to that exposure over all possible orders. These average contributions have been termed partial attributable risks (Eide and Gefeller 1995) and represent another potentially useful measure. Methods for visualizing sequential and partial ARs are provided by Eide and Heuch (2001). An illustration is given by Fig. 2.1. A detailed review of properties, interpretation, and variants of sequential and partial ARs was provided by Land et al. (2001).

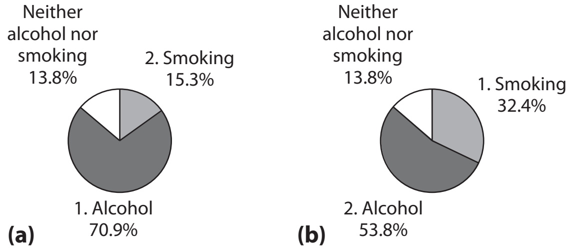

Figure 2.1. Sequential attributable risk estimates for elevated alcohol consumption (80+ g/day) and heavy smoking (10+ g/day) for two different orders of removal (a: alcohol, then smoking; b: smoking, then alcohol) – Case-control data on esophageal cancer (Tuyns, Pequignot and Jensen 1977; cf. Example 3)

Example 3. *(continued)*

Smoking is also a known risk factor of esophageal cancer so that it is important to estimate the impact of smoking and the joint impact of smoking and alcohol consumption on esophageal cancer in addition to that of alcohol consumption alone. Using the first category (i.e., 0–9 g/day) as the reference level of smoking, there were 78 cases in the reference level of smoking, 122 cases in the exposed level (i.e., 10+ g/day), 447 controls in the reference level and 328 controls in the exposed level. From these data, the crude odds ratio estimate for smoking at least 10 g/day was 2.1 and the crude AR estimate for smoking at least 10 g/day was 32.4%. Moreover, there were nine cases and 252 controls in the joint reference level of alcohol consumption and smoking (i.e., 0–39 g/day of alcohol and 0–9 g/day of tobacco), which yielded a crude joint odds ratio estimate of 10.2 and a crude joint AR estimate for drinking at least 40 g/day of alcohol or smoking at least 10 g/day of tobacco of 86.2%.

Furthermore, the crude AR estimate for alcohol consumption of at least 40 g/day was estimated at 70.9% in Sect. 2.5.1. Hence, considering the first order of risk factor removal (i.e., eliminating alcohol consumption above 39 g/day followed by eliminating smoking above 9 g/day) yields sequential AR estimates of 70.9% for elevated daily alcohol consumption and 86.2% − 70.9% = 15.3 percentage points for heavy smoking so that, once elevated alcohol consumption is eliminated, the additional impact of eliminating heavy smoking appears rather limited (Fig. 2.1a). Considering the second order (i.e., eliminating heavy smoking first) yields sequential AR estimates of 32.4% for heavy smoking and 86.2% − 32.4% = 53.8 percentage points for elevated alcohol consumption so that, once heavy smoking is eliminated, the additional impact of eliminating elevated alcohol consumption remains major (Fig. 2.1b). A summary of these results is provided by partial ARs for elevated alcohol consumption and heavy smoking, with estimated values of 62.4% and 23.9%, respectively, again reflecting the higher impact of elevated alcohol consumption on esophageal cancer. ♦

2.5.4 Preventable and Prevented Fractions

When considering a protective exposure or intervention, an appropriate alternative to AR is the preventable or prevented fraction (PF) defined as the ratio (Miettinen 1974):

$$\mathrm{PF} = \left\{\Pr\left(D|\overline{E}\right) - \Pr(D)\right\} / \Pr(D|\overline{E}) \; , \tag{2.15}$$

where $\Pr(D)$ is the probability of disease in the population, which may have some exposed (E) and some unexposed ($\overline{E}$) individuals, and $\Pr(D|\overline{E})$ is the hypothetical probability of disease in the same population but with all (protective) exposure eliminated. Depending on the context, these probabilities will refer to disease risk or may be replaced with incidence rates (see sections above). PF can be rewritten as:

$$\mathrm{PF} = p_E(1 - \mathrm{RR}) \; , \tag{2.16}$$

a function of both the prevalence of exposure, p_E, and the risk or rate ratio, RR. Thus, a strong association between exposure and disease may correspond to a high or low value of PF depending on the prevalence of exposure, as for AR. Moreover, portability is not a typical property of PF, as for AR. As for AR again, it may be useful to compare PF estimates among population subgroups to target prevention efforts to specific subgroups with a potentially high impact (as measured by the PF).

For a protective factor ($\mathrm{RR} < 1$), PF lies between 0 and 1 and increases with the prevalence of exposure and the strength of the association between exposure and disease.

PF measures the impact of an association between a protective exposure and disease at the population level. It has a public health interpretation as the proportion of disease cases averted ("prevented fraction") in relation to the presence of a protective exposure or intervention in the population, among the totality of cases that would have developed in the absence of that factor or intervention in the population. In this case, it is useful to assess prevention programs *a posteriori*. Alternatively, it can be used to assess prevention programs *a priori* by measuring the proportion of cases that could be potentially averted ("preventable fraction") if a protective exposure or intervention were introduced *de novo* in the population (Gargiullo et al. 1995). These interpretations are subject to the same limitations as corresponding interpretations for AR however (see Sect. 2.5.1).

PF and AR are mathematically related through (Walter 1976):

$$1 - \mathrm{PF} = 1/(1 - \mathrm{AR})\ . \tag{2.17}$$

From (2.17), it appears that, for a protective factor, PF estimates will usually differ from AR estimates obtained by reversing the coding of exposure. This follows from the respective definitions of AR and PF. While AR, with reverse coding, measures the potential reduction in disease occurrence that could be achieved if all subjects in the current population became exposed, PF measures the reduction in disease occurrence obtained from introducing exposure at the current prevalence in a formally unexposed population (Benichou 2000c).

In view of (2.17), estimability and estimation issues are similar for AR and PF. Specific PF adjusted point and confidence interval estimates were derived using the Mantel–Haenszel approach (Greenland 1987) and weighted-sum approaches (Gargiullo et al. 1995).

2.5.5 Generalized Impact Fraction

The generalized impact fraction (GIF) or generalized attributable fraction was introduced by Walter (1980), and Morgenstern and Bursic (1982) as the ratio:

$$\mathrm{GIF} = \left\{\mathrm{Pr}(D) - \mathrm{Pr}^*(D)\right\}/\mathrm{Pr}(D)\ , \tag{2.18}$$

where $\mathrm{Pr}(D)$ and $\mathrm{Pr}^*(D)$ respectively denote the probability of disease under the current distribution of exposure and under a modified distribution of exposure.

As for AR and PF, these probabilities denote risks or can be replaced by incidence rates depending on the context.

The generalized impact fraction not only depends on the association between exposure and disease as well as the current distribution (rather than just the prevalence) of exposure, but also on the target distribution of exposure considered that will yield $\Pr^*(D)$. It is a general measure of impact that includes AR and PF as special cases. AR contrasts the current distribution of exposure with a modified distribution defined by the absence of exposure. Conversely, PF contrasts a distribution defined by the absence of exposure with the current distribution of exposure (prevented fraction) or target distribution of exposure (preventable fraction).

The generalized impact fraction can be interpreted as the fractional reduction of disease occurrence that would result from changing the current distribution of exposure in the population to some modified distribution. Thus, it can be used to assess prevention programs or interventions, targeting all subjects or subjects at specified levels, and aimed at modifying or shifting the exposure distribution (reducing exposure), but not necessarily eliminating exposure. For instance, heavy smokers could be specifically targeted by interventions rather than all smokers. The special AR case corresponds to the complete elimination of exposure by considering a modified distribution putting unit mass on the lowest risk configuration and can be used to assess interventions aimed at eliminating (rather than reducing) exposure. Alternatively, the general impact fraction could be used to assess the increase in disease occurrence as a result of exposure changes in the population, such as the increase in breast cancer incidence as a result of delayed childbearing (Kleinbaum et al. 1982, Chap. 9). Such interpretations are subject to the same limitations as for AR and PF (see Sect. 2.5.1).

The generalized impact fraction has been used for instance by Lubin and Boice (1989) who considered the impact on lung cancer of a modification in the distribution of radon exposure consisting in truncating the current distribution at various thresholds and by Wahrendorf (1987) who examined the impact of various changes in dietary habits on colo-rectal and stomach cancers.

Issues of estimability are similar to those for AR and PF. Methods to estimate the generalized impact fraction are similar to methods for estimating AR and PF. However, unlike for AR or PF, it might be useful to retain the continuous nature of exposures to define the modification of the distribution considered (for instance a shift in the distribution), and extensions of methods for estimating AR for continuous factors (Benichou and Gail 1990b) are relevant in this context. Drescher and Becher (1997) proposed extending model-based approaches of Bruzzi et al. (1985) and Greenland and Drescher (1993) to estimate the generalized impact fraction in case-control studies and considered continuous as well as categorical exposures.

2.5.6 Person-Years of Life Lost

Person-years of life lost (or potential years of life lost, PYLL) for a given cause of death is a measure defined as the difference between current life expectancy of the population and potential life expectancy with the cause of death eliminated (Smith 1998). For instance, one may be interested in PYLL due to prostate cancer in men, breast cancer in women, or cancer as a whole (all sites) in men and women. Methods for estimating PYLL rely on calculating cause-deleted life tables. Total PYLL at the population level or average PYLL per person may be estimated. As an example, a recent report from the Surveillance, Epidemiology and End Results (SEER) estimated that 8.4 million years of life overall were lost due to cancer in the US population (both sexes, all races) in the year 2001, with an average value of potential years of life lost per person of 15.1 years. Corresponding numbers were 779,900 years overall and 18.8 years on average for breast cancer in women, and 275,200 years overall and 9.0 years on average for prostate cancer in men (Ries et al. 2004).

PYLL represents an assessment of the impact of a given disease. Thus, it is not directly interpretable as a measure of exposure impact, except perhaps for diseases with a dominating risk factor, such as asbestos exposure for mesothelioma or human papilloma virus for cervical cancer.

However, it is possible to obtain a corresponding measure of the impact of a given exposure by converting PYLL due to a particular cause of death to PYLL due to a particular exposure. Estimation of an exposure-specific PYLL is obtained through applying an AR estimate for that exposure to the disease-specific PYLL, namely calculating the product PYLL times AR, which yields the fraction of PYLL attributable to exposure. In this process, several causes of deaths may have to be considered. For instance, the fractions of PYLL for mesothelioma and lung cancer would need to be added in order to obtain the overall PYLL for asbestos exposure. In contrast with AR that provides a measure of exposure impact as a fraction of disease incidence (or death), such calculations of PYLL will provide a measure of exposure impact on the life expectancy scale. As for AR, the impact of a given exposure on the PYLL scale will depend on the prevalence of exposure in the population and strength of association between exposure and disease(s). Moreover, it will depend critically on the age-distribution of exposure-associated diseases and their severity, i.e. case fatality.

2.6 Other Topics

2.6.1 Standardization of Risks and Rates

Risks and rates can usually not be directly compared between countries, regions or time periods because of differences in age structure. For example, an older population may appear to have higher rates of certain cancers, not because of the presence of risk factors, but because of the higher age itself. This is a form of confounding. In the tradition of demography, so called standardization is applied to reported rates and risks to adjust for differences in age and possibly other confounders. Direct standardization is the most commonly used technique. It proceeds by forming a weighted average of age specific rates or risks, where the weights reflect a known population structure. This structure is typically chosen as that of a country in a given census year, the so-called *standard population*. A directly standardized rate can be written:

$$\mathrm{SR} = \sum n_j^S h_j / \sum n_j^S = \sum w_j^S h_j \ , \tag{2.19}$$

where n_j^S is the number of individuals in age group j in the standard population, h_j are age specific rates in the population under study, and w_j^S are weights such that $\sum w_j^S = 1$. A standardized risk can be computed in the same manner. Since the weights are fixed and not estimated, the variance of the estimated standardized rate is

$$\mathrm{var}\left(\sum w_j^S h_j\right) = \sum \left(w_j^S\right)^2 h_j \ , \tag{2.20}$$

based on the Poisson assumption for the age specific rates. For risks, the binomial assumption may be used for the age specific risks.

When age-specific rates or risks are not available in the population under study, *indirect standardization* may be used. This technique is less common, but requires knowledge only of the age distribution, and not the age-specific rates, in the population under study. The indirectly standardized rate is obtained by (SMR)(CR^0), where SMR is the *standardized mortality or morbidity ratio* (see below), and CR^0 is the *crude* (i.e., original overall) rate in a reference population that provides stratum-specific rates.

The standardized mortality or morbidity ratio is a ratio between observed and expected event counts, where the expected count is based on age specific rates or risks in a reference population, which is a non-exposed or general population group. Then the standardized mortality (or morbidity or incidence) ratio

$$\text{SMR or SIR} = D_E/E_0 = \sum n_j h_j / \sum n_j h_{0j} \ ,$$

where D_E is the number of events in the exposed and E_0 is the expected number of events obtained from the rates h_{0j} in the unexposed applied to the sample composition of the exposed. The SMR can also be re-written as a weighted average of sex- and age-specific (say) rate ratios h_j/h_{0j} with weights $w_j = n_j h_{0j}$. It can be shown that these weights minimize the standard error of the weighted average (Breslow and Day 1987, Chap. 2) as long as the rates in the reference population are assumed to be known rather than estimated. Stratified analyses as discussed above, on the other hand, choose weights that minimize the standard errors when the rates are estimated among both the exposed and the unexposed. The SMR has the advantage that age- and sex-specific rates are not needed for the exposed group.

The denominator of the SMR is generally obtained from age- and sex-specific rates in the entire regional population. This allows the random variation of the denominator to be considered to be none, and confidence intervals can be based on the estimate $1/D_E^{1/2}$ for the standard error of ln(SMR). This standard error computation also assumes that the events among the exposed are uncorrelated (do not cluster), or more specifically, that the event count follows a Poisson distribution.

It may be noted that directly standardized rates are based on a choice of standard population to generate weights. While the weights used for the SMR result from the composition of the comparison group and do not involve a true standard population, the weights used in direct standardization are external as they result from information outside the samples being compared. In principle, the latter weights are similar to survey weights applied in for examples the National Health and Nutrition Examination Survey (NHANES), where the sample must be standardized to the US population to account for the methodology used in drawing it. While improving external validity, weights from direct standardization and survey weights always result in loss of statistical efficiency, i.e., standard errors will be larger than for crude, or non-weighted rates and risks. In contrast, many of the methods to adjust for confounding discussed in Sect. 2.4 are internal to the specific comparison and designed to optimize statistical efficiency.

Measures Based on Prevalence 2.6.2

Prevalence is the number of cases either at a given point in time (point prevalence) or over a time period (period prevalence) divided by the population size. Prevalence can be easier to obtain than incidence. For example, a population survey can determine how many individuals in a population suffer from a given illness or health condition at a point in time.

Measures of association based on prevalence parallel those for risk (for point prevalence) or incidence rates (for period prevalence). For example, one can form prevalence ratios, prevalence differences and prevalence odds ratios. Measures of impact based on prevalence can also be obtained.

Prevalence and the measures of association based on it are useful entities for health policy planning and for determining the level of services needed for individuals with a given health condition in the population. It is usually considered less useful for studying the etiology of a disease. The reason for this is that under certain assumptions prevalence of a disease equals its incidence multiplied by its duration (Kleinbaum et al. 1982, Chap. 8). These assumptions are that the population is stable, and that both the incidence and prevalence remain constant. Under more general conditions, prevalence still reflects both incidence and duration, but in a more complex manner. For a potentially fatal or incurable disease, duration means survival, and the exposures that increase incidence may reduce or increase survival and hence the association of an exposure with prevalence may be very different than its association with incidence. On the other hand, when a disease or condition can be of limited duration due to recovery or cure, and its duration is maintained by the same exposures that caused it, prevalence can be more meaningful than incidence. For example, it is conceivable that weight gain in a person may have caused hypertension, and when the person loses the same amount of weight she/he moves out of being hypertensive. In this latter case, the prevalence ratio between the percentages with hypertension in those exposed and unexposed to the risk factor captures the increase in the risk of living with the condition caused by the exposure, while the incidence ratio captures only part of the etiologic association.

2.7 Conclusions

Disease frequency is measured through the computation of incidence rates or estimation of disease risk. Both measures are directly accessible from cohort data. They can be obtained from case-control data only if they are complemented by follow-up or population data. Using regression techniques, methods are available to derive incidence rates or risk estimates specific to a given exposure profile. Exposure-specific risk estimates are useful in individual prediction.

A wide variety of options and techniques are available for measuring association. The odds ratio is presently the most often used measure of association for both cohort and case control studies. Adjustment for confounding is key in all analyses of observational studies, and can be pursued by standardization, stratification and by regression techniques. The flexibility of the latter, especially in the generalized linear model framework, and availability of computer software, has made it widely applied in the last several years.

Several measures are available to assess the impact of an exposure in terms of the occurrence of new disease cases at the population level, among which the attributable risk is the most commonly used. Several approaches have been developed to derive adjusted estimates of the attributable risk from case-control as well as cohort data, either based on stratification or on more flexible regression techniques. The concept of attributable risk has been extended to handle preventive exposures, multiple exposures, as well as assessing the impact of various modifications of the exposure distribution rather than the mere elimination of exposure.

References

Aalen O (1978) Nonparametric estimation of partial transition probabilities in multiple decrement models. Annals of Statistics 6:534–545

Aalen O, Johansen S (1978) An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics 5:141–150

American Cancer Society (1992) Cancer facts and figures. American Cancer Society, Atlanta, Georgia

Ames BN, Gold LS, Willett WC (1995) The causes and prevention of cancer. Proceedings of the National Academy of Sciences of the United States of America 254:1131–1138

Andersen PK, Gill RD (1982) Cox's regression models for counting processes: A large-sample study. Annals of Statistics 4:1100–1120

Anderson KM, Wilson PW, Odell PM, Kannel WB (1991) Cardiovascular disease risk profiles. A statement for health professionals. Circulation 83:356–362

Anderson SJ, Ahnn S, Duff K (1992) NSABP Breast Cancer Prevention Trial Risk Assessment Program, Version 2. University of Pittsburgh Department of Biostatistics, Pittsburgh, PA

Basu S, Landis JR (1995) Model-based estimation of population attributable risk under cross-sectional sampling. American Journal of Epidemiology 142:1338–1343

Begg CB (2001) The search for cancer risk factors: When can we stop looking? American Journal of Public Health 91:360–364

Begg CB, Satagopan JM, Berwick M (1998) A new strategy for evaluating the impact of epidemiologic risk factors for cancer with applications to melanoma. Journal of the American Statistical Association 93: 415–426

Benichou J (1991) Methods of adjustment for estimating the attributable risk in case-control studies: A review. Statistics in Medicine 10:1753–1773

Benichou J (1993a) A computer program for estimating individualized probabilities of breast cancer. Computers and Biomedical Research 26:373–382

Benichou J (1993b) Re: "Methods of adjustment for estimating the attributable risk in case-control studies: A review" (letter). Statistics in Medicine 12:94–96

Benichou J (2000a) Absolute risk. In: Gail MH, Benichou J (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 1–17

Benichou J (2000b) Attributable risk. In: Gail MH, Benichou J (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 50–63

Benichou J (2000c) Preventable fraction. In: Gail MH, Benichou J (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 736–737

Benichou J (2001) A review of adjusted estimators of the attributable risk. Statistical Methods in Medical Research 10:195–216

Benichou J, Gail MH (1989) A delta-method for implicitely defined random variables. American Statistician 43:41–44

Benichou J, Gail MH (1990a) Estimates of absolute cause-specific risk in cohort studies. Biometrics 46:813–826

Benichou J, Gail MH (1990b) Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. Biometrics 46:991–1003

Benichou J, Gail MH (1995) Methods of inference for estimates of absolute risk derived from population-based case-control studies. Biometrics 51:182–194

Benichou J, Wacholder S (1994) A comparison of three approaches to estimate exposure-specific incidence rates from population-based case-control data. Statistics in Medicine 13:651–661

Benichou J, Gail MH, Mulvihill JJ (1996) Graphs to estimate an individualized risk of breast cancer. Journal of Clinical Oncology 14:103–110

Benichou J, Byrne C, Gail MH (1997) An approach to estimating exposure-specific rates of breast cancer from a two-stage case-control study within a cohort. Statistics in Medicine 16:133–151

Berkson J (1958) Smoking and lung cancer. Some observations on two recent reports. Journal of the American Statistical Association 53:28–38

Birch MW (1964) The detection of partial associations, I: The 2 × 2 case. Journal of the Royal Statistical Society, Series B 27:313–324

Borgan Ø (1998) Nelson-Aalen estimator. In: Armitage P, Colton T (eds) Encyclopedia of epidemiologic methods. Wiley, Chichester, pp 2967–2972

Breslow NE (1981) Odds ratio estimators when the data are sparse. Biometrika 68:73–84

Breslow NE, Day NE (1980) Statistical methods in cancer research vol I: The analysis of case-control studies. International Agency for Research on Cancer Scientific Publications No. 32, Lyon

Breslow NE, Day NE (1987) Statistical methods in cancer research vol II: The design and analysis of cohort studies. International Agency for Research on Cancer Scientific Publications No. 82, Lyon

Breslow NE, Lubin JH, Marek P, Langholz B (1983) Multiplicative models and cohort analysis. Journal of the American Statistical Association 78:1–12

Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C (1985) Estimating the population attributable risk for multiple risk factors using case-control data. American Journal of Epidemiology 122:904–914

Chiang CL (1968) Introduction to stochastic processes in biostatistics. Wiley, New York

Colditz G, DeJong W, Hunter D, Trichopoulos D, Willett W (eds) (1996) Harvard report on cancer prevention, vol 1. Cancer Causes and Control 7(suppl.):S3–S59

Colditz G, DeJong W, Hunter D, Trichopoulos D, Willett W (eds) (1997) Harvard report on cancer prevention, vol 2. Cancer Causes and Control 8(suppl.):S1–S50

Cole P, MacMahon B (1971) Attributable risk percent in case-control studies. British Journal of Preventive and Social Medicine 25:242–244

Cornfield J (1951) A method for estimating comparative rates from clinical data: Applications to cancer of the lung, breast and cervix. Journal of the National Cancer Institute 11:1269–1275

Cornfield J (1956) A statistical problem arising from retrospective studies. In: Neyman J (ed) Proceedings of the Third Berkeley Symposium, vol IV. University of California Press, Monterey, pp 133–148

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EI (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. Journal of the National Cancer Institute 22:173–203

Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS (1999) Validation studies for models projecting the risk of invasive and total breast cancer incidence. Journal of the National Cancer Institute 91:1541–1548

Coughlin SS, Benichou J, Weed DL (1994) Attributable risk estimation in case-control studies. Epidemiologic Reviews 16:51–64

Cox DR (1972) Regression models and lifetables (with discussion). Journal of the Royal Statistical Society, Series B 34:187–220

Cox DR (1975) Partial likelihood. Biometrika 62:269–276

Cox LA (1984) Probability of causation and the attributable proportion of risk. Risk Analysis 4:221–230

Cox LA (1985) A new measure of attributable risk for public health applications. Management Science 7:800–813

Cutler SJ, Ederer F (1958) Maximum utilization of the life table method in analyzing survival. Journal of Chronic Diseases 8:699–712

Daly LE (1998) Confidence limits made easy: interval estimation using a substitution method. American Journal of Epidemiology 147:783–790

Deubner DC, Wilkinson WE, Helms MJ, Tyroler HA, Hames CG (1980) Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. American Journal of Epidemiology 112:135–143

Doll R, Peto R (1981) The causes of cancer. Oxford University Press, New York

Dorey FJ, Korn EL (1987) Effective sample sizes for confidence intervals for survival probabilities. Statistics in Medicine 6:679–687

Drescher K, Becher H (1997) Estimating the generalized attributable fraction from case-control data. Biometrics 53:1170–1176

Dupont DW (1989) Converting relative risks to absolute risks: A graphical approach. Statistics in Medicine 8:641–651

Easton DF, Peto J, Babiker AG (1991) Floating absolute risk: An alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. Statistics in Medicine 10:1025–1035

Eide GE, Gefeller O (1995) Sequential and average attributable fractions as aids in the selection of preventive strategies. Journal of Clinical Epidemiology 48:645–655

Eide GE, Heuch I (2001) Attributable fractions: fundamental concepts and their vizualization. Statistical Methods in Medical Research 10:159–193

Ejigou A (1979) Estimation of attributable risk in the presence of confounding. Biometrical Journal 21:155–165

Elandt-Johnson RC (1977) Various estimators of conditional probabilities of death in follow-up studies. Summary of results. Journal of Chronic Diseases 30:247–256

Elveback L (1958) Estimation of survivorship in chronic disease: The "actuarial" method. Journal of the American Statistical Association 53:420–440

Fleiss JL (1979) Inference about population attributable risk from cross-sectional studies. American Journal of Epidemiology 110:103–104

Fleiss JL, Dunner DL, Stallone F, Fieve RR (1976) The life table: A method for analyzing longitudinal studies. Archives of General Psychiatry 33:107–112

Fisher B, Costantino JP, Wickerham L, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, Daly M, Wieand S, Tan-Chiu E, Ford L, Womark N, other National Surgical Adjuvant Breast and Bowel project Investigators (1998) Tamoxifen for prevention of breast cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. Journal of the National Cancer Institute 90:1371–1388

Gail MH (1975) Measuring the benefit of reduced exposure to environmental carcinogens. Journal of Chronic Diseases 28:135–147

Gail MH, Benichou J (1994) Validation studies on a model for breast cancer risk (editorial). Journal of the National Cancer Institute 86:573–575

Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute 81:1879–1886

Gail MH, Costantino JP, Bruant J, Croyle R, Freedman L, Helzsouer K, Vogel V (1999) Weighing the risks and benefits of Tamoxifen treatment for preventing breast cancer. Journal of the National Cancer Institute 91:1829–1846

Gargiullo PM, Rothenberg R, Wilson HG (1995) Confidence intervals, hypothesis tests, and sample sizes for the prevented fraction in cross-sectional studies. Statistics in Medicine 14:51–72

Gefeller O (1990) Theory and application of attributable risk estimation in cross-sectional studies. Statistica Applicata 2:323–331

Gefeller O (1992) The bootstrap method for standard errors and confidence intervals of the adjusted attributable risk [letter]. Epidemiology 3:271–272

Gefeller O (1995) Definitions of attributable risk-revisited. Public Health Reviews 23:343–355

Gefeller O, Land M, Eide GE (1998) Averaging attributable fractions in the multifactorial situation: Assumptions and interpretation. Journal of Clinical Epidemiology 51:437–451

Gray RJ (1988) A class of k-sample tests for comparing the cumulative incidence of a competing risk. Annals of Statistics 16:1141–1151

Greenland S (1981) Multivariate estimation of exposure-specific incidence from case-control studies. Journal of Chronic Diseases 34:445–453

Greenland S (1984) Bias in methods for deriving standardized mortality ratio and attributable fraction estimates. Statistics in Medicine 3:131–141

Greenland S (1987) Variance estimators for attributable fraction estimates, consistent in both large strata and sparse data. Statistics in Medicine 6:701–708

Greenland S (1992) The bootstrap method for standard errors and confidence intervals of the adjusted attributable risk [letter]. Epidemiology 3:271

Greenland S (2001) Attributable fractions: bias from broad definition of exposure. Epidemiology 12:518–520

Greenland S, Drescher K (1993) Maximum-likelihood estimation of the attributable fraction from logistic models. Biometrics 49:865–872

Greenland S, Morgenstern H (1983) Morgenstern corrects a conceptual error [letter]. American Journal of Public Health 73:703–704

Greenland S, Robins JM (1988) Conceptual problems in the definition and interpretation of attributable fractions. American Journal of Epidemiology 128:1185–1197

Greenland S, Thomas DC (1982) On the need for the rare disease assumption. American Journal of Epidemiology 116:547–553

Hartman LC, Sellers TA, Schaid DJ, Franks TS, Soderberg CL, Sitta DL, Frost MH, Grant CS, Donohue JH, Woods JE, McDonnell SK, Vockley CW, Deffenbaugh A, Couch FJ, Jenkins RB (2001) Efficacy of bilateral prophylactic mastectomy in BRCA1 and BRCA2 gene mutation carriers. Journal of the National Cancer Institute 93:1633–1637

Henderson BE, Ross RK, Pike MC (1991) Toward the primary prevention of cancer. Science 254:1131–1138

Holford TR (1980) The analysis of rates and of survivorship using log-linear models. Biometrics 36:299–305

Hosmer D, Lemeshow S (1999) Applied survival analysis: Regression modeling of time to event data. John Wiley & Sons, Hoboken, New Jersey

Hoskins KF, Stopfer JE, Calzone K, Merajver SD, Rebbeck TR, Garber JE, Weber BL (1995) Assessment and counseling for women with a family history of breast cancer. A guide for clinicians. Journal of the American Medical Association 273:577–585

Johnson NL, Kotz S (1970) Distributions in statistics, vol 2. Houghton-Mifflin, Boston

Kahn HA, Sempos CT (1989) Statistical methods in epidemiology. Monographs in epidemiology and biostatistics, vol 12, Oxford University Press, Oxford, New York

Kahn MJ, O'Fallon WM, Sicks JD (1998) Generalized population attributable risk estimation. Technical Report #54, Mayo Foundation, Rochester, Minnesota

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53:457–481

Katz D, Baptista J, Azen SP, Pike MC (1978) Obtaining confidence intervals for the risk ratio in a cohort study. Biometrics 34:469–474

Kay R, Schumacher M (1983) Unbiased assessment of treatment effects on disease recurrence and survival in clinical trials. Statistics in Medicine 2:41–58

Keiding N, Andersen PK (1989) Nonparametric estimation of transition intensities and transition probabilities: A case study of a two-state Markov process. Applied Statistics 38:319–329

Kleinbaum DG, Kupper LL, Morgenstern H (1982) Epidemiologic research: Principles and quantitative methods. Lifetime Learning Publications, Belmont

Kooperberg C, Petitti DB (1991) Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study. Epidemiology 2:363–366

Korn EL, Dorey FJ (1992) Applications of crude incidence curves. Statistics in Medicine 11:813–829

Kuritz SJ, Landis JR (1987) Attributable risk estimation from matched-pairs case-control data. American Journal of Epidemiology 125:324–328

Kuritz SJ, Landis JR (1988a) Summary attributable risk estimation from unmatched case-control data. Statistics in Medicine 7:507–517

Kuritz SJ, Landis JR (1988b) Attributable risk estimation from matched case-control data. Biometrics 44:355–367

Lagakos SW, Mosteller F (1986) Assigned shares in compensation for radiation-related cancers (with discussion). Risk Analysis 6:345–380

Laird N, Oliver D (1981) Covariance analysis of censored survival data using log-linear analysis techniques. Journal of the American Statistical Association 76:231–240

Land M, Gefeller O (1997) A game-theoretic approach to partitioning attributable risks in epidemiology. Biometrical Journal 39:777–792

Land M, Vogel C, Gefeller O (2001) Partitioning methods for multifactorial risk attribution. Statistical Methods in Medical Research 10:217–230

Landis JR, Heyman ER, Koch GG (1978) Average partial association in three-way contingency tables: A review and discussion of alternative tests. International Statistical Review 46:237–254

Landis JR, Sharp TJ, Kuritz SJ, Koch G (2000) Mantel–Haenszel methods. In: Gail MH, Benichou J (eds) Encyclopedia of epidemiologic methods, Wiley, Chichester, pp 499–512

Langholz B, Borgan Ø (1997) Estimation of absolute risk from nested case-control data. Biometrics 53:767–774

Last JM (1983) A dictionary of epidemiology. Oxford University Press, New York

Leung HM, Kupper LL (1981) Comparison of confidence intervals for attributable risk. Biometrics 37:293–302

Levin ML (1953) The occurrence of lung cancer in man. Acta Unio Internationalis contra Cancrum 9:531–541

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Liddell JC, McDonald JC, Thomas DC (1977) Methods of cohort analysis: Appraisal by application to asbestos mining (with discussion). Journal of the Royal Statistical Society, Series A 140:469–491

Lin DY, Psaty BM, Kronmal RA (1998) Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 54:948–963

Littell AS (1952) Estimation of the t-year survival rate from follow-up studies over a limited period of time. Human Biology 24:87–116

Llorca J, Delgado-Rodriguez M (2000) A comparison of several procedures to estimate the confidence interval for attributable risk in case-control studies. Statistics in Medicine 19:1089–1099

Lubin JH, Boice JD Jr (1989) Estimating Rn-induced lung cancer in the United States. Health Physics 57:417–427

Lui KJ (2001a) Interval estimation of the attributable risk in case-control studies with matched pairs. Journal of Epidemiology and Community Health 55:885–890

Lui KJ (2001b) Notes on interval estimation of the attributable risk in cross-sectional sampling. Statistics in Medicine 20:1797–1809

Lui KJ (2003) Interval estimation of the attributable risk for multiple exposure levels in case-control studies with confounders. Statistics in Medicine 22:2443–2557

Lynch HT, Lynch JF, Rubinstein WS (2001) Prophylactic mastectomy: obstacles and benefits (editorial). Journal of the National Cancer Institute 93:1586–1587

MacMahon B (1962) Prenatal X-ray exposure and childhood cancer. Journal of the National Cancer Institute 28:1173–1191

MacMahon B, Pugh TF (1970) Epidemiology: Principles and methods. Little, Brown and Co, Boston

Madigan MP, Ziegler RG, Benichou J, Byrne C, Hoover RN (1995) Proportion of breast cancer cases in the United States explained by well-established risk factors. Journal of the National Cancer Institute 87:1681–1685

Mantel N (1973) Synthetic retrospective studies and related topics. Biometrics 29:479–486

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 22:719–748

Markush RE (1977) Levin's attributable risk statistic for analytic studies and vital statistics. American Journal of Epidemiology 105:401–406

Matthews DE (1988) Likelihood-based confidence intervals for functions of many parameters. Biometrika 75:139–144

Mausner JS, Bahn AK (1974) Epidemiology: An introductory text. W.B. Saunders, Philadelphia

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. CRC Press, Boca Raton

McElduff P, Attia J, Ewald B, Cockburn J, Heller R (2002) Estimating the contribution of individual risk factors to disease in a person with more than one risk factor. Journal of Clinical Epidemiology 55:588–592

Mehta CR, Patel R, Senchaudhuri P (2000) Efficient Monte Carlo methods for conditional logistic regression. Journal of the American Statistical Association 95:99–108

Mezzetti M, Ferraroni M, Decarli A, La Vecchia C, Benichou J (1996) Software for attributable risk and confidence interval estimation in case-control studies. Computers and Biomedical Research 29:63–75

Miettinen OS (1972) Components of the crude risk ratio. American Journal of Epidemiology 96:168–172

Miettinen OS (1974) Proportion of disease caused or prevented by a given exposure, trait or intervention. American Journal of Epidemiology 99:325–332

Miettinen OS (1976) Estimability and estimation in case-referent studies. American Journal of Epidemiology 103:226–235

Morgenstern H (1982) Uses of ecologic analysis in epidemiological research. American Journal of Public Health 72:1336–1344

Morgenstern H, Bursic ES (1982) A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. Journal of Community Health 7:292–309

Morgenstern H, Kleinbaum D, Kupper LL (1980) Measures of disease incidence used in epidemiologic research. International Journal of Epidemiology 9:97–104

Neuhaus JM, Jewell NP (1990) The effect of retrospective sampling on binary regression models for clustered data. Biometrics 46:977–990

Neutra RR, Drolette ME (1978) Estimating exposure-specific disease rates from case-control studies using Bayes' theorem. American Journal of Epidemiology 108:214–222

Oakes D (1981) Survival times: Aspects of partial likelihood (with discussion). International Statistical Review 49:235–264

Ouellet BL, Romeder JM, Lance JM (1979) Premature mortality attributable to smoking and hazardous drinking in Canada. American Journal of Epidemiology 109:451–463

Palta M, Lin C-Y, Chao W (1997) Effect of confounding and other misspecification in models for longitudinal data. In: Modeling longitudinal and spatially correlated data. Lecture Notes in Statistics Series 122. Proceeding of the Nantucket Conference on Longitudinal and Correlated Data. Springer-Verlag, Heidelberg, New York, pp 77–88

Palta M, Lin C-Y (1999) Latent variables, measurement error and methods for analyzing longitudinal binary and ordinal data. Statistics in Medicine 18:385–396

Palta M (2003) Quantitative methods in population health: Extensions of ordinary regression. John Wiley & Sons, Hoboken, New Jersey

Prentice RL, Breslow NE (1978) Retrospective studies and failure time models. Biometrika 65:153–158

Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE (1978) The analysis of failure times in the presence of competing risks. Biometrics 34:541–554

Prentice RL, Pyke R (1979) Logistic disease incidence models and case-control studies. Biometrika 66:403–411

Rao CR (1965) Linear statistical inference and its application. John Wiley, New York, pp 319–322

Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, Mariotto A, Feuer EJ, Edwards BK (eds) (2004) SEER Cancer Statistics Review, 1975–2001, National Cancer Institute. Bethesda, MD. (http://seer.cancer.gov/csr/1975_2001) Accessed May 21, 2004

Robins JM, Greenland S (1989) Estimability and estimation of excess and etiologic fractions. Statistics in Medicine 8:845–859

Rockhill B, Newman B, Weinberg C (1998) Use and misuse of population attributable fractions. American Journal of Public Health 88:15–21

Rockhill B, Weinberg C, Newman B (1998) Population attributable fraction estimation for established breast cancer risk factors: considering the issues of high prevalence and unmodifyability. American Journal of Epidemiology 147:826–833

Rothman KJ, Greenland S (1998) Modern epidemiology. Lippincott-Raven, Philadelphia.

SAS Institute Inc. (1999) SAS/STAT user's guide. Version 8. SAS Institute Inc, Cary, NC

Schlesselman JJ (1982) Case-control studies. Design, conduct and analysis. Oxford University Press, New York

Seiler FA (1986) Attributable risk, probability of causation, assigned shares, and uncertainty. Environment International 12:635–641

Seiler FA, Scott BR (1986) Attributable risk, probability of causation, assigned shares, and uncertainty. Environment International 12:635–641

Siemiatycki J, Wacholder S, Dewar R, Cardis E, Greenwood C, Richardson L (1988) Degree of confounding bias related to smoking, ethnic group and SES in estimates of the associations between occupation and cancer. J Occup Med 30:617–625

Smith L (1998) Person-years of life lost. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics. Wiley, Chichester, pp 3324–3325

Spiegelman D, Colditz GA, Hunter D, Hetrzmark E (1994) Validation of the Gail et al model for predicting individual breast cancer risk. Journal of the National Cancer Institute 86:600–607

Sturmans F, Mulder PGH, Walkenburg HA (1977) Estimation of the possible effect of interventive measures in the area of ischemic heart diseases by the attributable risk percentage. American Journal of Epidemiology 105:281–289

Tarone RE (1981) On summary estimators of relative risk. Journal of Chronic Diseases 34:463–468

Tsiatis AA (1981) A large-sample study of Cox's regression model. Annals of Statistics 9:93–108

Tuyns AJ, Pequignot G, Jensen OM (1977) Le cancer de l'œsophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac. Bulletin of Cancer 64:45–60

US National Cancer Institute (2004) Breast cancer risk assessment tool. An interactive tool to measure a woman's risk of invasive breast cancer. (http://bcra.nci.nih.gov/brc) Accessed May 12, 2004

Uter W, Pfahlberg A (1999) The concept of attributable risk in epidemiological practice. Biometrical Journal 41:985–999

Uter W, Pfahlberg A (2001) The application of methods to quantify attributable risk in medical practice. Statistical Methods in Medical Research 10:231–237

Wacholder S, Benichou J, Heineman EF, Hartge P, Hoover RN (1994) Attributable risk: Advantages of a broad definition of exposure. American Journal of Epidemiology 140:303–309

Wahrendorf J (1987) An estimate of the proportion of colo-rectal and stomach cancers which might be prevented by certain changes in dietary habits. International Journal of Cancer 40:625–628

Walter SD (1975) The distribution of Levin's measure of attributable risk. Biometrika 62:371–374

Walter SD (1976) The estimation and interpretation of attributable risk in health research. Biometrics 32:829–849

Walter SD (1980) Prevention for multifactorial diseases. American Journal of Epidemiology 112:409–416

Walter SD (1983) Effects of interaction, confounding and observational error on attributable risk estimation. American Journal of Epidemiology 117:598–604

Whittemore AS (1982) Statistical methods for estimating attributable risk from retrospective data. Statistics in Medicine 1:229–243

Whittemore AS (1983) Estimating attributable risk from case-control studies. American Journal of Epidemiology 117:76–85

Wooldridge JM (2001) Econometric analysis of cross section and panel data. MIT Press, Boston

Woolf B (1955) On estimating the relationship between blood group and disease. Annals of Human Genetics 19:251–253

Wu K, Brown P (2003) Is low-dose Tamoxifen useful for the treatment and prevention of breast cancer (editorial)? Journal of the National Cancer Institute 95:766–767

Zhang J, Yu KF (1998) What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 280:1690–1691