7. Relating risk factors to health outcomes

Quantifying relationships between two factors or one factor and the occurrence, presence, severity, or course of disease

The "Big Picture"

At this point in the course, it will be good to take stock of where we are and where we are going. After a brief overview of population and health, we have thoughtfully considered the phenomenon of disease in relation to how epidemiologists study disease. Under that topic we examined issues of definition, classification, and natural history. We then turned to the question of how to measure disease frequency and extent in populations. We examined some general issues in numeracy and descriptive statistics, and then took up the fundamental epidemiologic measures of prevalence and incidence, with the latter approached as a proportion or as a rate. From there we took up the topic of standardization, which facilitates comparisons between prevalence and incidence across populations with different demographic composition, and we saw how these various measures and concepts are used in descriptive epidemiology and surveillance.

For the next section of the course we will be concerned with how to investigate associations between health outcomes and potential risk factors. That task involves questions of study design, measures of association, validity, inference and interpretation. The topics of study design and measures of association are so intertwined that whichever one we begin with, it always seems that we should have begun with the other! Analytic studies provide the data for estimating measures of association and impact, but measures of association and impact motivate the design of the studies.

However, the basic epidemiologic approach to relating risk factors to health outcomes is more general than the specifics of either topic. Consider a population in which a disease or some other condition occurs throughout the population but more often in persons with characteristic A. We are likely to be interested in how the existence (prevalence) or occurrence (incidence) of the disease among people with characteristic A compares with that for the population as a whole and for people with some other characteristic B (which could simply be the absence of A). To make this comparison we:

a. Measure the frequency - prevalence, CI, ID - of the disease or condition in each group (and perhaps in the entire population);
b. Compare the frequencies (fairly! - e.g., after standardization if necessary))
c. Quantify the comparison with a measure of association
d. Quantify the potential impact of the characteristic on the condition, if we are willing to posit a causal relationship.

We have already discussed measures of frequency and extent. Now we turn to measures of association and impact.

Measuring the strength of a relationship

The question that summarized the preceding topic could be stated as "How much of a factor is there?" or "How often does a disease (or other phenomenon) occur?". However, much of epidemiology is concerned with relationships among factors, particularly with the effect of an "exposure" on "a disease". Therefore the present topic addresses the question "How strong is the relationship between two factors?" or "How strong is the relationship between a study factor and an outcome?" A relationship may be "strong" without being "causal", and vice versa. Nevertheless, two factors that are strongly associated are more likely to be causally related.

There are a number of ways in which the strength of the relationship between two variables can be assessed. We can, for example, assess the extent to which a change in one variable is accompanied by a change in the other variable or, equivalently, the extent to which the distribution of one variable differs according to the value of the other variable. For this assessment, epidemiologists use a measure of association*.

A second perspective is the extent to which the level of one of the factors might account for the value of the second factor, as in the question of how much of a disease is attributable to a factor that influences its occurrence. Epidemiologists use measures of impact to address this question.

Most of the measures we will cover in this topic apply to relationships between a factor that is dichotomous (binary, having two possible values) and a measure of frequency or extent, in particular, a rate, risk, or odds. Such measures are the most commonly used in epidemiology. We will also touch on measures that are used in other situations.

Measures of association

A measure of association provides an index of how strongly two factors under study vary in concert. The more tightly they are so linked, the more evidence that they are causally related to each other (though not necessarily that one causes the other, since they might both be caused by a third factor).

Association - two factors are associated when the distribution of one is different for some value of the other. To say that two factors are associated means, essentially, that knowing the value of one variable implies a different distribution of the other. Consider the following two (hypothetical) tables:

* Although this term and "measure of effect" have frequently been used interchangeably (e.g., in this text), Rothman and Greenland (2000:58-59) draw the following distinction: associations involve comparisons between groups or populations; effects involve comparisons of the same population [hypothetically] observed in two different conditions; measures of association are typically used to estimate measures of effect.

CHD and oral contraceptives in women age 35 years or more

	OC	No OC	Total
CHD	30	20	50
Non-case	30	70	100
Total	60	90	150

Breast cancer (BC) and oral contraceptives

	OC	No OC	Total
Cancer	15	35	50
Non-case	30	70	100
Total	45	105	150

Consider first the table on the left (CHD and OC). The overall proportion of OC users is 60/150 = 0.40, but that among CHD cases is 30/50 = 0.60 while that among noncases is 30/100 = 0.30. The distribution of values of OC use ("users", "nonusers") is therefore different for different CHD values ("case", "noncase"). Similarly, the distribution of values of CHD is different for different values of OC (30/60 of OC users have CHD; 20/90 of non-users of OC have CHD).

If asked to estimate the proportion of OC users in a sample of 40 women selected at random from the table on the left, would we want to know how many in the sample had CHD and how many did not? Indeed we would.

We know that the proportion of OC users must be no lower than 0.30 (if the sample consists entirely of noncases) and no greater than 0.60 (if the sample consists entirely of cases). In the absence of knowing the proportion of cases, our best estimate would be the overall proportion in the population, 0.40. But is we knew the proportion of cases in the sample, we could move our estimate up (if more than one-third were cases) or down (if fewer than one-third were cases). [Now, verify that to estimate the proportion of CHD cases in a random sample, we would want to know the proportion of OC users. What is the best estimate of the proportion with CHD if the sample consists of 22 OC users and 18 nonusers? – The answer is at the end of this chapter.]

Thus, in the data in the left-hand table, there is an association between OC use and CHD. In contrast, in the table on the right (BC and OC), the distribution of OC use is the same for the cases, the noncases, and the entire group. Therefore, the data in the right-hand table show no association between breast cancer and use of OC's.

Correlation and Agreement

Association is a general term that encompasses many types of relationships. Other terms are used to indicate specific types of association. Two important ones are:

Correlation is a type of association in which the relationship is monotonic, i.e., it goes in one direction - the more of one factor, the more of the other (positive or direct correlation), OR the more of one factor, the less of the other (negative or inverse correlation). Linear correlation (measured by the Pearson product-moment correlation coefficient) assesses the extent to which the relationship can be summarized by a straight line. Nonparametric correlation coefficients, such as the Spearman rank correlation coefficient, assess the extent to which the two factors are correlated

but without regard to the size of the change in one that accompanies a change in the other, simply the direction.

Agreement is a type of correlation in which the two factors (generally two measures of the same phenomenon) are not only directly correlated with each other but have the same actual values. For example, two sphygmomanometers should give the same readings when used on the same person on the same occasion, not merely readings that are correlated. Two measurements of a stable phenomenon should agree with each other, not merely correlate. If one of the measures is known to be highly accurate and the other is being assessed, then we can assess validity of the latter, rather than merely agreement between the two.

ASIDE

Some sociological commentary

Since the factors studied by epidemiologists are often the occurrence of disease and the presence of exposure, the primary epidemiologic measures are proportions and rates of disease across different exposure groups. Indeed, because these measures are so familiar to epidemiologists and clinicians, even when the disease (e.g., blood pressure) and/or exposure are not represented by dichotomous (two-category) variables, it is common convert them into proportions or rates for at least some analyses. We will therefore spend most of our time on measures of association and impact involving rates and proportions. Bear in mind, though, that phenomena (e.g., physiologic measurements, nutrient intake, environmental exposures) that are capable of being measured as quantities are often more properly analyzed without dichotomizing.

The preference for rates and proportions is one reason for the different approaches to statistical analysis used by epidemiologists and social scientists who also study data on populations. But there are other differences in approach that presumably have a different basis, perhaps epidemiologists' focus on biological relationships.

One potential source of confusion – even conflict! – is the difference in the way that epidemiologists on the one hand and social scientists and biostatisticians look at associations. Epidemiologists tend to regard the strength of an association as a separate matter from the quantity of numerical evidence that the association would not easily arise by chance (i.e., its "statistical significance"). Other professions, however, often look first to the statistical significance of an association before considering any other characteristic. Thus, a biostatistician or psychologist might complete dismiss an association that an epidemiologist might characterize as "strong though potentially due to chance". Conversely, a psychologist or biostatistician may characterize as "highly significant" an association that an epidemiologist might dismiss as too weak to be biologically meaningful. As we will see later, various measures of association used in statistics (e.g., chi-squared statistics, correlation coefficients) are in a different category than the measures of association we will discuss now.

END OF ASIDE

Some basic measures

Before diving in to our discussion of how to measure associations, we may wish to begin with some basics. Suppose that an epidemiologist is asked to investigate the possible hazard from an inadequate air filtration system in a large school building in a poor urban neighborhood. The particular concern involves children with asthma, 400 of whom attend the school (school A). The epidemiologist is informed that on a particular day, 12 children suffered an asthmatic attack, whereas at a very similar nearby school (school B) with 500 asthmatic children, only 5 suffered an asthmatic attack on the same day.

The epidemiologist first arranges the data in a 2 × 2 table:

Cumulative incidence of asthmatic attack during one school day

Had an asthma attack	School A	School B	Total
Yes	12	5	17
No	388	495	883
Total	400	500	900

The first step is to compute the incidence in each school:

1-day cumulative incidence in school A: 12 cases / 400 children at risk = 0.03 or 3%

1-day cumulative incidence in school B: 5 cases / 500 children at risk = 0.01 or 1%

School A does in fact have a higher incidence of asthma attacks on the study day.

In order to assess the strength of the association between school and asthma incidence, the next step is to compute a measure of strength of association. The most common measure computed in this situation is the ratio of the two cumulative incidences (the "cumulative incidence ratio", CIR, also called the "risk ratio"). The CIR is simply 0.03/0.01 = 3.0, which is often interpreted as indicating a "moderately strong" association. The epidemiologist cumulative incidence difference (CID) might also compute the difference between the CI's (a "cumulative incidence difference", CID), and report that having inadequate air filtration was associated with a two percentage point greater asthma incidence during the 7-hour school day. Armed with this basic example, let us examine the concepts that underlie these measures.

Absolute versus relative effects

When we have incidence rates or proportions from two different populations (e.g., PC-users and Mac-users), it is easy to tell which rate is larger. But quantifying how much larger raises the question of how to compare the two rates. A basic question is whether or not the amount by which the larger rate exceeds the smaller one should be relative to the size of one of the rates.

If you ask a 10-year old how much older she is than her 5-year old brother, she will probably answer "5 years". But if she is mathematically-inclined, she may say that she is "twice his age" or "100% older". Both statements accurately quantify the amount by which she is older, yet they have different "flavors". Do we have a reason to prefer one or the other?

We might be inclined to prefer the answer "5 years". "Might", because the choice of a measure depends on our purpose, and we have not specified an objective. But two reasons come to mind why we might prefer the absolute difference (5 years) to the relative difference (100% older) or ratio (twice his age).

For one, "5 years" will remain accurate indefinitely, whereas "twice" (or "100% more") are accurate only this year. In that sense "5 years" provides a better summary of the relation between the children's respective ages. For another, human growth and aging, at least from a societal point of view and perhaps from a biological point of view as well, are processes which are marked by absolute increases, not relative ones. For example, we generally think of school entrance and graduation, puberty, eligibility for a drivers' license, presbyopia, and retirement in terms of specific age ranges, not proportional increases. We say "in 15 years you will probably need bifocals", rather than "when your age is 50% greater". In contrast, when adjusting a recipe for a larger or smaller number of guests, we multiply or divide the amounts of each ingredient by a common factor, rather than subtract a common amount from each one. For scaling a recipe, we are interested in proportionate (relative) increases.

Similarly, when we quantify the comparison of two incidences (or two prevalences), we can take the absolute difference (incidence difference) or the relative difference (excess risk). Which one, absolute or relative, is of greater interest to us in quantifying the comparison of two measures of occurrence or extent? This question has inspired no small amount of debate in the early days of modern epidemiology (ca. 1955) and, as so often happens, a case can be made for both approaches. The choice depends on our objective, our concept of the phenomena, and the availability of data.

One problem with using the absolute difference (variously called "risk difference", "rate difference", "cumulative incidence difference", "incidence density difference", "attributable risk", according to fashion, the group of epidemiologists with which the epidemiologist wishes to identify him/herself, the decade in which she/he learned epidemiology, or whether the comparison involves incidence rates, incidence proportions, prevalences, or mortality rates) as a measure of strength of association is that if the incidences themselves are small, as will always be the case for a rare disease, then the difference must also be small. For example, if the annual mortality rate for a rare disease such as esophageal cancer is 60/100,000 in persons with low vitamin C intake and 20/100,000 in persons with high vitamin C intake, the difference is only 40/100,000. In contrast, the difference for an association involving a more common disease, such as vitamin E and CHD, might be 1,200/100,000 for low vitamin E intake and 800/100,000 for high vitamin E intake = 400/100,000, an order of magnitude greater.

The much greater size of the second difference indicates that if these two vitamins are causal factors many more lives could be saved from increasing vitamin E intake than from increasing vitamin C intake. Vitamin E appears to have a greater public health impact. But is it logical to conclude from the greater difference for vitamin E that its association with CHD is stronger than vitamin C's with

esophageal cancer? First, if we did draw that conclusion it would imply that nearly any association involving a common disease must be stronger than all associations involving very rare diseases. Second, since the actual incidence of most conditions varies by all sorts of factors (age, gender, economic resources, smoking, alcohol intake, physical activity, diet, genetics, cofactors), the absolute difference is very likely to vary, possibly greatly, across populations (however, the relative difference may also vary).

In contrast, expressing the incidence differences relative to the size of the actual incidences produces measures of association that appear to be comparable. Thus we can compute a relative difference in incidence of esophageal cancer mortality in relation to vitamin C as $(I_1 - I_0)/I_0 = (0.00060 - 0.00020)/0.00020 = 2.0$ and a relative difference for CHD mortality in relation to vitamin E as $(I_1 - I_0)/I_0 = (0.01200 - 0.00800 / 0.00800) = 0.50$. On this basis, the association involving vitamin C is substantially greater than that involving vitamin E. This relative difference measure is often called the excess risk (or "excess rate", since the data are rates, not proportions). If we add 1.0 to the excess risk or rate, we obtain an even simpler relative measure, I_1/I_0, which is variously termed relative risk, risk ratio, rate ratio, cumulative incidence ratio, incidence density ratio, or, for prevalences, prevalence ratio.

Relative versus Absolute Measures of Association

Here are two real-life examples that contrast relative and absolute measures of association. The first is based on data from a follow-up study by Mann *et al.* (presented in a seminar at UNC-CH by Bruce Stadel):

Incidence of myocardial infarction (MI) in oral contraceptive (OC) users per 100,000 women-years, by age and smoking

Age (years)	Cigarettes/day	Oral contraceptive users	Non-users	RR**	AR***
30-39	0-14	6	2	3	4
	15 +	30	11	3	19
40-44	0-14	47	12	4	35
	15 +	246	61	4	185

Notes:

* RR=relative risk (rate ratio)

** AR=attributable risk (rate difference, absolute difference)

In this table, the incidence of MI is clearly greater for OC users, since in each age-smoking stratum the OC users have a higher incidence (ID) than do the nonusers. Moreover, the ratio of the two incidences (the RR) is nearly constant across strata, a desirable property for a summary measure, whereas the rate difference (AR) varies widely. According to Breslow and Day, the rate ratio tends

to be more stable across strata, supporting its desirability as a measure of association. Not all quantitative epidemiologists agree with this assertion.

The second example comes from a follow-up study of lung cancer and coronary artery disease in relation to cigarette smoking:

Mortality rates per 100,000 person-years from lung cancer and coronary artery disease for smokers and nonsmokers of cigarettes

	Smokers	Nonsmokers	Ratio	Difference
Cancer of the lung	48.3	4.5	10.8	44
Coronary artery disease	294.7	169.5	1.7	125

Source: 1964 *Surgeon General's Report on Smoking and Health*, page 110, quoted in Joseph Fleiss, *Statistical methods for rates and proportions*, 2nd edition, page 91

The rate ratio for the relation between smoking and lung cancer mortality is much larger than that between smoking and coronary artery disease mortality, but the rate difference is much larger for coronary artery disease mortality. These figures are usually interpreted to mean that lung cancer mortality is more closely associated with cigarette smoking than is coronary artery disease mortality; elimination of cigarette smoking would lead to a proportionate reduction in lung cancer mortality greater than the proportionate reduction in coronary artery disease mortality. However, the reduction in the number of deaths from lung cancer would be smaller in magnitude than the reduction in deaths from coronary artery disease. These issues will be explored in detail in the section Measures of Impact, later in this chapter.

Concept of relative risk

Nevertheless, for the most part we use relative risk as the basic measure of strength of association between a characteristic and the development of a condition.

The concept of relative risk is operationalized by :

a. Cumulative incidence ratio (CIR), also called risk ratio
b. Incidence density ratio (IDR), also called rate ratio
c. Odds ratio (OR), which estimates CIR and IDR under certain circumstances.

General formula:

$$\text{Incidence ratio} = \frac{\text{Incidence in "exposed"}}{\text{Incidence in "unexposed"}} = \frac{I_1}{I_0}$$

You may recall from the chapter on standardization that the SMR can be thought of as a ratio of "observed" to "expected" mortality rates. In fact, the concept of observed and expected can be brought in here as well. When we contrast the incidence rates in exposed and unexposed groups, we are typically using the unexposed incidence as a barometer of what incidence we might find in the exposed group if exposure had no effect. In that sense, the incidence in the unexposed constitutes an "expected", while the incidence in the exposed group constitutes an "observed".

The concept of relative risk can also be applied in situations where incidence estimates are unavailable or not even of greatest interest. For example, a direct estimate of the incidence ratio can be obtained in a case-control study with incident (newly-occurring) cases if the controls are selected in a suitable manner (as explained in the chapter on Analytic Study Designs). In situations where we want to estimate incidence ratios but only prevalence data are available, the prevalence ratio (PR) or prevalence odds ratio (POR) may provide a solution. The reason is the relation among prevalence, incidence, and duration, presented in the chapter on Measuring Disease and Exposure (in a stationary population, prevalence odds = incidence × average duration, or for a rare outcome, prevalence ≈ incidence × average duration). A key question is whether duration is the same in all groups being compared, since if it is not then the comparison of prevalences will provide a distorted picture of a comparison of incidences.

The PR may also be a logical choice for quantifying associations between exposures and conditions whose duration is as or more important than their incidence. For example, a large proportion of a population experience emotions or conditions such as anxiety, fatigue, or unhappiness from time to time. Since point prevalence will count mostly people in whom the condition persists, prevalence may be as or more useful than incidence as a measure of frequency in such cases. (The PR is also the straightforward choice for simple descriptive statements, such as "smoking was twice as common among persons with less than a high school education".)

Interpretation of relative risk

Example: Incidence ratio of 2.0 means that:

- "The incidence in the exposed population is twice that in the unexposed population"
- "The exposure is associated with a 100% increase in incidence."
- "The exposure is associated with a two-fold greater incidence." (although commonly encountered, this rendition should probably be avoided since "two-fold greater" might also be interpreted as 200% greater, which corresponds to an incidence ratio of 3.0)

Descriptive adjectives for magnitude of association (as commonly used)

1.0	No association (null value)
1.1-1.3	Weak
1.4-1.7	Modest
1.8-3.0	Moderate
3-8	Strong

For inverse associations (incidence ratio is less than 1.0), take the reciprocal and look in above table, e.g., reciprocal of 0.5 is 2.0, which corresponds to a "moderate" association.

Two-by-two tables

The most basic data layout in epidemiology is the two-by-two table:

	Exposure			
Disease	Yes	No	Total	
Yes	a	b	m_1	(a + b)
No	c	d	m_2	(c + d)
Total	n_1	n_0	n	
	(a + c)	(b + d)		

One major epidemiologic controversy is whether the disease should be shown in the rows, as above, or in the columns. Kleinbaum, Kupper, and Morgenstern use the above format. Hennekens and Buring place the disease categories in the columns and the exposure in the rows. Some authors use one presentation for cohort studies and the other for case-control studies. As you can see, epidemiology is not yet really a discipline (or not yet disciplined).

The above form of the 2 × 2 table is used to present data from a study (e.g., cohort, cross-sectional, case-control) with count data. When the study uses person-years data (e.g., to estimate incidence density), then the "no disease" column is removed and person-time totals (PY_1, PY_0) occupy the right-hand marginal:

	Exposure		
Disease	Yes	No	Total
Yes	a	b	m_1
Person-time	PY_1	PY_0	PY

Armed with our tables (whatever their orientation), we will now define the three major relative risk measures, about which there is much less controversy:

1. Cumulative incidence ratio (CIR)
2. Incidence density ratio (IDR)
3. Odds ratio (OR)

Cumulative incidence ratio (also called "risk ratio" or "relative risk")

The cumulative incidence ratio (CIR) addresses the question "by how many times does the risk in exposed persons exceed that for unexposed persons?" If the CIR is 3, we can say that exposed persons have 3 times the risk of unexposed persons. We can also say that the average exposed individual has three times the risk of disease as the average unexposed individual. This is often just what we want to know. The mathematical definition is:

$$\text{Cumulative incidence ratio} = \frac{\text{Cumulative incidence in "exposed"}}{\text{Cumulative incidence in "unexposed"}} = \frac{CI_1}{CI_0}$$

Since the CIR is based on estimates of CI or risk, the CIR can be estimated directly only from a cohort study. It is, however, possible to estimate it indirectly in other situations.

Incidence density ratio (also called "rate ratio")

The incidence density ratio (IDR) addresses the question "how many times does the rate of disease in exposed persons exceed that in unexposed persons?". If the IDR is 3 we can say the the rate in the exposed is 3 times that in the unexposed. There is not an obvious interpretation at the individual level, but the IDR is of prime importance for studies of dynamic populations and lengthy cohorts. The mathematical definition is:

$$\text{Incidence density ratio} = \frac{\text{Incidence density in "exposed"}}{\text{Incidence density in "unexposed"}} = \frac{ID_1}{ID_0}$$

The IDR is used in situations where the outcome is the length of time until an event (e.g., death) occurs and is mathematically equivalent to the hazard ratio of survivorship analysis. The IDR can be estimated directly in a follow-up study (of a fixed cohort or a dynamic population).

(Risk) odds ratio

The odds ratio (OR) is a ratio of "odds", which are transformations of risks or probabilities.

odds = p/(1-p), where p = probability

The OR addresses the question "how many times greater is the odds of disease for exposed persons than for unexposed persons?" Since odds have a different scale of measurement than risk, the

answer to this question can sometimes differ from the answer to the corresponding question about risk. Often, however, we are concerned with rare diseases, for which risk and odds are very close and CIR's and OR's (and IDR's) are very close. Since the OR can be defined in terms of odds of disease among exposed or odds of exposure among cases, there are two mathematical formulations:

$$\text{Odds ratio} = \frac{\text{Odds in "exposed"}}{\text{Odds in "unexposed"}}$$

The odds is simply an algebraic transformation of probability, so any probability (which must, of course, be less than 1.0) can be expressed as "odds". The probability that something may happen, especially something bad, is often referred to as a "risk". Odds derived from a risk are termed, appropriately, risk odds, so that a ratio of two risk odds is a ***risk odds ratio***, or ***ROR***.

(Exposure) odds ratio

A prevalence is commonly referred to as an estimate of probability (e.g., of exposure). A justification for this usage is that if we were to select an individual at random from the group, the probability that that individual would have a certain characteristic is estimated by the prevalence in the group. Odds that correspond to the probability of exposure are called "exposure odds", so their ratio is an ***exposure odds ratio***, or ***EOR***. Although conceptually distinct, for a two-by-two table these two odds ratios are algebraically identical, as we shall see. Thus, our ability to estimate an (exposure) odds ratio in a situation where we do not know disease incidence is a powerful tool for examining *associations* involving disease incidence even where we do not have incidence data, as was first presented in a classic paper by Jerome Cornfield (see the chapter on Analytic Study Designs for elaboration).

$$OR_r = \text{Risk odds ratio} = \frac{\text{Risk odds in "exposed"}}{\text{Risk odds in "unexposed"}} = \frac{odds_1}{odds_0} = \frac{CI_1 / (1\text{-}CI_1)}{CI_0 / (1\text{-}CI_0)}$$

$$OR_e = \text{Exposure odds ratio} = \frac{\text{Exposure odds in "cases"}}{\text{Exposure in "noncases"}} = \frac{odds_1}{odds_0}$$

Relation of the odds ratio to the risk ratio

When incidences are small (i.e., the outcome under study is rare in the population), the odds ratio closely approximates both the risk ratio and the incidence density ratio. The conventional guideline for classifying a disease as "rare" is an incidence below 10%. A good way to assess the extent of divergence of the odds ratio and risk ratio is to examine a spreadsheet with sample incidences and computed relative risks and odds ratios (e.g., the guideline suggested by Zhang and Yu [1998] of incidence below 10% and risk ratio below 2.5 allows the odds ratio to be only 20% greater than the risk ratio).

If one feels that the OR exaggerates the strength of association objectionably, it is a simple matter to derive a corresponding risk ratio estimate **if** one has additional information – overall exposure prevalence, overall disease incidence, disease incidence in the exposed, or disease incidence in the unexposed (Hogue, Gaylor, and Schulz, 1983). The simplest conversion is available if one knows the incidence in the unexposed group, e.g.:

$$RR = \frac{OR}{(1 - CI_0) + (CI_0 \times OR)}$$

where CI_0 is the incidence in the unexposed group [Zhang and Yu (1998), adapted to the notation used here]. A prevalence odds ratio can be converted into a prevalence ratio by substituting prevalence in the unexposed in place of CI_0 in the above formula. The divergence between the OR and the IDR will generally be less than that between the OR and the CIR. The reason is that all three measures of incidence (ID, CI, odds) have the identical numerator (new cases), but as incidence increases the denominators of ID and odds decrease, whereas the denominator for CI does not change.

Ratios of proportions versus ratios of odds

In case-control studies without additional information, the OR is often the only measure of association that can be estimated. Also, when the outcome is rare, all three measures of relative risk – the OR, CIR, and IDR – have approximately the same value. In other situations (i.e., cohort or cross-sectional data with non-rare outcomes), the appropriateness of the OR as a epidemiologic measure of association has been the subject of considerable debate.

Proponents of the OR point to several desirable mathematical properties it has compared to the risk ratio, including the fact that the strength of association is not affected by reversing the definition of the outcome (Walter, 2000). For example, in a smoking cessation trial, the OR for success will be the reciprocal of the odds ratio for failure; the "risk" ratio (CIR) for success, however, will be very different from the CIR for failure. Also, the prevalence odds ratio (POR) can in principle be used to estimate the the incidence rate ratio from cross-sectional data, assuming that disease duration is unrelated to exposure and that the incidences and durations in exposed and unexposed groups have been constant long enough to achieve a steady state condition. Moreover the popularity of multiple logistic regression, which estimates the OR controlling for multiple variables (see chapter on Data Analysis and Interpretation), has been a strong motivation for many investigators to estimate odds ratios even in cohort studies where incidence can be estimated directly.

As software tools for estimating the CIR and the PR have become available (e.g., SAS PROC GENMOD), however, the use of the odds ratio in cohort and cross-sectional studies is becoming less accepted, especially for non-rare outcomes (Thompson, Myers, and Kriebel, 1997). Its value in cross-sectional data is somewhat undercut by the difficulty of accepting that the stationary population (steady-state) assumption holds.

Critics have termed the OR "incomprehensible" (Lee, 1994:201) and as lacking "intelligibility" (Lee and Chia, 1994). Indeed, after a controversy erupted about news reports of a study by Kevin Schulman (Schulman *et al.*, 1999), the editors of the *New England Journal of Medicine* apologized for having allowed he use of the OR in the study's abstract (*New Engl J Med* 1999;341:287). One follow-up report in Brillscontent.com quoted one of the study's authors (Jesse Berlin, professor of biostatistics at the University of Pennsylvania School of Medicine) as saying "Unless you're a professional statistician, you're not likely to have the slightest clue what an odds ratio means. The truth is, it's confusing for a lot of people, including physicians."

In the Schulman *et al.* study, primary care physicians attending professional meetings viewed videotaped interviews of hypothetical patients (portrayed by actors) and received additional medical data, and then indicated whether or not they would refer the patient for cardiac catheterization. A central finding was that the physicians recommended catheterization for 84.7% of the presentations when the actor was an African American compared to 90.6% of the presentations when the actor was a European American. The finding was presented as an OR of 0.6, which was then reported by the news media as indicating that black patients were "40 percent less likely" to be referred as were white patients (see Table 2 in *Schwartz et al.*, 1999 for a summary of news reports).

Schwartz *et al.* (1999) explained that because the outcome was so common, the actual risk ratio (0.93, indicating a weak association) was greatly overstated by the OR, which contributed to the media's overstatement of the association. However, the risk ratio for **not** being referred is also 0.6 (0.09/0.15), indicating that white patients were only 60% as likely not to be referred as were black patients or that black patients were 60% more likely not to be referred as were white patients (RR of 1.6 = 1/0.6). So whether the impression given by the news media was exaggerated or not is debatable, at least with respect to the OR (see Schwartz *et al.* for other limitations in the study).

Greenland (1987) asserts that the OR's relevance for epidemiology derives solely from its ability to estimate of the rate ratio (IDR) or cumulative incidence ratio (CIR). His objection to the OR as a measure of effect lies in the lack of a simple correspondence between the odds for a population and the odds for an individual. Whereas "incidence proportion" (i.e., CI) is equivalent to a simple average of the risk for each individual in the population and incidence density (ID) is equivalent to a simple average of the "hazard" for each individual in the population, incidence odds is not equivalent to a simple average of the disease odds for each individual in the population (Greenland, 1987). Thus, the OR is not a ratio of averages interpretable at the individual level. It turns out that this property ("noncollapsibility") of the OR can make its use misleading when one attempts to examine an association with control for other factors (see chapter on Data Analysis and Interpretation).

Although one can take refuge in the assertion that "qualitative judgments based on interpretating odds ratios as though they were relative risks are unlikely to be seriously in error" (Davies, Crombie, and Tavakoli, 1998:991), it is safer to avoid the OR when incidence or prevalence ratios can be estimated.

Two typically unstated assumptions

Stable exposure status

The above discussion assumes that the population being studied is reasonably stable in respect to exposure status. When this is not the case it may be necessary to change individuals' exposure status during the observation period, assigning their follow-up to one or another exposure group, if the exposure effect is believed not to persist. For example, a subject may exercise, stop, and begin again. If the effect of exercise is believed to terminate shortly after exercise is stopped and to begin again shortly after resumption of exercise, then follow-up time (person-time) can be accumulated in the appropriate exercise category for each part of the follow-up period of an incidence density measure. (An alternative approach is to place such "switchers" in a category of their own.)

Absence of "contagion"

The above discussion also assumes that exposure and outcome are independent, i.e., one person's disease does not affect another person's risk. This assumption is violated, of course, for contagious diseases, such as sexually transmitted infections, and for arthropod-borne pathogens, e.g. malaria, where humans serve as a reservoir. Here, the spread of disease increases the exposure of unaffected individuals so that their risk increases. These so-called "dependent happenings" can result in distortion, or at least marked variability over time, in the above measures of association (see, for example, Koopman JS *et al.*, 1991). Dependent happenings are by no means confined to communicable diseases, inasmuch as personal and community behaviors are frequently affected by what other people and communities are doing. Some examples are smoking cessation, dietary change, suicide attempts, driving behavior, road safety regulations, and intensity of disease detection and reporting.

More on risk and relative risk

The ***excess risk*** gives the proportionate increase in incidence (an analogous measure can be constructed using incidence density or odds). It is a slight modification of the CIR and useful in a variety of circumstances including measures of relative impact, to be discussed shortly. The algebraic definition is:

$$\text{Excess risk} = \text{CIR} - 1 = \frac{CI_1}{CI_0} - 1 = \frac{CI_1 - CI_0}{CI_0}$$

For diseases with an extended risk period, as duration of follow-up increases, risk and CI become larger. Being cumulative and predicated on the population remaining at risk, CI is an increasing function whose limit is 1.0 – if we remain at risk forever, then eventually we will all become cases. As CI_1 and CI_0 both increase towards their limit of 1.0, then the CIR also approaches 1.0. Therefore the value of the CIR can change as the duration of follow-up lengthens. It is also possible for the IDR to change with duration of follow-up, but that is a function of the natural history of the disease rather than the the IDR's mathematical properties.

When the CI is low, due to a rare disease and/or short follow-up period:

$$CI \approx ID \times T \quad \text{(where T = follow-up time)}$$

$$OR \approx IDR \approx CIR$$

because if CI is $\approx ID \times T$, then $CI_1 = ID_1 \times T$ and $ID_0 = ID_0 \times T$, so:

$$CIR \approx \frac{ID_1 \times T}{ID_0 \times T} = \frac{ID_1}{ID_0} = IDR$$

As follow-up time becomes shorter, then CI becomes smaller, eventually reaching 0. But as the CI becomes smaller its value becomes increasingly the same as $ID \times T$. For this reason, the limit of the CIR as the follow-time becomes vanishingly short ($T \rightarrow 0$) is the IDR. For this reason the IDR is sometimes referred to as the "instantaneous CIR".

In a steady-state (constant size and age distribution, constant incidence density, prevalence, and duration of disease) dynamic population:

Prevalence odds = Incidence × Duration (see previous chapter)

From this we can see that the prevalence odds ratio (POR) estimates the IDR if duration is unrelated to exposure, because:

$$POR = \frac{odds_1}{odds_0} = \frac{ID_1 \times T}{ID_0 \times T} = \frac{ID_1}{ID_0} = IDR$$

where T here is duration in exposed and unexposed cases.

Estimating relative risk (via the odds ratio) from data from a case-control study

1. Construct (2x2, four-fold) table

	Exposure			
Disease	Yes	No	Total	
Yes	a	b	m_1	(a + b)
No	c	d	m_2	(c + d)
Total	n_1 (a + c)	n_0 (b + d)	n	

2. Odds of Exposure in cases

$$\text{Odds} = \frac{\text{Proportion of cases who are exposed}}{\text{Proportion of cases who are unexposed}} = \frac{a / (a + b)}{b / (a + b)} = \frac{a}{b}$$

3. Odds of exposure in controls

$$\text{Odds} = \frac{\text{Proportion of controls who are exposed}}{\text{Proportion of controls who are unexposed}} = \frac{c / (c + d)}{d / (c + d)} = \frac{c}{d}$$

4. Exposure odds ratio (OR_e)

$$OR_e = \frac{\text{Odds of exposure in cases}}{\text{Odds of exposure in controls}} = \frac{a / b}{c / d} = \frac{ad}{bc}$$

If the data had come from a cross-sectional or cohort study, we could instead have estimated the risk odds ratio (OR_r), as the odds of disease in exposed persons divided by odds of disease in unexposed persons. Algebraically, the exposure and disease odds ratios are identical.

Note that the odds ratio can be computed from proportions or percentages as readily as from the actual numbers, since in computing the odds ratio the first step (see above) is to convert the numbers into proportions and then to convert the proportions into odds.

Difference measures

Measures based on the difference between two proportions or rates are the other principal form of comparison for rates and proportions. They are often used as measures of impact, as we will discuss in the next section. The formulas and terms for differences of cumulative incidences (or risks) and incidence rates are:

$$CID = CI_1 - CI_0$$ ("***Cumulative incidence difference***",

also known as the "Risk difference" or "Attributable risk")

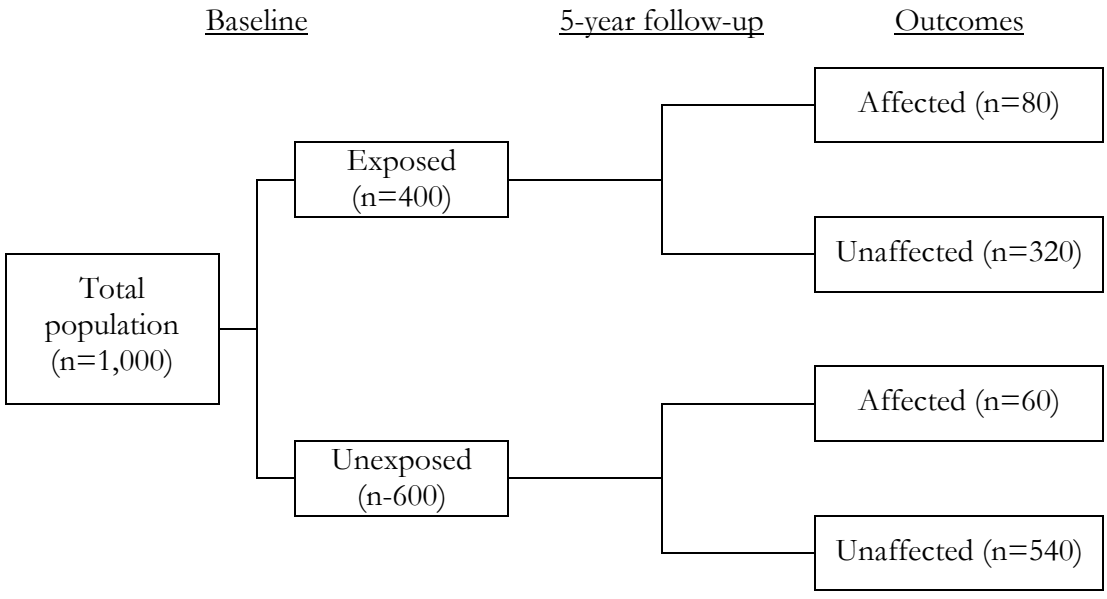
$$IDD = ID_1 - ID_0$$ ("***Incidence density difference***",

also known as the "Rate difference")

These difference measures, of course, can be derived directly only from a cohort or follow-up study. If we lack information on the size of the population at risk, as in a case-control study with no additional information, we have no way to estimate either CI or ID, so we cannot estimate risk or rate differences. In a cross-sectional study, we cannot estimate incidence at all, though by analogy with CID and IDD we can estimate the ***prevalence difference***, $P_1 - P_0$.

Examples of computations

Follow-up of a fixed cohort



(Assume no losses to follow-up, including deaths from other causes.)

The above data are often summarized into a 2 × 2 table:

5 - year incidence of disease

	Exposure		
Disease	Yes	No	Total
Yes	80	60	140
No	320	540	860
Total	400	600	1000

Note on terminology: The four numbers (80, 60, 320, 540) in the interior of the table are referred to as the "cells"; the row and column totals (140, 860, 400, 600) are referred to as the "marginals". The cells are often referred to as "a", "b", "c", and "d" in zig-zag fashion beginning with the upper left cell.

CI (crude) = 140 / 1000 = .14 (i.e., the overall 5-year cumulative incidence was 14/100)

CI_1 = 80 / 400 = .20, CI_0 = 60 / 600 = .10

CIR = CI_1 / CI_0 = .20 / .10 = 2.0 (the exposure was associated with a doubling of risk)

CID = CI_1 - CI_0 = .20 - .10 = .10 (see below for interpretation)

Excess risk = CIR - 1 = 1.0 (i.e., the exposure was associated with a 100% increase in risk)

$$OR_r = \frac{CI_1 / (1 - CI_1)}{CI_0 / (1 - CI_0)} = \frac{0.20 / 0.80}{0.10 / 0.90} = \frac{0.25}{0.11} = 2.25$$

Note that the OR is more extreme than the CIR.

Average incidence density measures could be computed from the above table by making the assumption that cases occurred evenly throughout the period, or equivalently, that all cases occurred at the midpoint of the follow-up period, 2.5 years:

$$ID = \frac{\text{cases}}{\text{person-years}} = \frac{140}{(860)(5) + (140)(2.5)} = \frac{140}{4300 + 350} = 0.030 \text{ cases/py}$$

(Total person-years at risk comprises 860 persons followed for 5 years and 140 persons followed for 2.5 years – once the disease occurred, that subject was deemed no longer at risk. If that situation were not the case, then person-years would be computed differently.)

$ID_1 = 80 / [(320)(5) + (80)(2.5)] = 80 / 1800 = 0.044$ cases / person-year

$ID_0 = 60 / [(540)(5) + (60)(2.5)] = 60 / 2850 = 0.021$ cases / person-year

IDR = 0.044 / 0.021 = 2.095 = 2.1 (compare to CIR of 2.0)

IDD = 0.044 - 0.021 = .023 cases / person-yr OR 23 cases / 1000 person-yrs

Note that each ID is very close to the corresponding CI divided by the number of years (5). When the incidence is low, the CI approximately equals ID × (time interval).

Measures of association – non-dichotomous exposure

Ratio measures of association are suited to dichotomous (i.e., two-category) measures, such as presence of disease (yes or no) or exposure (yes or no). If the exposure has multiple categories (for example, different types of industrial solvents or several levels of exposure), a ratio measure of effect can be computed for each type or level compared to the unexposed group (if there is no unexposed group, then one exposure or level can be selected as a reference category). Consider, for example, the classic study by Wynder and Graham (1950) on lung cancer and cigarette smoking. In this case, "None (less than 1 per day)" is selected as the reference category, and the odds ratio is computed for each higher level of smoking relative to the reference level.

Cigarette smoking histories of 605 male lung cancer patients and 780 controls

Amount of cigarette smoking for 20+ years.* (percent distribution)	Lung cancer patients [N=605]	Controls [N=780]	OR
None (less than 1 per day)	1.3	14.6	1.0**
Light (1-9 per day)	2.3	11.5	2.2
Moderately heavy (10-15 per day)	10.1	19.0	6.0
Heavy (16-20 per day)	35.2	35.6	11.1
Excessive (21-34 per day)	30.9	11.5	30.2
Chain (35+ per day)	20.3	7.6	30.0

* includes pipe and cigar smokers, with a conversion formula.

** reference category.

The odds ratios (OR) are obtained by forming a 2x2 table for each exposure level relative to the reference level. For example, for "Heavy (16-20 per day)" compared to "None":

	Lung cancer	Control
Heavy	35.2	35.6
None	1.3	14.6

$$OR = \frac{35.2 \times 14.6}{35.6 \times 1.3} = 11.1$$

(As stated earlier, the OR calculation can be done just as easily from percentages as from the actual numbers of cases and controls, since the first step is to derive proportions from the numbers. The fact that these percentages are age-adjusted actually means that the ORs are age-adjusted as well.)

The odds ratios reveal the existence of a marked dose-response relationship.

Measures of association – non-dichotomous disease

When the disease or outcome variable is not dichotomous (e.g., body mass index) but the exposure is, the outcome variable can be categorized (e.g., "above or below 30% greater than ideal weight") to enable computation of ratio measures of association. Alternatively, a summary statistic (e.g., mean body mass) can be computed for each category of the exposure, but then we have no measure that can be interpreted as relative risk.

When both disease and exposure have multiple ordered categories (e.g., injury severity rating with several levels (an ordinal variable), parity (a count variable), or blood pressure (a continuous measure), categorization can be imposed to obtain a ratio measure of effect. Alternatively, the relationship between outcome and exposure can be plotted, and the slope used as a measure of the strength of the relationship (e.g., a 2 mmHg increase in diastolic blood pressure for every 14 grams of alcohol consumed is stronger than a 1 mmHg increase for every 14 grams). Linear regression coefficients are used to estimate the slope of the relationship and provide a satisfactory index of strength of association for continuous variables, though one that cannot readily be compared to measures of relative risk. We will return to regression coefficients later in the course.

Correlation coefficients are often used as measures of association between ordinal or continuous variables, but as explained below, these are not regarded as epidemiologic measures of strength of association.

Other measures of association

> "When I use a word, it means precisely what I want it to, neither more nor less" (Lewis Carroll, Alice in Wonderland)

As mentioned earlier, a point of confusion for the learner is the difference between what epidemiologists mean by a measure of association and what is measured by various statistics that are also referred to as measures of association. To clarify this unsatisfactory state of affairs, we will discuss two measures that are widely used in both epidemiology and other disciplines, but which epidemiologists regard as very different from the measures of association we have discussed above.

Chi-square for association

A nearly ubiquitous statistic in epidemiology is the chi-square for association. The chi-square and its associated p-value address the question of the degree to which an association observed in a sample is likely to reflect an association in the population from which the sample was obtained, rather than simply have arisen due to sampling variability. The p-value estimates the probability that variability of random sampling can result in two variables being associated in a sample even if they are entirely independent in the population. Although there is obviously a connection between the question addressed by the chi-square and the question addressed by the relative risk, the two questions are by no means interchangeable. For example, consider the table at the very beginning of this chapter.

CHD and oral contraceptives (OC) in women age 35 years or more

	OC	$\overline{\text{OC}}$	Total
CHD	30	20	50
$\overline{\text{CHD}}$	30	70	100
Total	60	90	150

Regarding these data as having come from a hypothetical case-control study, we select the odds ratio (OR) as the appropriate measure of strength of association. Since CHD is a rare disease, the OR will estimate the CIR as well as the IDR. The OR for the above table is:

$$\text{OR} = \frac{30 \times 70}{20 \times 30} = 3.5$$

i.e., the observed association suggests that the risk of CHD in women 35 years or older who use OC is 3.5 times that of similarly aged women who do not use OC.

The chi-squared statistic for this table will yield a p-value that approximates the probability that a table with an OR of 3.5 or stronger will arise from a random draw of 50 women (who will be called "cases") from a population of 60 OC users and 90 nonusers. That chi-squared statistic is 12.4, which corresponds to a very small probability – much lower then 0.0001, or 1 in a thousand draws (the computation will be covered in a later part of the course). Suppose instead that the study that yielded the above table had been only one-fifth as large. Keeping the same proportion in each of the four cells, we would then have this table:

CHD and oral contraceptives (OC)in women age 35 years or more

	OC	$\overline{OC}$	Total
CHD	6	4	10
$\overline{CHD}$	6	14	20
Total	12	18	30

The odds ratio for this table is still 3.5, but the chi-squared statistic is now only 2.42, which corresponds to a p-value of 0.12. The greater p-value results from the fact that it is much easier to obtain an association with OR of 3.5 or greater by randomly drawing 10 "cases" from a room with 12 OC users and 18 nonusers than by randomly drawing 50 "cases" from a room with 60 OC users and 90 nonusers.

Since the OR remains identical but the chi-squared statistic and its p-value change dramatically, clearly the epidemiologic measure of association and the chi-square are measuring different features of the data. The chi-squared statistic is used to evaluate the degree of numerical evidence that the observed association was not a chance finding. The epidemiologic measure of association is used to quantify the strength of association as evidence of a causal relationship.

Correlation coefficients

Correlation coefficients are measures of linear or monotonic associations, but again not in the same sense as measures of relative risk. The linear correlation coefficient (Pearson or product-moment correlation, usually abbreviated "r") measures the degree to which the association between two variables is linear. An r of zero means that the two variables are not at all linearly related (they may nevertheless be associated in some other fashion, e.g., a U-shaped relationship). An r of +1 or -1 means that every pair of observations of the two variables corresponds to a point on a straight line drawn on ordinary graph paper. However, knowing whether or not the relationship is linear tells us nothing about the steepness of the line, e.g., how much increase in blood pressure results from a 5% increase in body mass. Other correlation coefficients (e.g., Spearman) measure the degree to which a relationship is monotonic (i.e., the two variables covary, without regard to whether the pairs of observations correspond to a straight line or a curve).

Epidemiologists think of the relationships between variables as indications of mechanistic processes, so for an epidemiologist, strength of association means how large a change in risk or some other outcome results from a given absolute or relative change in an exposure. If the assumption is correct, the strength should not depend upon the range of exposures measured or other aspects of the distribution. In contrast, r is affected by the range and distribution of the two variables and therefore has no epidemiologic interpretation (Rothman, p.303). Standardized regression coefficients are also not recommended for epidemiologic analysis for similar reasons (see Greenland, Schlesselman, and Criqui, 1986).

Correlation coefficients between dichotomous variables — Correlation coefficients can be particularly problematic when used to quantify the relationship between two dichotomous (binary) factors, especially when one or both of them are rare. The reason is that correlation coefficients between binary variables cannot attain the theoretical minimum (-1) and maximum (+1) values except in the special case when the both factors are present half of the time and absent half of the time (Peduzzi, Peter N., Katherine M. Detre, Yick-Kwong Chan. Upper and lower bounds for correlations in 2 × 2 tables–revisited. *J Chron Dis* 1983;36:491-496). If one or both factors are rare, even if the two variables are very strongly related, the correlation coefficient may be restricted to a modest value. In such a case an apparently small correlation coefficient (e.g., 0.15) may actually be large in comparison with the maximum value obtainable for given marginal proportions.

For example, the correlation coefficient between smoking and lung cancer cannot be large when the proportion of lung cancer cases is small but that of smokers is large, as shown in the following example (Peduzzi PN, Detre KM, Chan YK. Upper and lower bounds for correlations in 2 × 2 tables–revisited. *J Chron Dis* 1983;36:491-496) based on data from Allegheny County, PA:

	Smoker	Nonsmoker	Total
Lung cancer	20	2	22
No lung cancer	14,550	9,576	24,126
Total	14,570	9,578	24,148
Lung cancer incidence			0.001
Smoking prevalence			0.60
Odds ratio			6.6
	Correlation coefficient (r)	R-square (R^2)	
Based on above data	0.019	0.00036	
If all cases were smokers	0.024	0.00058	
If no cases were smokers	-0.037	0.00157	

Here, the correlation coefficient (r) is a meagre 0.019, with a corresponding R^2 ("proportion of variance explained") of 0.000356. Even if all 22 lung cancer cases were smokers, the correlation coefficient would rise only to 0.024 (with $R^2 = 0.0006$), and if no lung cancer cases smoked r falls only to -0.037. In contrast, the OR is 6.6, indicating a strong relationship (the RR and IDR are essentially the same, since the outcome is so rare). Therefore the correlation coefficient and proportion of variance explained are not readily applicable to relationships between dichotomous variables, especially when the row or column totals are very different.

Measures of Impact

Concept

Relative risk measures compare the risk (or rate) in an exposed group to that in an unexposed group in a manner that assesses the strength of association between the exposure and outcome for the purpose of evaluating whether the association is a causal one, as we will see in the chapter on Causal Inference. But when we have decided (or assumed) that the exposure causes the outcome, we often wish to assess the individual and/or public health importance of a relationship, i.e.,

- How much of a disease can be attributed to a causative factor?
- What is the potential benefit from intervening to modify the factor?

The answers to these questions enter into public health policy-making and, in principle, individual decision-making, since they indicate the amount or proportion of the burden of a disease that can be prevented by eliminating the presumed causal factor (e.g., pollution control) or by carrying out a preventive intervention (e.g., fortification of foods). Examples of the kind of questions that prompt the use of measures of impact are:

1. Now that I am 35 years old, my CHD risk from taking oral contraceptives is twice as great as when I was 25. But *how much more risk* do I have due to taking the pill?
2. In HIV-discordant couples in which a condom is not used and one partner has a bacterial sexually transmitted disease, *how much of the risk* of heterosexual transmission of HIV is due to presence of the sexually transmitted disease and therefore might be eliminated through STD control measures?
3. *How many cases* of asthma are due to ambient sulfur dioxide?
4. *What proportion of motor vehicular deaths* can be prevented by mandatory seat belt use.
5. *What proportion of perinatal HIV transmission that has or would have occurred* has been prevented through the use of prenatal, intrapartum, and neonatal zidovudine?

To answer these questions we employ ***attributable fractions***, which are measures of ***impact*** or ***attributable risk***. The concept of attributable risk is of central importance for public health, since it addresses the question of "so what?". Although some students find the topic of attributable risk a source of confusion, at least some of their confusion is attributable (!) as much to the terminology as to the basic concept. There are, however, a number of subtleties and legitimate sources of confusion related to attributable risk. To introduce the concept we make the simplifying assumptions that the exposure in question has either adverse or beneficial effects but not both, that the exposed and unexposed groups are identical except for the exposure, and that no person is susceptible to getting the outcome from both the exposure and some other causal factor (e.g., people who would have become ill from the exposure might, if not exposed, have developed the outcome from another cause during the follow-up period, so the number of cases would be the same). We also begin by focusing on risks and proportions, rather than on rates.

One more prefatory note: at the risk of provoking a reaction of "Duh!", I will note that questions of attributable risk arise only in situations where more than one factor can cause the outcome under consideration. When the outcome has a necessary causal factor (which is always the case when the outcome is defined in terms of the etiologic agent, as with infectious diseases), all of the cases must be attributable to that factor. Eliminating the factor would avoid all risk. If a necessary cause ("C") requires a co-factor or susceptibility factor ("S") for the effect to occur, then all of the cases are attributable both to "C" and to "S". This last point also illustrates that attributable fractions do not sum to 1.0, even though they are often expressed as percentages.

Perspectives

There are a variety of different measures of impact, and at least twice that many names for them. (For example, the term "attributable risk" is sometimes used to refer to the risk difference, sometimes to the population attribute risk proportion described below, and sometimes to the class of measures of impact. See Greenland and Robins (1988) and Rothman and Greenland for various usages, with citations) One reason for the multiplicity of measures is simply to have a measure for each of the various ways to ask a question about impact. That is, the question can be asked in absolute ("How much" risk) or relative ("What proportion" of risk) terms. It can be asked with reference specifically to persons exposed to the factor or with reference to the whole population. Also, the factor being considered may cause or prevent the outcome. Various combinations of these alternatives call for different measures. The justification for having more names than measures (and for using the same name for different measures) is harder to divine.

Absolute perspective

The **absolute perspective** for attributable risk is expressed by the questions, "**How much** of the risk is attributable to the factor?" and "How many cases might be avoided if the factor were absent?" The answer is obtained by estimating the **risk difference** for exposed and unexposed persons. The risk difference provides an estimate of the **amount of risk in exposed persons** that is "attributable" to the factor (assuming causality and equivalent risk for exposed and unexposed groups from factors other than the exposure). If we are interested in the amount of risk that is attributable to the exposure **in the total population** (assuming causality), we multiply the risk difference by the **exposure prevalence** in the population. If we are interested in the **actual number of cases** that are attributable, i.e., that could have been avoided by complete elimination of the exposure (before any irreversible effects have occurred), we can multiply the risk difference by the population size.

Relative perspective

The **relative perspective** for attributable risk is expressed by the question, "**What proportion** of the risk is attributable to the factor?" and "What proportion of the cases of the disease might be avoided if the factor were absent?". Here, we need to express the amount of risk attributable to the factor relative to the total risk in exposed persons or in the total population. The measure for the exposed population is sometimes referred to as the "***attributable fraction***" (AF) or "***attributable risk proportion***" (ARP). Greenland and Robins (1988) explain the distinction between the "***excess fraction***" and the "***etiologic fraction***". The latter, which encompasses all cases in which the

exposure plays a causal role, cannot be estimated without strong biological assumptions (see Greenland and Robins for examples). So unless otherwise indicated, the attributable fractions discussed here are excess fractions, not etiologic fractions. The measure for the entire population is sometimes referred to as the "***population attributable fraction***" (PAF) or "population attributable risk proportion."

Attributable fraction

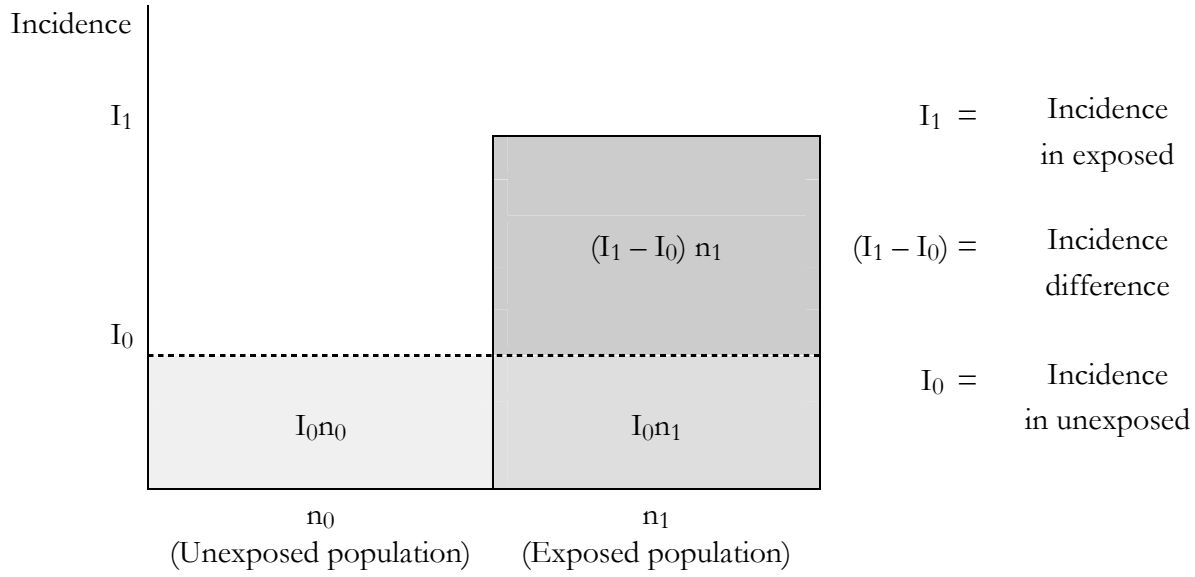
The AF is directly related to the strength of the association between the exposure and the disease – if the exposure doubles the risk, then half of the risk is attributable to the exposure; if the exposure triples the risk, then two-thirds of the risk is attributable to the exposure; if the exposure multiples the risk fourfold, then the AF is three-fourths, etc.

Population attributable fraction

The PAF reflects not only the strength of association but also the prevalence of the exposure in the population. Obviously an exposure can do more damage (have more impact) if many people are exposed to it. The PAF adds this consideration to the AF. Note that many older texts and articles refer to the PAF simply as "attributable risk".

The following diagram displays the basis for the various measures of attributable risk. The basic idea is, simply, that if we observe incidence I_1 in an exposed population and a lower incidence I_0 in a comparable unexposed population, and we make the assumption that the exposure is causing the higher incidence in the exposed population, then it is logical to suppose that the difference, $I_1 - I_0$, is the amount of incidence that is due to the exposure. Then, depending on the way in which we are asking the question, this "attributable incidence" is expressed as an absolute difference or as a relative difference, and in relation to exposed persons only or to the entire population.

Diagrammatic representation of attributable risk in a population



P_1 = Proportion exposed

In the above diagram:

n_0 and n_1 represent, respectively, the numbers unexposed and exposed persons, or the amounts of unexposed and exposed population-time; $n = n_0 + n_1$

P_0 and P_1 represent, respectively, the proportions of unexposed and exposed persons or population time (i.e., $P_1 = n_1/n$)

I_0 is the incidence proportion (cumulative incidence) of disease in unexposed persons, so I_0n_0 is the expected number of cases among unexposed persons, i.e., the area of the most lightly shaded rectangle.

I_1n_1 is, similarly, the expected number of cases among exposed persons, i.e., the combined area of the two more darkly-shaded rectangles.

$(I_1 - I_0)$ is the incidence difference or "attributable risk." $(I_1 - I_0)$ gives the amount of incidence that exposed persons experience over and above the incidence they would be expected to have had (I_0) in the absence of exposure. That excess risk is therefore "attributable" to exposure.

$(I_1–I_0)\ n_1$ is the expected number of cases among exposed persons beyond those expected from their background incidence (I_0), i.e., attributable cases (the area of the darkest rectangle). Attributable cases are simply the attributable risk multiplied by the number of exposed persons.

RR is the relative risk (risk ratio, CIR), I_1/I_0

The attributable fraction in exposed persons [AF] is the proportion of exposed cases that is "attributable" to the exposure. This proportion is:

$$AF = \frac{\text{"Attributable cases"}}{\text{All } \textit{exposed} \text{ cases}} = \frac{(I_1 - I_0)\, n_1}{I_1 n_1} = \frac{I_1 - I_0}{I_1} = \frac{RR - 1}{RR}$$

(the RR's are obtained by dividing numerator and denominator by I_0).

Similarly, the population attributable fraction [PAF], the proportion of all cases that is attributable to exposure, is:

$$PAF = \frac{\text{"Attributable cases"}}{\text{All cases}} = \frac{(I_1 - I_0)\, n_1}{I_1 n_1 + I_0 n_0} = \frac{I_1 n_1 - I_0 n_1}{I_1 n_1 + I_0 n_0} = \frac{P_1(RR-1)}{1 + P_1\,(RR-1)}$$

The right-hand formula (see the assignment solution for its derivation) displays the relationship of the PAF to exposure prevalence and "excess risk" (RR-1). The denominator cannot be less than 1, so if the numerator is very small (e.g., very low exposure prevalence and/or weak association), then the PAF will also be very small. Conversely, for a very prevalent exposure (e.g., P_1=0.80) and very strong association (e.g., RR=9), then the numerator [0.80 × (9-1)] will be large (6.4). The denominator will be close to this value, since the 1 will have little influence. Thus, the PAF will show that a large proportion (i.e., close to 1.0) of the cases are attributable. As the prevalence rises, the PAF comes closer to the AF (when P_1=1, as it does in the exposed population, the PAF formula reduces to that for the AF [(RR – 1)/RR]).

The joint influence of strength of association and prevalence of exposure on the PAF may be easier to see in the following algebraic reformulation:

$$PAF = \frac{1}{1 + 1/[P_1(RR-1)]}$$

Definitions and formulas

Attributable risk [absolute]: the amount of the risk in the exposed group that is related to their exposure. Attributable risk is estimated by the cumulative incidence difference or incidence density difference:

$$AR = I_1 - I_0$$

Population attributable risk [absolute]: the amount of risk in the population (i.e., in exposed and unexposed persons taken together) that is related to exposure. Population attributable risk is equal to the attributable risk multiplied by the prevalence of the exposure:

$$PAR = AR \times P_1 = (I_1 - I_0)\, P_1 = I - I_0$$

[This measure is not often used, but is helpful here to complete the pattern. "I" without a subscript refers to the total, or crude incidence. The equivalence of the middle and right-hand terms in the above expression can be seen by substituting $(I_1P_1 + I_0P_0)$ for I and $(I_0P_0 + I_0P_1)$ for I_0.]

Attributable fraction (often expressed as a percent) [AF]: the proportion (percent) of the risk in the exposed group that is related to their exposure.

$$AF = \frac{I_1 - I_0}{I_1} = \frac{RR - 1}{RR} = \frac{AR}{I_1}$$

Population attributable fraction (also often expressed as a percent) [PAF]: the proportion (percent) of the risk in the population that is related to the exposure.

$$PAF = \frac{P_1\,(RR - 1)}{1 + P_1\,(RR - 1)} = \frac{I - I_0}{I} = \frac{PAR}{I}$$

The following alternate formula for the PAF can be used with data available from most case-control studies and, when necessary, with an adjusted RR (e.g., an age-standardized RR) when it is necessary to control for differences between exposed and unexposed groups.

$$PAF = \frac{P_{E|D}\,(RR - 1)}{RR} = (P_{E|D}) \times AF$$

where PE|D is the proportion of cases who are exposed. The derivation of this expression is

$$PAF = \frac{\text{Attributable cases}}{\text{All cases}} = \frac{(\text{Exposed cases}) \times AF}{\text{All cases}} = (P_{E|D}) \times AF$$

The expression on the right can be estimated from a case-control study that provides an estimate of RR, since neither exposure prevalence in the population nor the actual disease incidence is needed.

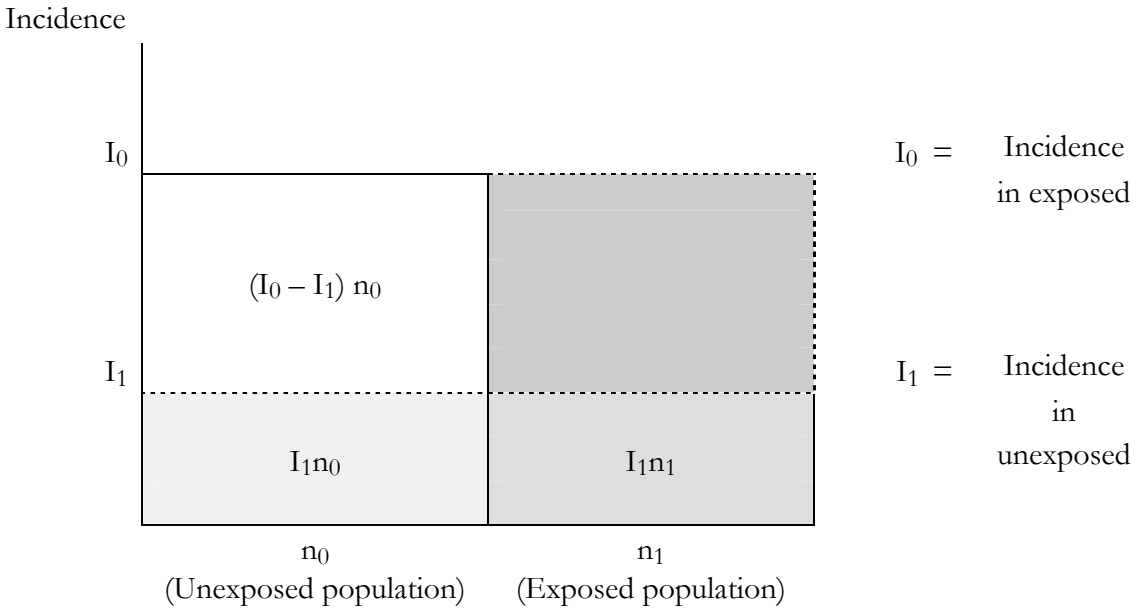
Preventive fraction

When $I_1 < I_0$ (e.g., for a vaccine, use of protective equipment, or pollution control devices), a "preventive" fraction is needed. Since a protective exposure (assuming causality) reduces the risk of

the outcome, we cannot think in terms of "attributable cases" since the "cases" have not occurred! Instead, we define a preventive fraction as a proportion of "potential cases" that were prevented, i.e., that did ***not*** occur because of the protective exposure. For vaccines, this proportion is referred to as ***vaccine efficacy*** or effectiveness.

As with attributable fractions, there are two variants, one for those exposed to the preventive intervention and one for the population as a whole (both are based on the "relative" perspective; the absolute perspective does not appear to be used). The following diagram, similar to that for attributable fractions, will be used.

Diagrammatic representation of preventive fraction in a population



P_1 = Proportion exposed

Where n_1, n_0, I_1, I_0 are as before and $(I_0 - I_1)\, n_1$ denotes the "prevented cases", i.e., the number of potential cases that would have occurred if the exposure were not associated with lower incidence (recall that I_0 is greater than I_1) or had not been present. $I_1 n_1$ are the cases that occurred in spite of the intervention.

Therefore, the preventive fraction in the exposed (PF_1) quantifies the prevented cases as a proportion of all potential cases in exposed persons. The preventive fraction in the population (PF) expresses the prevented cases as a proportion of all potential cases in the entire population. In each case, the "prevented cases" are cases that would have occurred but for the preventive exposure; the "potential cases" are prevented cases plus actual cases.

From the diagram:

Preventive fraction in exposed

(PF_1 - for those exposed to the preventive measure)

$$PF_1 = \frac{\text{"Prevented } \textit{potential} \text{ cases"}}{\text{All } \textit{potential exposed cases}} = \frac{(I_0 - I_1)\, n_1}{I_0 n_1} = \frac{(I_0 - I_1)}{I_0} = 1 - RR$$

(since $I_1 < I_0$, $RR < 1.0$).

Preventive fraction in the population (PF)

$$PF = \frac{\text{"Prevented } \textit{potential} \text{ cases"}}{\text{All } \textit{potential cases}} = \frac{(I_0 - I_1)\, n_1}{I_0\, n} = \frac{(I_0 - I_1)\, P_1}{I_0} = P_1\, PF_1$$

(recall that n_1/n is the proportion exposed, P_1).

The preventive fraction represents the proportion (or percent) of the potential burden of disease which was prevented by the protective factor. The following formula displays this aspect clearly:

$$PF = \frac{(I_0 - I_1)\, n_1}{I_0\, n} = \frac{(I_0 n_1 - I_1 n_1) + (-I_0 n_0 + I_0 n_0)}{I_0\, n} = \frac{(I_0 n_1 + I_0 n_0)}{I_0\, n} - \frac{(I_1 n_1 + I_0 n_0)}{I_0\, n} = \frac{I_0 - I}{I_0}$$

I_0 is the risk in people unexposed to the preventive measure. If no one received its benefits, then the risk in the entire population would also be I_0. The actual overall risk, I, represents an average of the risks for those exposed to the preventive measure and those not exposed, weighted by their respective sizes ($I_1 n_1 + I_0 n_0$). So ($I_0 - I$) is the difference between the risk that *could have been* observed and the risk that *was* observed, which difference is assumed to be attributable to effectiveness of the preventive measure and its dissemination. The last formula expresses this difference as a proportion of the risk in the absence of the preventive measure.

In all of these measures, of course, the assumption is made, at least for purposes of discussion, that the relationship is causal, and in some cases, that removing the cause (or introducing the preventive factor) is fully and immediately effective. In any specific example, of course, the latter assumption can be varied.

Unified approach to attributable risk and preventive fraction

Although there are many subtleties, the basic idea of attributable and prevented fractions is relatively straightforward. The following conceptualization brings out the underlying simplicity and may be the easiest way to derive formulas when needed.

The objective is to quantify the impact of an exposure or preventive measure in terms of the burden of a disease. Large impacts come from:

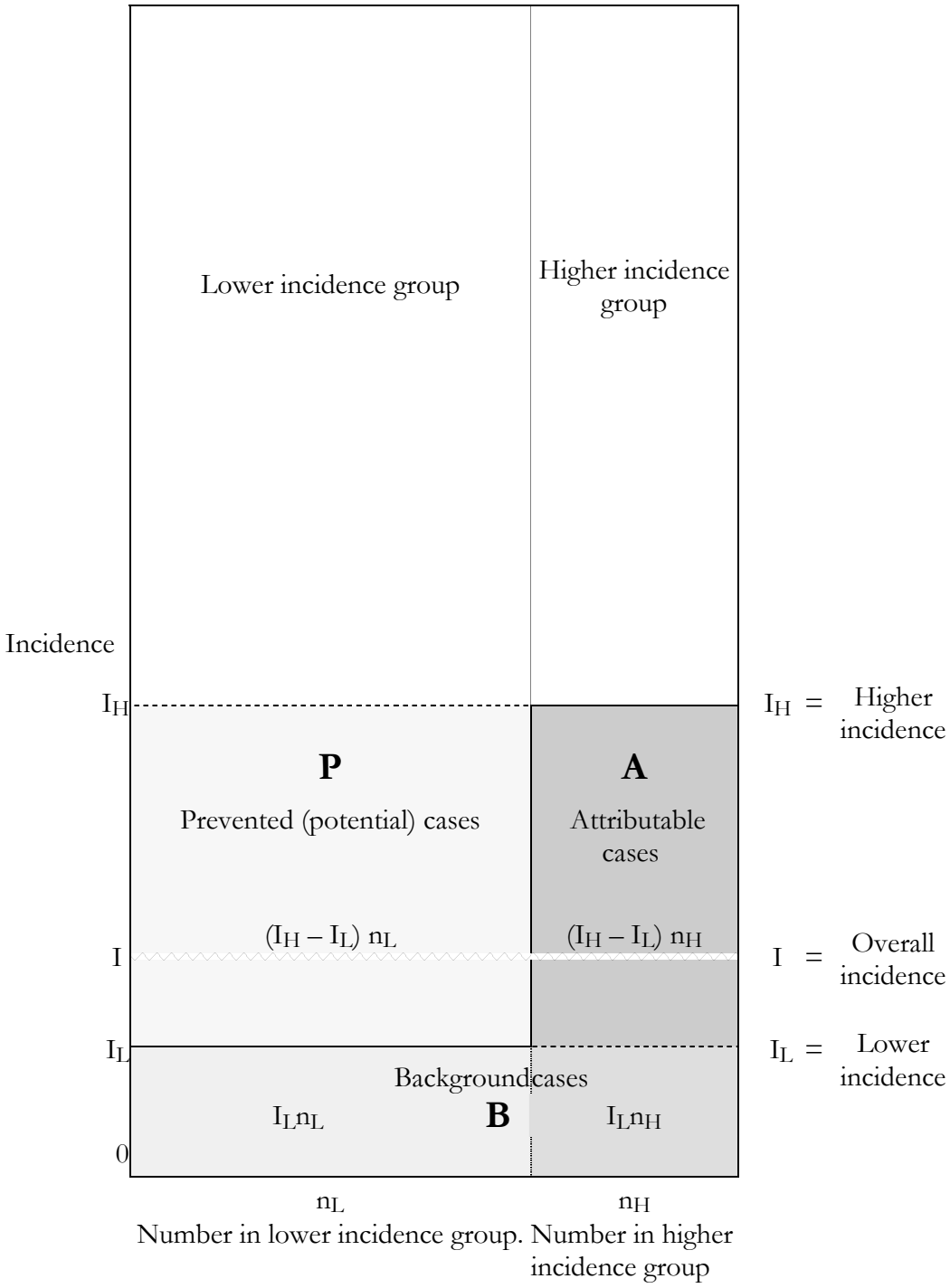
1. high frequency of disease
2. powerful risk or preventive factor
3. large proportion of people exposed to the factor

One aspect that complicates the formulas is the fact that incidence in people exposed to a risk factor is greater than in the unexposed, but incidence in people exposed to a preventive factor is lower than in the unexposed. We can side-step this difference by thinking in terms of the higher incidence and the lower incidence.

The diagram on the following page represents a population at risk in which people can be classified into two exposure groups, one with lower incidence (e.g., physically active) and the other with higher incidence (e.g., sedentary). The width of each tall rectangle indicates the number of people in the corresponding exposure group. Physical activity and sedentary lifestyle make a good example, because they will work as well for the risk factor (attributable fraction) perspective and the preventive factor (prevented fraction) perspective. Let us use I_L and I_H to represent the lower and higher incidence, and N_L and N_H to represent the number of people or amount of population time in the lower (physically active) and higher incidence (sedentary) categories, respectively.

In this diagram, rectangle **A** [$N_H (I_H - I_L)$] represents the **attributable** caseload. This is the caseload that would not have occurred were it not for the risk factor (or for the absence of the preventive factor). Rectangle **P** [$N_L (I_H - I_L)$] represents the **prevented** caseload. This caseload is only potential, since it has not occurred. These cases would have occurred had the preventive factor (physically activity) not been present (or if the risk factor – sedentary lifestyle – were to spread to the lower incidence group). Note that both the attributable caseload and the prevented caseload are based on $(I_H - I_L)$, the difference in the two incidences.

Diagrammatic representation attributable and prevented cases



n_1 = Number with the risk or preventive factor.

P_1 = Proportion with the risk or preventive factor.

Rectangle **B** [$N_L\ I_L + N_H\ I_L$] represents the unavoidable (background) caseload. This is the caseload that occurs in spite of the presence of the preventive factor and absence of the risk factor. The total observed caseload is represented by the sum of the rectangles for the two exposure groups [$N_L I_L + N_H I_H$]. If I is the overall (crude) incidence, then the total observed caseload can also be written as [$(N_L + N_H)\ I$]. The total potential caseload (i.e., observed plus prevented) corresponds to [$(N_L + N_H)\ I_H$], the result of subjecting the entire population to the higher incidence.

With this diagram and notation we can express both attributable and prevented fractions in a more parallel manner. The population attributable fraction (PAF) is simply $\mathbf{A}/(\mathbf{A} + \mathbf{B})$ and the prevented fraction (PF) is simply $\mathbf{P}/(\mathbf{A} + \mathbf{B} + \mathbf{P})$. We can therefore write the formulas we derived earlier as:

$$PAF = \frac{\text{“Attributable cases”}}{\text{All cases}} = \frac{\mathbf{A}}{\text{All cases}} = \frac{N_H\,(I_H - I_L)}{N_L\,I_L + N_H\,I_H} = \frac{P_H\,(I_H - I_L)}{I}$$

The last step made use of the facts that $N_H/(N_L + N_H)$ is the prevalence of the exposure and that the overall incidence I equals the total cases divided by the total population. Since the attributable fraction (AF) concerns only the exposed group, $P_H = 1$ and $AF = (I_H - I_L)/I_H$, which also equals $(RR - 1)/RR$.

Similarly, the prevented fraction is:

$$PF = \frac{\text{“Prevented cases”}}{\text{All } \textit{potential} \text{ cases}} = \frac{\mathbf{P}}{(N_L + N_H)\ I_H} = \frac{N_L\,(I_H - I_L)}{(N_L + N_H)\ I_H} = \frac{P_L\,(I_H - I_L)}{I_H}$$

If we divide numerator and denominator by I_H, we obtain $P_L\,(1 - RR)$. The prevented fraction in those exposed to the preventive intervention concerns only the lower incidence group, so $P_L = 1$ and $PF_1 = (1 - RR)$.

With this notation we can see the essential equivalence between the PAF and the PF. They both involve the risk difference times the number of people in the “exposed” group. They both are expressed as a proportion of the full caseload, except that for the PF we need to include the potential caseload in the denominator – otherwise it would not be a proportion.

Note that the attributable fraction for the group at higher risk and the prevented fraction for the group at lower risk are the same, since both fractions are simply the absolute difference in the two

risks, relative to the higher one. The attributable and prevented fractions for the population, however, depend upon the prevalence of the exposure as well as the risk for each group. If the prevalence of the exposure were 50%, so that equal proportions of the cohort were exposed to the higher and lower risks, the numerators of the attributable and prevented fractions would be equivalent (attributable caseload = prevented caseload). But since the denominator for the attributable fraction involves exposure prevalence but the denominator for the prevented fraction does not, the two fractions would still not be equal.

PAF in a case-control study

Case-control studies, as will be discussed in the following chapter, do not provide an estimate of incidence unless additional information is available. But case-control studies do yield estimates of the odds ratio (OR), which for a rare outcome approximates the RR. As long as the cases are representative of all cases of interest (e.g., in a defined target population), we can combine this RR estimate with the the prevalence of exposure in the cases to obtain the PAF. The prevalence of exposure in casee is simply the number of exposed cases ($N_H I_H$) divided by the total number of cases ($N_L I_L + N_H I_H$).

Note the similarity between this prevalence, $(N_H I_H)/(N_L I_L + N_H I_H)$, and the intermediate expression shown above for the PAF. The only difference is in the numerators: $N_H I_H$ versus $N_H (I_H - I_L)$. We can get from the exposure prevalence expression to the PAF by multiplying the former by $(I_H - I_L)/I_H$ or, equivalently, by $(RR - 1)/RR$, which we can estimate by $(OR - 1)/OR$ if the disease is rare. So we can estimate PARP for a rare disease from a case-control study that can measure neither incidences nor exposure prevalence by using OR to estimate RR in:

$$PAF = \frac{N_H (I_H - I_L)}{N_L I_L + N_H I_H} = \frac{P_{H|D} (RR - 1)}{RR} = P_{H|D} (AF)$$

This diagram and these formulas are worth becoming very familiar with, since doing so will help to develop an in-depth understanding of incidence, prevalence, relative risk, impact, and weighted averages and also to derive any of the basic attributable risk or preventive fraction formulas. Picture how the diagram will change as the prevalence of the exposure increases or decreases. How will the wavy line (the overall incidence, I) move as the other variables change? How is it related to I_L, I_H, N_L, N_H? (This is definitely a key relationship to know). What happens when everyone is exposed? When no one is exposed? What is the prevalence of exposure in cases? How will it change as incidence and/or overall exposure prevalence change?

Interpreting attributable fractions

Although the basic concept of attributable risk is intuitively meaningful, it turns out that it has many subtleties and nuances. My own appreciation of the subtleties is still developing, so that much of

what I have written (as some written by other authors) is not completely accurate. The confusion is aggravated by the multitude of terms that have been introduced, with usages that differ from one author to another, as well as the use of the same term to mean different things. Therefore, if you find yourself being confused by something you are reading in this area, always consider the possibility that what you are reading may be confused as well.

One big conceptual complication arises when we try to interpret attributable fractions (e.g., AF, PAF) in etiologic (causal) terms, which is of course what we were interested in doing. Consider the following two questions, which figure prominently in product liability litigation, where courts have held that recovery requires a finding that the plaintiff's disease was "more likely than not" a consequence of exposure to the product (e.g., asbestos, prescription drugs, silicone breast implants, tobacco).

- Among nonsmokers exposed to X, *what proportion of Y was caused by* X?
- What is the probability that person Z's case of Y *resulted from* X?

What distinguishes these two questions from the illustrative ones at the beginning of the section is the use of causal terminology ("caused by", "resulted from") instead of the more general (and vaguer) "attributed to". Incidence and measures derived from incidence show only overall, or net effects, not the causal processes that produce them. Even though, for example, a sedentary lifestyle increases the risk of coronary heart disease, physical exercise can acutely increase the risk of a cardiac event. When we compare the rate of cardiac events in a sedentary group to the rate in a physically active group, the difference in incidence rates measures the increased rate of cardiac events *associated with* a sedentary llifestyle. But if some of the incidence of cardiac events in exercisers actually *results from exercising*, then the difference in incidence between the two groups measures the *net* harm from a sedentary lifestyle, rather than the *total* effect. By comparing the incidence rates we are letting the cardiac events in the exercisers offset some of the events in the sedentary group, with the relative size of benefit and harm depending upon the kinds of people (e.g., genetic characteristics or distributions of other exposures) who are exercise and do not exercise. In general, epidemiologic data will not reveal what contributes to the net incidence difference.

Similarly, if the action of one causal factor can preempt the opportunity for another factor to cause the disease (because the disease has already occurred), then there is no way to know from epidemiologic data which factor caused the disease in a person or population exposed to both causal factors. For this reason, it is problematic to interpret attributable risk measures as ***etiologic fractions***, although many writers have used the terminology interchangeable (see Greenland and Robins, 1988 for an explanation with hypothetical examples). According to Greenland (1999: 1167), the "key fallacy in much of the literature and testimony regarding the probability of causation is the use of the following generally incorrect equations: Etiologic Fraction = Rate Fraction and Probability of Causation = Rate Fraction . . .", where the etiologic fraction (EF) is "the fraction of these individuals for whom exposure was a contributory cause of the disease" (Greenland, 1999: 1166) and the rate fraction (RF) is the incidence rate difference divided by the incidence rate in the exposed (analogous to the AF, except derived from incidence rates rather than incidence proportions) (p1167). In algebraic terms, $EF=(A_1+A_2)/A_T$, where A_1 are exposed persons who would have developed the disease at some point but whose disease process was accelerated due to the exposure, A_2 are exposed persons whose disease would never have occurred without the

exposure, and A_T is A_1+A_2 plus exposed persons who develop the disease completely independently of exposure. The EF estimates the ***probability of causation***, since $(A_1+A_2)/A_T$ is the probability that a person randomly selected from A_T had his/her disease accelerated by (A_1) or completely caused by (A_2) the exposure. The proportion A_2/A_T is the ***excess fraction***, since it gives the proportion of the total caseload that would not have occurred without the exposure (Greenland, 1999), regardless of time to occurrence. Greenland observes that the failure to distinguish the excess fraction from the etiologic fraction is a "major problem in most of the literature", and regards the term "attributable risk" as particularly misleading even though it "dominates the American literature", both in biostatistics and epidemiology [p.1168].

Adjusting a population attributable fraction for another risk factor

An adjusted population attributable fraction (PAF) can also be estimated from stratified data, which provides a way of controlling for differences between exposure groups in the distributions of a factor such as age. In crude data, the PAF can be estimated as:

$$PAF = \frac{\text{Attributable cases}}{\text{All cases}} = \frac{\text{Number of cases observed} - \text{Number of cases expected}}{\text{All cases}}$$

with the number of cases expected in the absence of exposure being estimated from the observed incidence in unexposed persons. With stratified data, the number of expected cases is estimated separately for each stratum and then summed across strata. The expected number of cases in each stratum is simply $I_{0i}\, n_i$, where I_{0i} is the incidence in unexposed persons in stratum i and n_i is the number of persons in stratum i, so the total number of expected cases is $\sum(I_{0i}\, n_i)$. Thus, the PAF for stratified data is:

$$PAF = \frac{\text{All cases} - \sum (I_{0i}\, n_i)}{\text{All cases}} = \frac{I - \sum (I_{0i}\, n_i / n)}{I}$$

where I is the overall (crude) incidence, obtained by dividing numerator and denominator by the total number of persons (n) summed across all strata. This formula can be found in Rockhill, Newman, and Weinberg (1998), though it appears in probability notation with an unfortunate typographical error (an extraneous overbar above the C).

A simpler formula, that can also be estimated from case-control data, is:

$$PAF = \frac{\text{Attributable cases}}{\text{All cases}} = \frac{\sum (\text{Attributable cases})_i}{\text{All cases}}$$

$$= \frac{\sum [\, c_i \times PAF_i \,]}{C} = \sum [\, (c_i/C)\, PAF_i \,]$$

where c_i is the number of cases in stratum i and C is the total number of cases. Thus, the overall PAF can be obtained as a weighted average of the stratum-specific PAFs, with the fraction of cases in each stratum as the weights (Hanley, 2001). Since the attributable cases in a stratum can also be written in terms of the attributable fraction (among exposed cases), the above expression can also be written in terms of stratum-specific attributable fractions:

$$= \frac{\sum [\, (\text{Exposed cases})_i \times AF_i \,]}{\text{All cases}} = \sum [\, (c_i/C)\, (P_{E|D})_i\, AF_i \,]$$

which is a weighted average using the same weights as above, with $(P_{E|D})_i$ representing the exposure prevalence among cases in stratum i. (Thanks to Beverly Rockhill Levine for comments on this section.)

Answer to question at beginning of the chapter about the association between CHD and OC:

The proportion of CHD cases in the sample of 40 must be somewhere between 30/60 = 0.5 (the proportion of cases among OC users) and 20/90 = 0.2222 (the proportion among nonusers). If the sample consists of 22 users and 18 nonusers, then the best estimate of the sample proportion of CHD cases is:

$$\text{Proportion with CHD} = 0.5\left(\frac{22}{40}\right) + 0.2222\left(\frac{18}{40}\right) = 0.5(0.55)+0.2222(0.45) = 0.375$$

Therefore, the best estimate of the overall proportion with CHD is approximately 0.375 or 15 women in the sample of 40.

Summary

There are three categories of measures: Frequency/extent, association, impact

(1) Measures of frequency or extent (especially prevalence and incidence)

In epidemiology, incidence is the occurrence of any new health-related event (e.g., disease, death, recovery). Incidence is quantified as a:

PROPORTION: the proportion of a population who experience the event; also called "RISK", since it estimates the average risk per person for the period. [Risk] ODDS are simply a transformation of risk [risk/(1-risk)].

RATE: the number of health events per person per unit time; corresponds to the average risk per person per unit time.

MEASURE	EPIDEMIOLOGIC ESTIMATOR	UNITS	LIMITS
Risk	Cumulative Incidence (CI)	Dimensionless	0 to 1
Rate	Incidence Density (ID)	1/time	0 to "infinity"
$Odds_r$	CI / (1-CI)	Dimensionless	0 to "infinity"

CI (a proportion) is used to estimate an individual's risk of developing a disease. ID (a rate) is used to estimate the force intensity of occurrences. Risk and rate are related, since the greater the intensity of occurrences in a population, the greater the risk of an event to any member of the population. When CI is small (i.e., because of a low the intensity of disease or a short time interval), ID is approximately equal to CI divided by the number of years of followup. When CI is not small, the relationship is more mathematically complex.

Application

The choice of an incidence measure (either CI or ID) depends upon:

a. OBJECTIVES OF THE STUDY

CI provides a direct estimate of an individual's risk, as may be useful for making clinical and/or personal decisions;

ID is often preferred for assessment of the population impact of a health event or for testing etiologic hypotheses.

b. PRACTICAL CONSIDERATIONS

CI may be preferred:

- if the health event has a restricted risk period
- if it is difficult to ascertain time of change in health status
- for ease of comprehension.

ID may be preferred:

- if the health event has an extended risk period
- if lengths of follow-up vary
- if there is a large loss to follow-up
- if the health event can recur (e.g., bone fractures).

A ratio of two risk estimates is a "risk ratio" (RR). A ratio of two rate estimates is a "rate ratio" (RR). A ratio of two odds is an "odds ratio" (OR). All these measures are sometimes referred to as "relative risk" (RR), though strictly speaking only the first pertains to risk.

(2) Measures of association

e.g. Ratios of proportions (CIR), ratios of rates (IDR), ratios of odds (OR_r and OR_e)

$$\text{CIR} = \frac{a / (a + c)}{b / (b + d)} = \frac{\text{CI in exposed}}{\text{CI in unexposed}} = \text{"risk ratio", "relative risk"}$$

where a=exposed cases, b=unexposed cases, c=exposed noncases, d=etc.

In probability terms, CIR = Pr(D|E) / Pr(D|E)

How do we interpret the CIR?

a. If CIR = 1, then no association between exposure and disease.
b. If CIR > 1 then exposure appears to be associated with increased risk of disease, i.e., exposure may be harmful.
c. If CIR < 1 then exposure appears to be associated with decreased risk of disease, i.e., exposure may be protective.

(CIR's less than 1.0 can be awkward to think about, so it many cases it is helpful to reverse the disease or exposure category to obtain the reciprocal of the CIR. A CIR of 0.4 then becomes a CIR of 2.5)

CIR can be directly estimated if the exposure status is known before the occurrence of disease, as in a prospective followup study or a retrospective cohort study.

When a disease is rare, the OR_r approximates the CIR – a useful thing to know because logistic regression models may be employed to estimate odds ratios:

$$OR_r = \frac{\text{Odds of disease in exposed}}{\text{Odds of disease in unexposed}} = CIR$$

The OR (whether the "risk OR" or "exposure OR") is easy to calculate as the cross-product ratio: $(a \times d) / (b \times c)$.

The risk and exposure OR's are calculated identically from a 2×2 table, but that doesn't mean they are equivalent "epidemiologically". Remember that the numbers in the 2×2 table are only an abstraction from the actual study experience and must be used with the design in mind (i.e., a case-control design is not equivalent to a longitudinal design). In a cohort study, we typically compute a CIR or an average IDR. In a follow-up study without a fixed cohort, we typically compute an IDR. In a case-control study we typically compute an OR. In a cross-sectional study, we typically compute a prevalence ratio or a prevalence OR. If one of the cumulative incidences is known, the OR estimate (e.g., from a logistic regression model – see chapter on Data Analysis and Interpretation) can be converted to a risk ratio estimate by the following formula (Zhang and Yu, 1998; notation changed to match that used in this chapter):

$$RR = \frac{OR}{(1 - CI_0) + (CI_0 \times OR)}$$

A prevalence ratio can be estimated from a prevalence odds ratio in the same manner, if the prevalence in the unexposed is known.

3) Measures of impact

"How much" of a disease can be attributed to an exposure can be considered as:

- an amount of the risk or incidence in the exposed (CID) or in the total population (usually presented as a number of cases)
- a proportion of the risk or incidence in the exposed (AF) or in the total population (PAF).

The contributors to impact measures are:

1. Strength of association – affects all measures of impact.
2. Level of background incidence – affects only amount of incidence (CID, IDD)
3. Population prevalence of the exposure – affects only impact in the population (e.g., PAF).

Appendix — Relating risk factors to health outcomes

Estimating exposure-specific incidence and attributable risk from a case-control when the crude incidence is known

This procedure makes use of the fact that the crude incidence can be expressed as a weighted average of exposure-specific incidences:

$$I = P_1 I_1 + P_0 I_0$$

where:

I = crude incidence

I_1 = incidence in exposed population

I_0 = incidence in unexposed population

P_1 = proportion of the population that is exposed

P_0 = proportion of the population that is unexposed

Since the RR (relative risk, CIR, IDR) = I_1/I_0, it is possible to substitute $RR \times I_0$ for I_1 in the above expression:

$$I = P_1 \times RR \times I_0 + P_0 I_0$$

Similarly, since $P_1 + P_0 = 1$, we can substitute for $(1 - P_1)$ for P_0:

$$I = P_1 \times RR \times I_0 + (1 - P_1) \times I_0$$

Solving for I_0 yields:

$$I_0 = \frac{I}{P_1 \times RR + (1 - P_1)} = \frac{I}{1 + P_1 (RR - 1)}$$

Since for a rare disease we can estimate RR by using OR, the final formula is:

$$I_0 = \frac{I}{1 + P_1 (OR - 1)}$$

This formula can be used for a case-control study where:

1. The control group has been selected in such a way that the proportion exposed estimates P_1 in the population;
2. There is information available to estimate the crude incidence of the disease;
3. The disease is sufficiently rare (e.g., incidence proportion, by the end of follow-up, is less than 10% in both exposure groups) so that the OR estimates the RR fairly well.

Once we have estimated I_0, we estimate I_1 by multiplying by the OR. From I_1 and I_0 we can estimate attributable risk.

Demonstration that OR estimates CIR when CI's are small

For this demonstration and the following one, a different notation will simplify the presentation. We will use D and E for disease and exposure, so that we can upper case for presence and lower case for absence. Thus, subscript E will refer to the presence of exposure, subscript e will refer to the absence of exposure. Similarly, subscript D refers to cases, subscript d to noncases. P stands for probability, which can also be interpreted as prevalence (when applied to exposure) or risk (when applied to disease). The vertical bar means "given" or "conditional on". Thus:

P_E = Probability of being exposed (i.e., prevalence of exposure)

P_e = Probability of being unexposed (1 - P_E)

P_D = Probability of disease (risk)

P_d = Probability of nondisease (1 - P_D)

$P_{E|D}$ = Probability of exposure conditional on disease (i.e., prevalence of exposure in cases)

$P_{e|D}$ = Probability of nonexposure conditional on disease (i.e., 1 - $P_{E|D}$)

$P_{E|d}$ = Probability of exposure conditional on non-disease (i.e., prevalence of exposure in noncases)

$P_{e|d}$ = Probability of nonexposure conditional on non-disease (i.e., 1 - $P_{E|d}$)

$P_{D|E}$ = Probability of disease (risk) conditional on exposure (i.e., risk of disease in the exposed)

$P_{d|E}$ = Probability of nondisease conditional on exposure (i.e., 1 - $P_{D|E}$)

$P_{D|e}$ = Probability of disease conditional on non-exposure (i.e., risk of disease in the unexposed)

$P_{d|e}$ = Probability of nondisease conditional on non-exposure (i.e., 1 - $P_{D|e}$)

By definition, $odds_r$ = risk / (1 - risk). The risk odds ratio (OR_r) is the ratio of odds for exposed persons to odds for unexposed persons. In probability notation:

$$OR_r = \frac{P_{D|E}/(1 - P_{D|E})}{P_{D|e}/(1 - P_{D|e})} = \frac{P_{D|E}}{P_{D|e}} \times \frac{(1 - P_{D|e})}{(1 - P_{D|E})} = RR \times \frac{(1 - P_{D|e})}{(1 - P_{D|E})}$$

since $P_{D|E} / P_{D|e} = RR$. When a disease is rare, $P_{D|E}$ and $P_{D|e}$ are both small, so $OR_r \approx RR$. Using CI to estimate risk and CIR to estimate RR, we have $OR_r \approx CIR$ when CI is small in both exposed and unexposed groups. To illustrate, make up some 2×2 tables to reflect various disease risks and CIR's, and compute the OR's. You can verify that the OR is always farther from 1.0 than the CIR but when the incidence is below about 10%, the OR deviates little from the CIR. The OR can also be expressed as $OR_r = CIR + (OR_r -)P_{D|E}$ (when $OR > 1$; see Hogue, Gaylor, and Schultz, 1983), which demonstrates that the absolute difference between OR and CIR is related to the size of the OR and the disease risk in exposed persons.

Bibliography

Textbook chapters (see preceding listing under Measuring Disease).

Breslow, N.E. and N.E. Day. *Statistical methods in cancer research: volume 1 – the analysis of case-control studies*. IARC Scientific Publications No. 32. Lyon, International Agency for Research on Cancer, 1980.

Davies, Huw Talfryn Oakley; Iain Kinloch Crombie, Manouche Tavakoli. When can odds ratios mislead? BMJ 1998;316:989-991.

Deubner, David C., Herman A. Tyroler, John C. Cassel, Curtis G. Hames, and Caroline Becker. Attributable risk, population attribution risk, and population attributable fraction of death associated with hypertension in a biracial population. *Circulation* 1975;52:901-908

Freeman, Jonathan; George B. Hutchison. Duration of disease, duration indicators, and estimation of the risk ratio. *Am J Epidemiol* 1986; 124:134-49. (Advanced)

Gladen, Beth C. On graphing rate ratios. *Am J Epidemiol* 1983; 118:905-908.

Greenland, Sander. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. *Am J Public Health* 1999 (August)

Greenland, Sander. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol* 1987; 125:761-768 and correspondence in *Am J Epidemiol* 1988; 128:1181-1184.

Greenland, Sander; James M. Robins. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol* 1988; 128:1185-1197. (Intermediate)

Greenland, Sander; Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am J Epidemiol* 1986; 123:203-208. (Advanced)

Hanley, J.A. A heuristic approach to the formulas for population attributable fraction. *J Epidemiol. Community Health* 2001; 55:508-514. (Very understandable)

Hebert, James R.; Donald R. Miller. Plotting and discussion of rate ratios and relative risk estimates. Letter. *J Clin Epidemiol* 1989; 42(3); 289-290.

Hogue, Carol J.R.; David W. Gaylor, and Kenneth F. Schulz. Estimators of relative risk for case-control studies. *Am J Epidemiol* 1983;118:396-407.

Koopman JS, Longini IM, Jacquez JA, Simon CP, et al. Assessing risk factors for transmission of infection. *Am J Epidemiol* 1991 (June); 133(12):1199-1209.

Lee, James. Odds ratio or relative risk for cross-sectional data? *Intl J Epidemiol* 1994;23:201-203.

Lee, James; KS Chia. Estimation of prevalence rate ratios for cross sectional data an example in occupational epidemiology. *Br J Ind Med* 1993;50:861-2.

Rockhill, Beverly; Beth Newman and Clarice Weinberg. Use and misuse of population attributable fractions. *Am J Public Health* 1998; 88(1):15-19. (Note that this article has a typographical error – an extraneous overbar above the C – in the first formula and its explanation.)

Rothman, Modern epidemiology, pp. 285-295; Schlesselman, Case-control studies, pp. 227-234.

Schulman KA, Berlin JA, Harless W, Kerner JF, Sistrunk S, Gersh BJ, Dubé R, Taleghani CK, Burke JE, Williams S, Eisenberg JM, Escarce JJ, Ayers W. The effect of race and sex on physicians' recommendations for cardiac catheterization. *N Engl J Med* 1999 (Feb 25); 340:618-626.

Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *N Engl J Med* 1999 (July 22);341:279-283.

Thompson, Mary Lou; J.E. Myers, D. Kriebel. Prevalence odds ratio or prevalence ratio in the analysis of cross-sectional data: what is to be done. *Occup Environ Med* 1998;55:272-277.

Walter, Stephen D. Calculation of attributable risks from epidemiological data. *Intl J of Epidemiol* 7:175-182, 1978.

Walter, Stephen D. Choice of effect measure for epidemiological data. *J Clinical Epidemiology* 2000;53:931-939.

Zhang, Jun; Kai F. Yu. What's the relative risk: a method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998;280(19):1690-1691.