

3

Measuring Associations Between Exposures and Outcomes

3.1 INTRODUCTION

Epidemiologists are often interested in assessing the presence of associations expressed by differences in disease frequency. Measures of association can be based on either *absolute differences* between measures of disease frequency in groups being compared (e.g., exposed versus unexposed) or *relative differences* or ratios (Table 3–1). Measures based on absolute differences are often preferred when public health or preventive activities are contemplated, as their main goal is often an absolute reduction in the risk of an undesirable outcome. In contrast, etiologic studies that are searching disease causes and determinants usually rely on relative differences in the occurrence of discrete outcomes, with the possible exception of instances in which the outcome of interest is continuous; in this situation, the assessment of mean absolute differences between exposed and unexposed individuals is also a frequently used method for the determination of an association (Table 3–1).

3.2 MEASURING ASSOCIATIONS IN A COHORT STUDY

In traditional prospective or cohort studies, study participants are selected in one of two ways: (1) a defined population or population sample is included in the study and classified according to level of exposure, or (2) exposed and unexposed individuals are specifically identified and included in the study. These individuals are then followed concurrently or nonconcurrently^{1,2} for ascertainment of the outcome(s), allowing for the estimation of an incidence measure in each group (see also Chapters 1 and 2).

So as to simplify the concepts described in this chapter, only two levels of exposure are considered in most of the examples that follow—exposed and unexposed. Furthermore, the length of follow-up is assumed to be complete in all individuals in the cohort (i.e., no censoring occurs). (The discussion that follows, however, also generally applies to risk and rate estimates that take into account incomplete follow-up and censoring, described in the previous chapter; Section 2.2.) For simplification purposes, this chapter focuses almost exclusively on the ratio of two simple incidence probabilities (proportions/risks) or odds (which

Table 3–1 Types of Measures of Association Used in Analytic Epidemiologic Studies

Type	Examples	Usual application
Absolute difference	Attributable risk in exposed	Primary prevention impact; search for causes
	Population attributable risk Effectiveness, Efficacy	Primary prevention impact Impact of intervention on recurrences, case fatality, etc
	Mean differences (continuous outcomes)	Search for determinants
Relative difference	Relative risk/rate	Search for causes
	Relative odds	Search for causes

are generically referred to in this chapter as relative risk and odds ratio, respectively) or on the absolute difference between two incidence probabilities (i.e., the attributable risk); however, concepts described in relationship to these measures also apply to a great extent to the other related association measures, such as the rate ratio and the hazard ratio. Finally, for the purposes of simplifying the description of measures of association, it is generally assumed that the estimates are not affected by either confounding or bias.

3.2.1 Relative Risk (Risk Ratio) and Odds Ratio

A classic two-by-two cross-tabulation of exposure and disease in a cohort study is shown in Table 3–2. Of a total of $(a + b)$ exposed and $(c + d)$ unexposed individuals, a exposed and c unexposed develop the disease of interest during the follow-up time. The corresponding risk and odds estimates are shown in the last two columns of Table 3–2. The probability odds of the disease (the ratio of the probability of disease to the probability of no disease) arithmetically reduces to the ratio of the number of diseased cases divided by the number of individuals who do not develop the disease for each exposure category.

The *relative risk* of developing the disease is expressed as the ratio of the risk (incidence) in exposed individuals (q_+) to that in unexposed (q_-):

[Equation 3.1]

$$\text{Relative risk (RR)} = \frac{q_+}{q_-} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

For methods on estimating confidence limits and p values for a relative risk, see Appendix A, Section A.3.

The *odds ratio* (or *relative odds*) of disease development is the ratio of the odds of developing the disease in exposed divided by that in unexposed individuals; because, in this example, it is based on the incidence proportions or probabilities, it is occasionally designated *probability relative odds*. The ratio of the probability odds of disease is equivalent to

the *cross-product ratio*, $(a \times d)/(b \times c)$. Using the notation in Table 3–2, note that the ratio of the probability odds of disease is equivalent to the *cross-product ratio*, $(a \times d)/(b \times c)$:

$$\text{Probability odds ratio (OR)} = \frac{\frac{q_+}{1 - q_+}}{\frac{q_-}{1 - q_-}} = \frac{\frac{\frac{a}{a+b}}{1 - \left(\frac{a}{a+b}\right)}}{\frac{\frac{c}{c+d}}{1 - \left(\frac{c}{c+d}\right)}} = \frac{\frac{\frac{a}{a+b}}{\frac{b}{a+b}}}{\frac{\frac{c}{c+d}}{\frac{d}{c+d}}} = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)}$$

Thus,

[Equation 3.2]

$$\text{OR} = \frac{a \times d}{b \times c}$$

For methods on obtaining confidence limits and p values for an odds ratio, see Appendix A, Section A.4.

In the hypothetical example shown in Table 3–3, severe hypertension and acute myocardial infarction are the exposure and the outcome of interest, respectively. The sample size for each level of exposure was arbitrarily set at 10,000 to facilitate the calculations. For these data, because the probability (risk) of myocardial infarction is low for both the exposed and the unexposed groups, the probability odds of developing the disease approximates the probabilities; as a result, the probability odds ratio of disease (exposed vs. unexposed) approximates the relative risk:

$$\text{RR} = \frac{\frac{180}{10,000}}{\frac{30}{10,000}} = \frac{0.0180}{0.0030} = 6.00$$

$$\text{Probability OR} = \frac{\frac{180}{9820}}{\frac{30}{9970}} = \frac{0.01833}{0.00301} = 6.09$$

Table 3–2 Cross-Tabulation of Exposure and Disease in a Cohort Study

Exposure	Disease Incidence (Risk)		Probability Odds of Disease
	Diseased	Nondiseased	
Present	a	b	$q_+ = \frac{a}{a+b}$ $\frac{q_+}{1 - q_+} = \frac{\frac{a}{a+b}}{1 - \left(\frac{a}{a+b}\right)} = \frac{a}{b}$
Absent	c	d	$q_- = \frac{c}{c+d}$ $\frac{q_-}{1 - q_-} = \frac{\frac{c}{c+d}}{1 - \left(\frac{c}{c+d}\right)} = \frac{c}{d}$

Table 3–3 Hypothetical Cohort Study of the 1-Year Incidence of Acute Myocardial Infarction in Individuals with Severe Systolic Hypertension (≥ 180 mm Hg) and Normal Systolic Blood Pressure (<120 mm Hg)

Blood Pressure Status	Myocardial Infarction				
	Number	Present	Absent	Probability	Probability Odds _{dis}
Severe hypertension	10,000	180	9820	180/10,000 = 0.0180	180/(10,000 – 180) = 180/9820 = 0.01833
Normal	10,000	30	9970	30/10,000 = 0.0030	30/(10,000 – 30) = 30/9970 = 0.00301

A different situation emerges when the probabilities of developing the outcome are high in exposed and unexposed individuals. For example, Seltser et al.³ examined the incidence of local reactions in individuals assigned randomly to either an injectable influenza vaccine or a placebo group. Table 3–4, based on this study, shows that as the probability (incidence) of local reactions is high, the probability odds estimates of local reactions do not approximate the probabilities (particularly in the group assigned to the vaccine). Thus, the probability odds ratio of local reactions (vaccine vs. placebo) is fairly different from the relative risk:

$$RR = \frac{\frac{650}{2570}}{\frac{170}{2410}} = \frac{0.2529}{0.0705} = 3.59 \quad OR = \frac{\frac{650}{1920}}{\frac{170}{2240}} = \frac{0.3385}{0.0759} = 4.46$$

When the condition of interest has a high incidence and when prospective data are available, as was the case in this vaccination trial, it is usually better to report the relative risk because it is a more easily understood measure of association between the risk factor and the outcome.

Although, as discussed later, the odds ratio is a valid measure of association in its own right, it is often used as an approximation of the relative risk. The use of the odds ratio *as an*

Table 3–4 Incidence of Local Reactions in the Vaccinated and Placebo Groups, Influenza Vaccination Trial

Group	Local Reaction				
	Number	Present	Absent	Probability	Probability Odds _{dis}
Vaccine	2570	650	1920	650/2570 = 0.2529	650/(2570 – 650) = 650/1920 = 0.3385
Placebo	2410	170	2240	170/2410 = 0.0705	170/(2410 – 170) = 170/2240 = 0.0759

Note: Based on data for individuals 40 years old or older in Seltser et al.³ To avoid rounding ambiguities in subsequent examples based on these data (Figure 3–4, Tables 3–7 and 3–9), the original sample sizes in Seltser et al.'s study (257 vaccinees and 241 placebo recipients) were multiplied by 10.

Source: Data from R Seltser, PE Sartwell, and JA Bell, A Controlled Test of Asian Influenza Vaccine in Population of Families, *American Journal of Hygiene*, Vol 75, pp 112–135, © 1962.

estimate of the relative risk biases it in a direction opposite to the null hypothesis: that is, it tends to exaggerate the magnitude of the association. When the disease is relatively rare, this “built-in” bias is negligible, as in the previous example from Table 3–3. When the incidence is high, however, as in the vaccine trial example (Table 3–4), the bias can be substantial.

An expression of the mathematical relationship between the odds ratio on the one hand and the relative risk on the other can be derived as follows. Assume that q_+ is the incidence (probability) in exposed (e.g., vaccinated) and q_- the incidence in unexposed individuals. The odds ratio is then

[Equation 3.3]

$$\begin{aligned} \text{OR} &= \frac{\left(\frac{q_+}{1-q_+}\right)}{\left(\frac{q_-}{1-q_-}\right)} = \frac{q_+}{1-q_+} \times \frac{1-q_-}{q_-} \\ &= \frac{q_+}{q_-} \times \left(\frac{1-q_-}{1-q_+}\right) \end{aligned}$$

The term q_+/q_- in Equation 3.3 is the relative risk. Thus, the term

$$\left(\frac{1-q_-}{1-q_+}\right)$$

defines the *bias* responsible for the discrepancy between the relative risk and the odds ratio estimates (*built-in bias*). If the association between the exposure and the outcome is positive, $q_- < q_+$, thus $(1-q_-) > (1-q_+)$. The bias term will therefore be greater than 1.0, leading to an overestimation of the relative risk by the odds ratio. By analogy, if the factor is related to a decrease in risk, the opposite occurs (i.e., $[1-q_-] < [1-q_+]$), and the odds ratio will again overestimate the strength of the association (in this case, by being smaller than the relative risk in absolute value). In general, the odds ratio tends to yield an estimate further away from 1.0 than the relative risk on both sides of the scale (above or below 1.0).

In the hypertension/myocardial infarction example (Table 3–3), the bias factor is of a small magnitude, and the odds ratio estimate, albeit a bit more distant from 1.0, still approximates the relative risk; using Equation 3.3:

$$\text{OR} = \text{RR} \times \text{“built-in bias”} = 6.0 \times \frac{1 - 0.0030}{1 - 0.0180} = 6.0 \times 1.015 = 6.09$$

In the example of local reactions to the influenza vaccine (Table 3–4), however, there is a considerable bias when using the odds ratio to estimate the relative risk:

$$\text{OR} = 3.59 \times \frac{1 - 0.0705}{1 - 0.2529} = 3.59 \times 1.244 = 4.46$$

Regardless of whether the odds ratio can properly estimate the relative risk, it is, as mentioned previously, a *bona fide* measure of association. Thus, a *built-in bias* can be only said to exist when the odds ratio is used as an estimate of the relative risk. The odds ratio is especially valuable because it can be measured in case-control (case–noncase) studies and because it is directly derived from logistic regression models (see Chapter 7, Section 7.4.3).

In addition, unlike the relative risk, the odds ratio of an event is the exact reciprocal of the odds ratio of the nonevent. For example, in the study of local reactions to the influenza vaccine discussed previously,³ the odds ratio of a local reaction

$$\text{OR}_{\text{local reaction (+)}} = \frac{\left(\frac{650}{1920}\right)}{\left(\frac{170}{2240}\right)} = 4.46$$

is the exact reciprocal of the odds ratio of not having a local reaction

$$\text{Probability OR}_{\text{local reaction (-)}} = \frac{\left(\frac{1920}{650}\right)}{\left(\frac{2240}{170}\right)} = 0.22 = \frac{1}{4.46}$$

This feature is not shared by the relative risk: using the same example

$$\text{RR}_{\text{local reaction (+)}} = \frac{\frac{650}{2570}}{\frac{170}{2410}} = 3.59$$

and

$$\text{RR}_{\text{local reaction (-)}} = \frac{\left(\frac{1920}{2570}\right)}{\left(\frac{2240}{2410}\right)} = 0.8 \neq \frac{1}{3.59}$$

This seemingly paradoxical finding results from the sensitivity of the relative risk to the absolute frequency of the condition of interest, with relative risks associated with *very* common endpoints approaching 1.0. This is easily appreciated when studying the complement of rare outcomes. For example, if the case fatality rates of patients undergoing surgery using a standard surgical technique and a new technique were 0.02 and 0.01, respectively, the relative risk for the relatively rare outcome “death” would be 0.02/0.01 = 2.0. The relative risk for survival, however, would be 0.98/0.99, which is virtually equal to 1.0, suggesting that the new surgical technique did not affect survival. On the other hand, the odds ratio of death would be

$$\text{OR}_{\text{death}} = \frac{\frac{0.02}{1.0 - 0.02}}{\frac{0.01}{1.0 - 0.01}} = 2.02$$

and that of survival would be

$$\text{OR}_{\text{survival}} = \frac{\frac{0.98}{1.0 - 0.98}}{\frac{0.99}{1.0 - 0.99}} = 0.495 = \frac{1.0}{2.02}$$

3.2.2 Attributable Risk

The *attributable risk* is a measure of association based on the absolute difference between two risk estimates. Thus, the attributable risk estimates the absolute excess risk associated with a given exposure. Because the attributable risk is often used to imply a cause–effect relationship, it should be interpreted as a true *etiologic fraction* only when there is reasonable certainty of a causal connection between exposure and outcome.^{4,5} The term *excess fraction* has been suggested as an alternative term when causality has not been firmly established.⁴ Also, although the formulas and examples in this section generally refer to attributable “risks,” they are also applicable to attributable rates or densities; that is, if incidence data based on person-time are used, an attributable rate among the exposed (see later here) can be calculated in units of rate per person-time.

As extensively discussed by Gordis,² the attributable risk assumes the following different formats:

Attributable Risk in Exposed Individuals

The attributable risk in the exposed is merely the difference between the risk estimates of different exposure levels and a reference exposure level; the latter is usually formed by the unexposed category. Assuming a binary exposure and letting risk in exposed equal q_+ and risk in unexposed equal q_- , the attributable risk in the exposed (AR_{exp}) is simply

[Equation 3.4]

$$AR_{\text{exp}} = q_+ - q_-$$

The attributable risk in the exposed measures the excess risk associated with a given exposure category. For example, based on the example in Table 3–3, the cumulative incidence of myocardial infarction among the hypertensive individuals (q_+) is 0.018 (or 1.8%), and that for normotensives (reference or unexposed category) (q_-) is 0.003 (or 0.3%); thus, the excess risk associated with exposure to hypertension is $0.018 - 0.003 = 0.015$ (or 1.5%). That is, assuming a causal association (and thus, no confounding or bias—see Chapters 4 and 5) and if the excess incidence were completely reversible, the cessation of the exposure (severe systolic hypertension) would lower the risk in the exposed group from 0.018 to 0.003. In Figure 3–1, the two bars represent the cumulative incidence in exposed and non-exposed individuals; thus, the attributable risk in the exposed (Equation 3.4) is the difference in height of these bars. Because it is the difference between two incidence measures, the attributable risk in the exposed is also an absolute incidence magnitude and therefore is measured using the same units. The estimated attributable risk in the exposed of 1.5% in the previous example represents the absolute excess incidence that would be prevented by eliminating severe hypertension.

Because most exposure effects are cumulative, cessation of exposure (even if causally related to the disease) usually does not reduce the risk in exposed individuals to the level found in those who were never exposed. Thus, the maximum risk reduction is usually achieved only through prevention of exposure rather than its cessation.

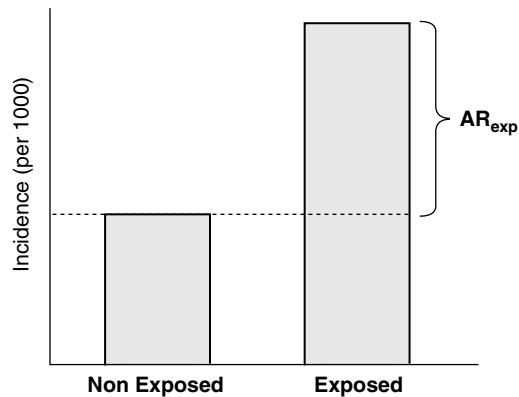


Figure 3–1 Attributable risk in the exposed.

Percent Attributable Risk in Exposed Individuals

A percent attributable risk in the exposed ($\%AR_{\text{exp}}$) is merely the AR_{exp} expressed as a percentage of the q_+ (i.e., the percentage of the total q_+ that can be attributed to the exposure). For a binary exposure variable, it is as follows:

[Equation 3.5]

$$\%AR_{\text{exp}} = \left(\frac{q_+ - q_-}{q_+} \right) \times 100$$

In the example shown in Table 3–3, the percent attributable risk in the exposed is

$$\%AR_{\text{exp}} = \frac{0.018 - 0.003}{0.018} \times 100 = 83.3\%$$

If causality had been established, this measure can be interpreted as the percentage of the total risk of myocardial infarction among hypertensives attributable to hypertension.

It may be useful to express Equation 3.5 in terms of the relative risk:

$$\%AR_{\text{exp}} = \left(\frac{q_+ - q_-}{q_+} \right) \times 100 = \left(1 - \frac{1}{RR} \right) \times 100 = \left(\frac{RR - 1.0}{RR} \right) \times 100$$

Thus, in the previous example, using the relative risk ($0.018/0.003 = 6.0$) in this formula produces the same result as when applying Equation 3.5:

$$\%AR_{\text{exp}} = \left(\frac{6.0 - 1.0}{6.0} \right) \times 100 = 83.3\%$$

The obvious advantage of the formula

[Equation 3.6]

$$\%AR_{\text{exp}} = \left(\frac{RR - 1.0}{RR} \right) \times 100$$

is that it can be used in case-control studies, in which incidence data (i.e., q_+ or q_-) are unavailable, but the odds ratio can be used as an estimate of the relative risk if the disease is relatively rare (see Section 3.2.1).

The percent attributable risk in the exposed is analogous to percentage *efficacy* when assessing an intervention such as a vaccine. The usual formula for efficacy is equivalent to the formula for percent attributable risk in the exposed (Equation 3.5) when q_+ is replaced by q_{cont} (risk in control group, e.g., the group receiving placebo) and q_- is replaced by q_{interv} (risk in those undergoing intervention):

[Equation 3.7]

$$\text{Efficacy} = \left(\frac{q_{\text{cont}} - q_{\text{interv}}}{q_{\text{cont}}} \right) \times 100$$

For example, in a randomized trial to evaluate the efficacy of a vaccine, the risks in persons receiving the vaccine and the placebo are 5% and 15%, respectively. Using Equation 3.7, efficacy is found to be 66.7%:

$$\text{Efficacy} = \left(\frac{15\% - 5\%}{15\%} \right) \times 100 = 66.7\%$$

Alternatively, Equation 3.6 can be used to estimate efficacy. In the previous example, the relative risk (placebo/vaccine) is $15\% \div 5\% = 3.0$. Thus,

$$\text{Efficacy} = \left(\frac{3.0 - 1.0}{3.0} \right) \times 100 = 66.7\%$$

The use of Equation 3.6 for the calculation of efficacy requires that, when calculating the relative risk, the group not receiving the intervention (e.g., placebo) be labeled “exposed” and the group receiving the active intervention (e.g., vaccine) be labeled as “unexposed.” A mathematically equivalent approach would consist of first obtaining the relative risk, but this time with the risk of those receiving the active intervention (e.g., vaccine) in the numerator and those not receiving it in the denominator (e.g., placebo). In this case, efficacy is calculated as the complement of the relative risk, that is, $(1.0 - \text{RR}) \times 100$. In the previous example, using this approach, the vaccine efficacy would be

$$\text{Efficacy} = \left[1.0 - \left(\frac{5\%}{15\%} \right) \right] \times 100 = 66.7\%$$

As for percent attributable risk, the correspondence between the relative risk and the odds ratio in most practical situations allows the estimation of efficacy in case-control studies using Equation 3.6.

Levin’s Population Attributable Risk

Levin’s population attributable risk estimates the proportion of the disease risk in the total population associated with the exposure.⁶ For example, let the exposure prevalence in the target population (p_e) be 0.40 (and, thus, prevalence of *nonexposure*, $[1 - p_e]$, be 0.60), and the risks in exposed and unexposed be $q_+ = 0.20$ and $q_- = 0.15$, respectively. Thus, the risk in the total population (q_{pop}) is as follows:

[Equation 3.8]

$$q_{\text{pop}} = [q_+ \times p_e] + [q_- \times (1 - p_e)]$$

?/Ed:
Eq. 3.7
larger
parens.
needed?

representing the weighted sum of the risks in the exposed and unexposed individuals in the population. In the example

$$q_{\text{pop}} = (0.20 \times 0.40) + (0.15 \times 0.60) = 0.17$$

The population attributable risk (Pop AR) is the difference between the risk in the total population and that in unexposed subjects:

$$\text{Pop AR} = q_{\text{pop}} - q_{-}$$

Thus, in the example, the population attributable risk is $0.17 - 0.15 = 0.02$. That is, if the relationship were causal and if the effect of the exposure were completely reversible, exposure cessation would be expected to result in a decrease in total population risk (q_{pop}) from 0.17 to 0.15 (i.e., to the level of risk of the unexposed group).

The Pop AR is usually expressed as the percent population attributable risk (%Pop AR):

[Equation 3.9]

$$\% \text{Pop AR} = \frac{(q_{\text{pop}} - q_{-})}{q_{\text{pop}}} \times 100$$

In the previous example, the percent population attributable risk is $(0.02/0.17) \times 100$, or approximately 12%.

As seen in Equation 3.8, the incidence in the total population is the sum of the incidence in the exposed and that in the unexposed, weighted for the proportions of exposed and unexposed individuals in the population. Thus, when the exposure prevalence is low, the population incidence will be closer to the incidence among the unexposed (Figure 3–2A). Similarly, if the exposure is highly prevalent (Figure 3–2B), the population incidence will

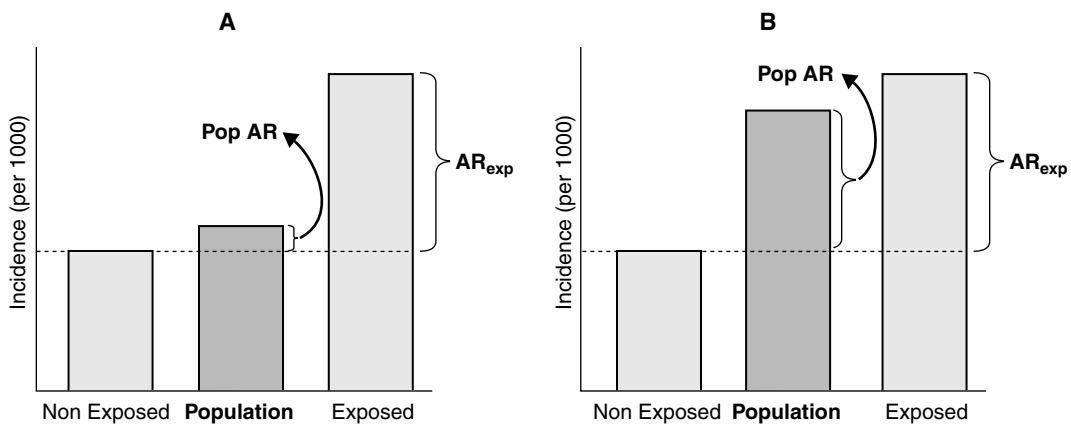


Figure 3–2 Population attributable risk and its dependence on the population prevalence of the exposure. As the population is composed of exposed and unexposed individuals, the incidence in the population is similar to the incidence in the unexposed when the exposure is rare (A) and is closer to that in the exposed when the exposure is common (B). Thus, for a fixed relative risk (eg, $RR \approx 2$ in the figure) the population attributable risk is heavily dependent on the prevalence of exposure.

be closer to the incidence among the exposed. As a result, the population attributable risk approximates the attributable risk when exposure prevalence is high.

After simple arithmetical manipulation,* Equation 3.9 can be expressed as a function of the exposure prevalence in the population and the relative risk, as first described by Levin:⁶

[Equation 3.10]

$$\% \text{Pop AR} = \frac{p_e \times (\text{RR} - 1)}{p_e \times (\text{RR} - 1) + 1} \times 100$$

Using the same example of a population with an exposure prevalence of 0.40 and a relative risk = 0.20/0.15 = 1.33, Equation 3.10 yields the same percent population attributable risk estimated previously:

$$\% \text{Pop AR} = \frac{0.40 \times (1.33 - 1.0)}{0.40 \times (1.33 - 1.0) + 1.0} \times 100 = \frac{0.40 \times 0.33}{0.40 \times 0.33 + 1.0} = 12\%$$

For a method of calculating the confidence limits of the population attributable risk, see Appendix A, Section A.5.

Levin's formula underscores the importance of the two critical elements contributing to the magnitude of the population attributable risk: the relative risk and the prevalence of exposure. The dependence of the population attributable risk on the exposure prevalence is further illustrated in Figure 3-3, which shows that for all values of the relative risk, the population attributable risk increases markedly as the exposure prevalence increases.

The application of Levin's formula in case-control studies requires using the odds ratio as an estimate of the relative risk and obtaining an estimate of exposure prevalence in the reference population, as discussed in more detail in Section 3.4.2.

*Using Equation 3.8, Equation 3.9 can be rewritten as a function of the prevalence of exposure (p_e) and the incidence in exposed (q_+) individuals, as follows:

$$\begin{aligned} \% \text{Pop AR} &= \frac{[q_+ \times p_e] + [q_- \times (1 - p_e)] - q_-}{[q_+ \times p_e] + [q_- \times (1 - p_e)]} \times 100 \\ &= \frac{[q_+ \times p_e] - [q_- \times p_e]}{[q_+ \times p_e] - [q_- \times p_e] + q_-} \times 100 \end{aligned}$$

This expression can be further simplified by dividing all the terms in numerator and denominator by q_-

$$\begin{aligned} \% \text{Pop AR} &= \frac{\frac{q_+}{q_-} \times p_e - p_e}{\frac{q_+}{q_-} \times p_e - p_e + 1} \times 100 = \frac{p_e \times \left(\frac{q_+}{q_-} - 1 \right)}{p_e \times \left(\frac{q_+}{q_-} - 1 \right) + 1} \times 100 \\ &= \frac{p_e \times (\text{RR} - 1)}{p_e \times (\text{RR} - 1) + 1} \times 100 \end{aligned}$$

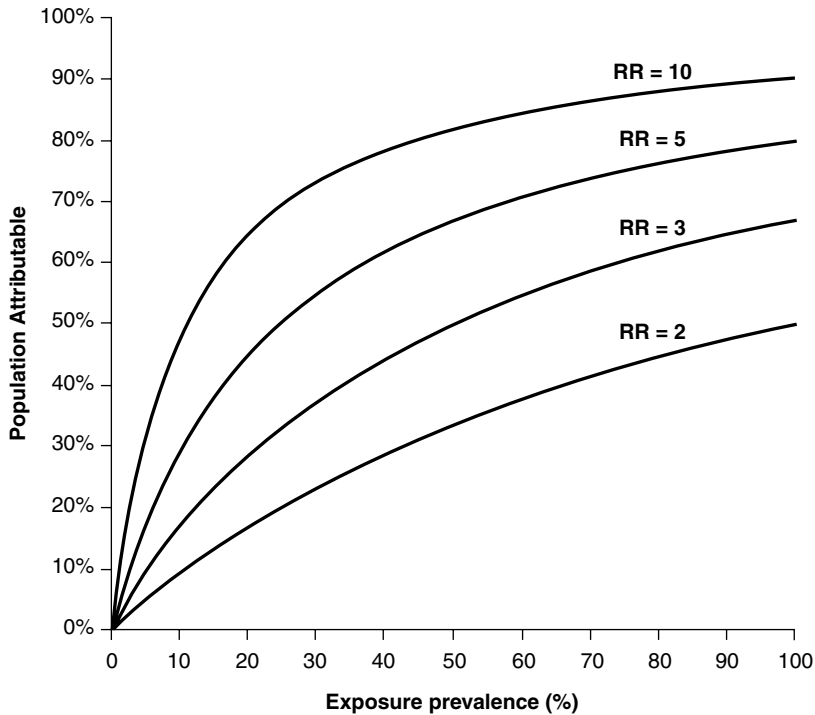


Figure 3-3 Population attributable risk: dependence on prevalence of exposure and relative risk.

All of the preceding discussion relates to a binary exposure variable (i.e., exposed vs. unexposed). When exposure has more than just two categories, an extension of Levin's formula has been derived by Walter.⁷

$$\begin{aligned} \% \text{Pop AR} &= \frac{\sum_{i=0}^k p_i \times (RR_i - 1)}{1 + \sum_{i=0}^k p_i \times (RR_i - 1)} \times 100 = \\ &= \left[1.0 - \frac{1.0}{\sum_{i=0}^k (p_i \times RR_i)} \right] \times 100 \end{aligned}$$

The subscript i denotes each exposure level; p_i is the proportion of the study population in the exposure level i , and "RR $_i$ " is the relative risk for the exposure level i compared with the unexposed (reference) level.

It is important to emphasize that both Levin's formula and Walter's extension for multilevel exposures assume that there is no confounding (see Chapter 5). If confounding is present, it is not correct to calculate the adjusted relative risk (using any of the approaches described in Chapter 7) and plug it into Levin's or Walter's formulas in order to obtain an "adjusted" population attributable risk.⁸ Detailed discussions on the estimation of the population attributable risk in the presence of confounding can be found elsewhere.^{7,9}

3.3 CROSS-SECTIONAL STUDIES: POINT PREVALENCE RATE RATIO

When cross-sectional data are available, often associations are assessed using the point prevalence rate ratio. The ability of the point prevalence ratio to estimate the relative risk is a function of the relationship between incidence and point prevalence, as discussed previously in Chapter 2 (Section 2.3, Equation 2.4):

$$\text{Point Prevalence} = \text{Incidence} \times \text{Duration} \times (1 - \text{Point Prevalence})$$

Using the notations “Prev” for point prevalence, “ q ” for incidence, and “Dur” for duration and denoting presence or absence of a given exposure by “+” or “-,” the point prevalence ratio (PRR) can be formulated as follows:

$$\text{PRR} = \frac{\text{Prev}_+}{\text{Prev}_-} = \frac{q_+ \times \text{Dur}_+ \times [1.0 - \text{Prev}_+]}{q_- \times \text{Dur}_- \times [1.0 - \text{Prev}_-]}$$

Because one of the components of this formula (q_+/q_-) is the relative risk, this equation can be written as

[Equation 3.11]

$$\text{PRR} = \text{RR} \times \left(\frac{\text{Dur}_+}{\text{Dur}_-} \right) \times \left(\frac{1 - \text{Prev}_+}{1 - \text{Prev}_-} \right)$$

Thus, if the point prevalence ratio is used to estimate the relative risk (e.g., in a cross-sectional study), two types of bias will differentiate these two measures: the ratio of the disease durations ($\text{Dur}_+/\text{Dur}_-$), and the ratio of the complements of the point prevalence estimates in the exposed and unexposed groups ($1 - \text{Prev}_+/1 - \text{Prev}_-$). Chapter 4 (Section 4.4.2) provides a discussion and examples of these biases.

3.4 MEASURING ASSOCIATIONS IN CASE-CONTROL STUDIES

3.4.1 Odds Ratio

One of the major advances in risk estimation in epidemiology occurred in 1951 when Cornfield pointed out that the *odds ratio of disease and the odds ratio of exposure are mathematically equivalent*.¹⁰ This is a simple concept, yet with important implications for the epidemiologist, as it is the basis for estimating the odds ratio of disease in case-control studies.

As seen previously in Equation 3.2, the ratio of the odds of disease development in exposed and unexposed individuals results in the cross-product ratio, $(a \times d)/(b \times c)$. Using the *prospective* data shown in Table 3–3, now reorganized as shown in Table 3–5, and assuming that the cells in the table represent the distribution of the cohort participants during a 1-year follow-up, it is possible to carry out a case-control analysis comparing the 210 individuals who developed a myocardial infarction (cases) with the 19,790 individuals who remained free of clinical coronary heart disease during the follow-up (controls). The

Table 3–5 Hypothetical Case-Control Study of Myocardial Infarction in Relation to Systolic Hypertension, Based on a 1-Year Complete Follow-up of the Study Population from Table 3–3

Systolic Blood Pressure Status*	Myocardial Infarction			
	Present		Absent	
Severe hypertension	180	(a)	9820	(b)
Normal	30	(c)	9970	(d)
Total	210	(a + c)	19790	(b + d)

*Severe systolic hypertension ≥ 180 mm Hg, and normal systolic blood pressure < 120 mm Hg.

absolute odds of exposure (Odds_{exp}) among cases and the analogous odds of exposure among controls are estimated as the ratio of the proportion of individuals exposed to the proportion of individuals unexposed:

$$\text{Odds}_{\text{exp/cases}} = \frac{\frac{a}{a+c}}{1 - \left(\frac{a}{a+c}\right)} = \frac{a}{c}$$

$$\text{Odds}_{\text{exp/controls}} = \frac{\frac{b}{b+d}}{1 - \left(\frac{b}{b+d}\right)} = \frac{b}{d}$$

The following derivation demonstrates that the odds ratio of exposure (OR_{exp}) is identical to the odds ratio of disease (OR_{dis}):

[Equation 3.12]

$$\text{OR}_{\text{exp}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \times d}{b \times c} = \frac{\frac{a}{b}}{\frac{c}{d}} = \text{OR}_{\text{dis}}$$

For the example shown in Table 3–5, the odds ratio of exposure is

$$\text{OR}_{\text{exp}} = \frac{\frac{180}{30}}{\frac{9820}{9970}} = \frac{180 \times 9970}{9820 \times 30} = 6.09 = \text{OR}_{\text{dis}}$$

In this example based on prospective data, all cases and noncases (controls) have been used for the estimation of the odds ratio; however, case-control studies are typically based on samples. If the total number of cases is small, as in the example shown in Table 3–5, the investigator may attempt to include all cases and a sample of controls. For example, if 100% of cases and a sample of approximately 10% of the noncases were studied (Table 3–6),

Table 3–6 Case-Control Study of the Relationship of Myocardial Infarction to Presence of Severe Systolic Hypertension Including All Cases and a 10% Sample of Noncases from Table 3–5

Systolic Blood Pressure Status*	Myocardial Infarction			
	Present		Absent	
Severe hypertension	180	(a)	982	(b)
Normal	30	(c)	997	(d)
Total	210	(a + c)	1979	(b + d)

*Severe systolic hypertension ≥ 180 mm Hg, and normal systolic blood pressure < 120 mm Hg.

assuming no random variability, results would be identical to those obtained when including all noncases, as in Table 3–5:

$$OR_{\text{exp}} = \frac{\frac{180}{30}}{\frac{982}{997}} = \frac{180 \times 997}{982 \times 30} = 6.09 = OR_{\text{dis}}$$

This example underscores the notion that the sampling fractions do not have to be the same in cases and controls. To obtain unbiased estimates of the absolute odds of exposure for cases and controls, however, sampling fractions must be independent of exposure: that is, they should apply equally to cells (a) and (c) for cases and cells (b) and (d) for controls. (Chapter 4, Section 4.2, presents a more detailed discussion of the validity implications for the OR estimate resulting from differential sampling fractions according to case and exposure status.)

In the example of local reactions to vaccination (Table 3–4), a case-control study could have been carried out including, for example, 80% of the cases that had local reactions and 50% of the controls. Assuming no random variability, data would be obtained as outlined in Figure 3–4 and shown in Table 3–7. If the sampling fractions apply equally to exposed (vaccinated) and unexposed (unvaccinated) cases and controls, the results are again identical to those seen in the total population, in which the (true) odds ratio is 4.46:

$$OR_{\text{exp}} = \frac{\left(\frac{520}{136}\right)}{\left(\frac{960}{1120}\right)} = 4.46 = OR_{\text{dis}}$$

The fact that the odds ratio of exposure is identical to the odds ratio of disease permits a “prospective” interpretation of the odds ratio in case-control studies (i.e., as a comparative measure of “disease odds” [as an approximation of the relative risk—discussed later here]). Thus, in the previous example based on a case-control strategy (and assuming that the study is unbiased and free of confounding), the interpretation of results is that for individuals who received the vaccine, the odds of developing local reactions is 4.46 times greater than the odds for those who received the placebo.

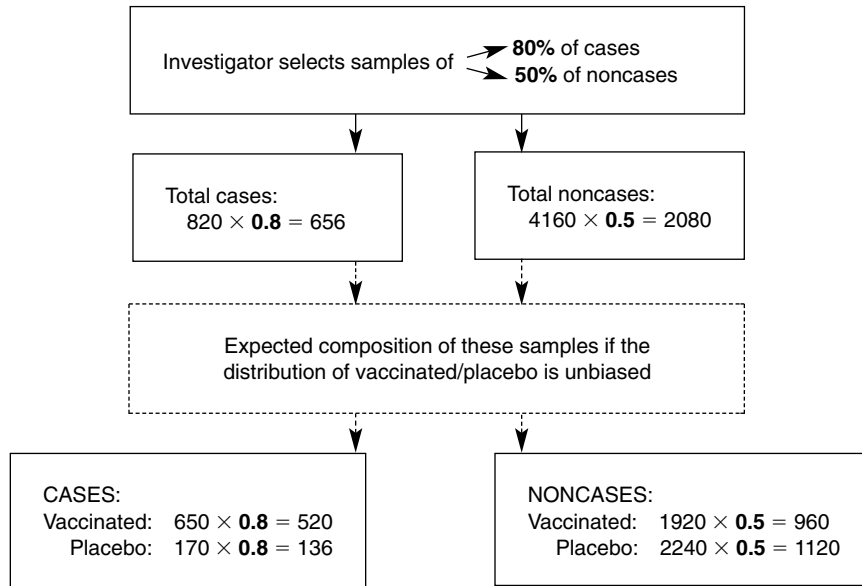


Figure 3–4 Selection of 80% of total cases and 50% of noncases in a case-control study from the study population shown in Table 3–4. Expected composition is assuming no random variability. *Source:* Data from R Seltser, PE Sartwell, and JA Bell, A Controlled Test of Asian Influenza Vaccine in a Population of Families, *American Journal of Hygiene*, Vol 75, pp 112–135, © 1962.

The use of the ratio of the odds of exposure for cases to that for controls,

$$OR_{\text{exp}} = \frac{\text{Odds}_{\text{exp cases}}}{\text{Odds}_{\text{exp controls}}}$$

is strongly recommended for the calculation of the odds ratio of exposure, rather than the cross-products ratio, so as to avoid confusion over different arrangements of the table, as, for example, when placing control data on the left and case data on the right:

<i>Exposure</i>	<i>Controls</i>	<i>Cases</i>
Yes	“a”	“b”
No	“c”	“d”

Table 3–7 Case-Control Study of the Relationship Between Occurrence of Local Reaction and Previous Influenza Immunization

<i>Vaccination</i>	<i>Cases of Local Reaction</i>	<i>Controls Without Local Reaction</i>
Yes	520	960
No	136	1120
Total	820 × 0.8 = 656	4160 × 0.5 = 2080

Note: Based on a perfectly representative sample of 80% of the cases and 50% of the controls from the study population shown in Table 3–4 (see Figure 3–4).

Source: Data from R Seltser, PE Sartwell, and JA Bell, A Controlled Test of Asian Influenza Vaccine in a Population of Families, *American Journal of Hygiene*, Vol 75, pp 112–135, © 1962.

In this example, the mechanical application of the cross-product ratio, $(a \times b)/(c \times d)$, results in an estimate of the odds ratio that is the inverse of the true relative odds. On the other hand, dividing the exposure odds in cases (b/d) by that in controls (a/c) results in the correct odds ratio.

Odds Ratio in Matched Case-Control Studies

In a matched paired case-control study in which the ratio of controls to cases is 1:1, an unbiased estimate of the odds ratio is obtained by dividing the number of pairs in which the case, but not the matched control, is exposed (case [+], control [-]), by the number of pairs in which the control, but not the case, is exposed (case [-], control [+]). The underlying intuitive logic for this calculation and an example of this approach are discussed in Chapter 7, Section 7.3.3.

Odds Ratio as an Estimate of the Relative Risk in Case-Control Studies: The Rarity Assumption

In a case-control study, the use of the odds ratio to estimate the relative risk is based on the assumption that the disease under study has a low incidence, thus resulting in a small built-in bias (Equation 3.3). As a corollary to the discussion in Section 3.2.1, it follows that when the disease that defines case-status in a case-control study is sufficiently rare, the estimated odds ratio will likely be a good approximation to the relative risk. On the other hand, when studying relatively common conditions, the built-in-bias might be large, and case-control studies may yield odds ratios that substantially overestimate the strength of the association vis-à-vis the relative risk. Based on Equation 3.3, the following expression of the relative risk value as a function of the odds ratio can be derived:

[Equation 3.13]

$$RR = \frac{OR}{1 - [q_- - (OR \times q_-)]}$$

It is evident from this equation that the relationship between the relative risk and odds ratio depends on the incidence of the outcome of interest (specifically q_- , i.e., the incidence in the unexposed in this particular formulation). As a corollary of this, Equation 3.13 also implies that, in order to estimate the value of the relative risk from an odds ratio obtained in a case-control study, an estimate of incidence obtained from prospective data will be necessary. This could be available from outside sources (e.g., published data from another cohort study judged to be comparable to the source population for the study in question); if the study is nested within a cohort, incidence data may be available from the parent cohort from which the case and comparison groups were drawn (see examples later here). Table 3–8 illustrates examples of this relationship for a range of incidence and odds ratio values. For outcomes with incidence in the range of less than 1% or 1 per 1000 (e.g., the majority of chronic or infectious diseases), the value of the relative risk is very close to that of the odds ratio. Even for fairly common outcomes with frequency ranging between 1% and 5%, the values of the relative risk and odds ratio are reasonably similar.

Table 3–8, however, shows that when the condition of interest (that defining case-status) is more common (e.g., incidence > 10% to 20%), the value of the odds ratio obtained in case-control studies, will be substantially different than that of the relative risk. This is not a limitation of the case-control design *per se*, but rather, it is a result from the mathematical relation between odds ratio and relative risk, irrespective of study design. It should be kept in mind when interpreting odds ratio values in studies of highly frequent conditions such as, for example, hypertension in a high-risk group (e.g., overweight adult individuals), 5-year mortality among lung cancer cases, or smoking relapse in a smoking cessation study. Additional examples of case-control studies assessing relatively common events can be provided for situations in which incidence data were also available, as follows:

- In a nested case-control study of predictors of surgical site infections after breast cancer surgery, 76 cases of infection were compared with 154 controls (no infection) with regard to obesity and other variables.¹¹ Data were available on the incidence of infection in the nonobese group (q_-), which was estimated to be approximately 30%. The study reported that the odds ratio of infection associated with obesity was 2.5. Thus, using Equation 3.13, the corresponding relative risk of infection comparing obese and nonobese can be estimated as 1.72.
- A case-control study investigated the relationship between genetic changes and prostate cancer progression.¹² Cases were individuals with a biochemical marker of cancer progression (prostate specific agent [PSA] > 0.4 ng/mL, $n = 26$) and were compared with 26 controls without biochemical progression. Loss of heterozygosity (LOH) was associated with an odds ratio of 5.54. Based on data from the study report, the approximate incidence of progression among non-LOH group was 60%; as a result, it is estimated that the 5.54 odds ratio obtained in the study corresponds to a relative risk of approximately 1.5.

Table 3–8 Relative risk equivalent to a given odds ratio as a function of the incidence of the condition that defined case status in a case-control study

Incidence in the unexposed population Odds ratio = 0.5 Odds ratio = 1.5 Odds ratio = 2.0 Odds ratio = 3.0

	<i>Relative Risk Equivalent</i>				
	Odds ratio = 0.5	Odds ratio = 1.5	Odds ratio = 2.0	Odds ratio = 3.0	
0.001	0.50	1.50	2.00	2.99	
0.01	0.50	1.49	1.98	2.94	
0.05	0.51	1.46	1.90	2.73	
0.1	0.53	1.43	1.82	2.50	
0.2	0.56	1.36	1.67	2.14	
0.3	0.59	1.30	1.54	1.88	
0.4	0.63	1.25	1.43	1.67*	

*Example of calculation: for an OR = 3 and $q_- = 0.4$ and using Equation 3.13, the relative risk is:

$$RR = \frac{3}{1 - (0.4 - 3 \times 0.4)} = 1.67$$

As in prospective studies (see Section 3.2.1), the rare-disease assumption *applies only to situations in which the odds ratio is used to estimate the relative risk*. When the odds ratio is used as a measure of association in itself, this assumption is obviously not needed. In the previous examples, there is nothing intrinsically incorrect about the odds ratio estimates; assuming no bias or random error, LOH is indeed associated with an odds ratio of biochemical prostate cancer progression of 5.54. Although it would be a mistake to interpret this estimate as a relative risk, it is perfectly correct to conclude that, compared with non-LOH, LOH multiplies the *odds* of biochemical progression by 5.54; this is as correct as concluding the LOH multiplies the *risk (incidence)* of biochemical progression by 1.5. Both conclusions are equally accurate.

When the Rarity Assumption Does Not Apply: Selecting Population Controls

The rare-disease assumption is irrelevant in situations in which the control group is a sample of the total population,¹³ which is the usual strategy in case-control studies within a defined cohort (Chapter 1, Section 1.4.2). In this situation, the odds ratio is a *direct estimate of the relative risk*, irrespective of the frequency of the outcome of interest.

The irrelevance of the rare-disease assumption when the control group is a sample of the total reference population can be demonstrated by comparing the calculation of the odds ratio using different types of control groups. Referring to the cross-tabulation, including all cases and all noncases in a defined population shown in Table 3–9, when noncases are used as the control group, as seen previously (Equation 3.12), the odds ratio of exposure is used to estimate the odds ratio of disease by dividing the odds of exposure in cases by that in controls:

$$OR_{\text{exp}} = \frac{\text{Odds}_{\text{exp cases}}}{\text{Odds}_{\text{exp noncases}}} = \frac{\frac{a}{c}}{\frac{b}{d}} = OR_{\text{dis}}$$

Another option is to use as a control group the total study population at baseline, rather than only the noncases. If this is done in the context of a cohort study, the case-control study is usually called a *case-cohort study* (Chapter 1, Section 1.4.2), and the division of the odds of exposure in cases by that in controls (i.e., the total population) yields the relative risk:

[Equation 3.14]

$$OR_{\text{exp}} = \frac{\text{Odds}_{\text{exp cases}}}{\text{Odds}_{\text{exp total population}}} = \frac{\left(\frac{a}{c}\right)}{\left(\frac{a+b}{c+d}\right)} = \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{c}{c+d}\right)} = RR$$

Table 3–9 Cross-Tabulation of a Defined Population by Exposure and Disease Development

<i>Exposure</i>	<i>Cases</i>	<i>Noncases</i>	<i>Total Population (Cases + Noncases)</i>
Present	a	b	a + b
Absent	c	d	c + d

Using again the local reaction/influenza vaccination investigation as an example (Table 3–4), a case-cohort study could be conducted using all cases and the total study population as the control group. The ratio of the exposure odds in cases ($\text{Odds}_{\text{exp cases}}$) to the exposure odds in the total study population ($\text{Odds}_{\text{exp pop}}$) yields the relative risk:

$$\text{OR}_{\text{exp}} = \frac{\text{Odds}_{\text{exp cases}}}{\text{Odds}_{\text{exp pop}}} = \frac{\left(\frac{650}{170}\right)}{\left(\frac{2570}{2410}\right)} = \frac{\left(\frac{650}{2570}\right)}{\left(\frac{170}{2410}\right)} = \frac{q_+}{q_-} = 3.59 = \text{RR}$$

where q_+ is the incidence in exposed and q_- the incidence in unexposed individuals.

In the estimation of the relative risk in the previous example, all cases and the total study population were included. Because unbiased sample estimates of $\text{Odds}_{\text{exp cases}}$ and $\text{Odds}_{\text{exp pop}}$ provide an unbiased estimate of the relative risk, however, a sample of cases and a sample of the total study population can be compared in a case-cohort study. For example, assuming no random variability, unbiased samples of 40% of the cases and 20% of the total population would produce the results shown in Table 3–10; the product of the division of the odds of exposure in the sample of cases by that in the study population sample can be shown to be identical to the relative risk obtained prospectively for the total cohort, as follows:

$$\text{OR}_{\text{exp}} = \frac{\frac{260}{68}}{\frac{514}{482}} = 3.59 = \text{RR}$$

Again, more commonly, because of the small number of cases relative to the study population size, case-cohort studies try to include all cases and a sample of the reference population.

One of the advantages of the case-cohort approach is that it allows direct estimation of the relative risk and thus does not have to rely on the rarity assumption. Another advantage is that because the control group is a sample of the total reference population, an unbiased estimate of the exposure prevalence (or distribution) needed for the estimation of Levin's population attributable risk (equation 3.10) can be obtained. A control group formed by an unbiased sample of the cohort also allows the assessment of relationships between different exposures or even between exposures and outcomes other than the outcome of interest in the cohort sample. To these analytical advantages, it can be added the practical advantage of the case-cohort design discussed in Chapter 1, Section 1.4.2, namely, the efficiency of

Table 3–10 Case-Cohort Study of the Relationship of Previous Vaccination to Local Reaction

<i>Previous Vaccination</i>	<i>Cases of Local Reaction</i>	<i>Cohort Sample</i>
Yes	260	514
No	68	482
Total	328	996

Note: Based on a random sample of the study population in Table 3–4, with sampling fractions of 40% for the cases and 20% for the cohort.

Source: Data from R Seltser, PE Sartwell, and JA Bell, A Controlled Test of Asian Influenza Vaccine in a Population of Families, *American Journal of Hygiene*, Vol 75, pp 112–135, © 1962.

selecting a single control group that can be compared with different types of cases identified on follow-up (e.g., myocardial infarction, stroke, and low extremity arterial disease).

In addition to these advantages connected with the choice of a sample of the cohort as the control group, there are other reasons why the traditional approach of selecting noncases as controls may not be always the best option. There are occasions in which excluding cases from the control group is logistically difficult and can add costs and participants' burden. For example, in diseases with a high proportion of a subclinical phase (e.g., chronic cholecystitis, prostate cancer), excluding cases from the pool of eligible controls (e.g., apparently healthy individuals) would require conducting more or less invasive and expensive examinations (e.g., contrasted X-rays, rectal exam). Thus, in these instances, a case-cohort approach might be indicated: selecting "controls" from the reference population irrespective of (i.e., ignoring) the possible presence of the disease (clinical or subclinical).

It is appropriate to conduct a case-cohort study only when a defined population (study base) from which the study cases originated can be identified, as when dealing with a defined cohort in the context of a prospective study. On the other hand, conducting case-cohort studies when dealing with "open" cohorts or populations at large requires assuming that these represented the source populations from which the cases originated (see Chapter 1, Section 1.4.2).

It should be also emphasized that when the disease is rare, the strategy of ignoring disease status when selecting controls would most likely result in few, if any, cases being actually included in the control group; thus, in practice, Equation 3.14 will be almost identical to Equation 3.12 because $(a + b) \approx b$ and $(c + d) \approx d$. For example, in the myocardial infarction/hypertension example shown in Table 3–3, the "case-cohort" strategy (selecting a 50% sample of cases and a 10% sample of total cohort as controls) would result in the following estimate of the odds ratio:

$$OR_{\text{exp}} = \frac{\text{Odds}_{\text{exp cases}}}{\text{Odds}_{\text{exp pop}}} = \frac{\left(\frac{90}{15}\right)}{\left(\frac{1000}{1000}\right)} = 6.00 = RR$$

In this same example, a case-"noncase" strategy would result in the following estimate:

$$OR_{\text{exp}} = \frac{\text{Odds}_{\text{exp cases}}}{\text{Odds}_{\text{exp noncases}}} = \frac{\left(\frac{90}{15}\right)}{\left(\frac{982}{997}\right)} = 6.09 = OR_{\text{dis}}$$

In other words, this situation is analogous to the situation discussed in Section 3.2.1 with regard to the similarity of the odds ratio and the relative risk when the disease is rare.

Influence of the Sampling Frame for Control Selection on the Parameter Estimated by the Odds Ratio of Exposure: Cumulative Incidence Versus Density Sampling

In addition to considering whether controls are selected from either noncases or the total study population, it is important to specify further the sampling frame for control selection. As discussed in Chapter 1 (Section 1.4.2), when controls are selected from a defined total

cohort, sampling frames may consist of either (1) individuals at risk when cases occur during the follow-up period (density sampling) or (2) the baseline cohort. The first alternative has been designated *nested case-control design* and the latter (exemplified by the “local reaction/influenza vaccination” analysis discussed previously here) *case-cohort design*.¹⁴ As demonstrated next, the nested case-control study and case-cohort designs allow the estimation of the rate ratio and the relative risk, respectively. An intuitive way to conceptualize which of these two parameters is being estimated by the odds ratio of exposure (i.e., rate ratio or relative risk) is to think of cases as the “numerator” and controls as the “denominator” of the absolute measure of disease frequency to which the parameter relates (see Chapter 2).

Density Sampling: The Nested Case-Control Design. As described previously (Section 1.4.2), the nested case-control design is based on incidence density sampling. It consists of selecting a control group that represents the sum of the subsamples of the cohort selected during the follow-up at the approximate times when cases occur (*risk set*) (Figure 1–20). These controls can be also regarded as a population sample “averaged” over all points in time when the events happen (see Chapter 2, Section 2.2.2) and could potentially include the case that defined the risk set. Therefore, the odds ratio of exposure thus obtained represents an estimate of the rate or density ratio (or relative rate or relative density). This strategy explicitly recognizes that there are losses (censored observations) during the follow-up of the cohort, in that cases and controls are chosen from the same reference populations excluding previous losses, thus matching cases and controls on duration of follow-up. When cases are excluded from the sampling frame of controls for each corresponding risk set, the odds ratio of exposure estimates the density odds ratio.

Selecting Controls From the Cohort at Baseline: The Case-Cohort Design. In the case-cohort design (also described in Chapter 1, Section 1.4.2), the case group is composed of cases identified during the follow-up period, and the control group is a sample of the total cohort at baseline (Figure 1–21). The cases and the sampling frame for controls can be regarded, respectively, as the type of numerator and denominator that would have been selected to calculate a probability based on the initial population, q . Thus, when these controls are selected, the odds ratio of exposure yields a ratio of the probability in exposed (q_+) to that in unexposed (q_-) individuals (i.e., the cumulative incidence ratio or relative risk) (Equation 3.14). Because the distribution of follow-up times in the sample of the initial cohort—which by definition includes those not lost as well as those subsequently lost to observation during follow-up—will be different from that of cases (whose “risk set”* by definition excludes previous losses), it is necessary to use survival analysis techniques to correct for losses that occur during the follow-up in a case-cohort study (see Section 2.2.1).

*Defined as the subset of the cohort members under observation at the time of each case’s occurrence (see Section 1.4.3, Figure 1-20).

It is also possible to exclude cases from the control group when sampling the cohort at baseline: that is, the sampling frame for controls would be formed by individuals who have remained disease-free through the duration of the follow-up. These are the persons who would have been selected as the denominator of the *odds based on the initial population*:

$$\left(\frac{q}{1-q} \right)$$

Thus, the calculation of the odds ratio of exposure when carrying out this strategy yields an estimate of the odds ratio of disease, (i.e., the ratio of the odds of developing the disease during the follow-up in individuals exposed and unexposed at baseline).

A summary of the effect of the specific sampling frame for control selection on the parameter estimated by the odds ratio of exposure is shown in Table 3–11.

Calculation of the Odds Ratio When There Are More Than Two Exposure Categories

Although the examples given so far in this chapter have referred to only two exposure categories, often more than two levels of exposure are assessed. Among the advantages of studying multiple exposure categories is the assessment of different exposure dimensions (e.g., “past” vs “current”) and of graded (“dose-response”) patterns.

In the example shown in Table 3–12, children with craniosynostosis undergoing craniectomy were compared with normal children in regard to maternal age.¹⁵ To calculate the odds ratio for the different maternal age categories, the youngest maternal age was chosen as the reference category. Next, for cases and controls separately, the odds for each maternal age category (vis-à-vis the reference category) were calculated (columns 4 and 5). The odds ratio is calculated as the ratio of the odds of each maternal age category in cases to the odds in controls (column 6). In this study, a graded and positive (direct) relationship was observed between maternal age and the odds of craniosynostosis.

When the multilevel exposure variable is ordinal (e.g., age categories in Table 3–12), it may be of interest to perform a trend test (see Appendix B).

Table 3–11 Summary of the Influence of Control Selection on the Parameter Estimated by the Odds Ratio of Exposure in Case-Control Studies Within a Defined Cohort

<i>Design</i>	<i>Population Frame for Control Selection</i>	<i>Exposure Odds Ratio Estimates</i>
Case-cohort	Total cohort at baseline (Total cohort at baseline minus cases that develop during follow-up)	Cumulative incidence ratio (relative risk) (Probability odds ratio)
Nested case-control	Population at approximate times when cases occur during follow-up (Population during follow-up minus cases)	Rate (density) ratio (Density odds ratio)

Table 3–12 Distribution of Cases of Craniosynostosis and Normal Controls According to Maternal Age

Maternal Age (Years) (1)	Cases (2)	Controls (3)	Odds of Specified Maternal Age vs Reference in Cases (4)	Odds of Specified Maternal Age vs Reference in Controls (5)	Odds Ratio (6) = (4)/(5)
<20*	12	89	12/12	89/89	1.00*
20–24	47	242	47/12	242/89	1.44
25–29	56	255	56/12	255/89	1.63
>29	58	173	58/12	173/89	2.49

*Reference category.

Source: Data from BW Alderman et al, An Epidemiologic Study of Craniosynostosis: Risk Indicators for the Occurrence of Craniosynostosis in Colorado, *American Journal of Epidemiology*, Vol 128, pp 431–438, © 1988, The Johns Hopkins Bloomberg School of Public Health.

3.4.2 Attributable Risk in Case-Control Studies

As noted previously (Section 3.2.2), percent attributable risk in the exposed can be obtained in traditional case-control (case–noncase) studies when the odds ratio is a reasonable estimate of the relative risk by replacing its corresponding value in Equation 3.6:

[Equation 3.15]

$$\%AR_{\text{exp}} = \left(\frac{OR - 1.0}{OR} \right) \times 100$$

In studies dealing with preventive interventions, the analogous measure is efficacy (see Section 3.2.2, Equation 3.7). The fact that the odds ratio is usually a good estimate of the relative risk makes it possible to use Equation 3.15 in case-control studies of the efficacy of an intervention such as screening.¹⁶

The same reasoning applies to the use of case-control studies to estimate the population attributable risk using a variation of Levin's formula:

[Equation 3.16]

$$\%Pop AR = \frac{p_e^* \times (OR - 1)}{p_e^* \times (OR - 1) + 1} \times 100$$

In Equation 3.16, the proportion of exposed subjects in the reference population (p_e) is represented as p_e^* because in the context of a case-control study this is often estimated by the exposure prevalence among controls. Such assumption is appropriate as long as the disease is rare and the control group is reasonably representative of all noncases in the reference population. Obviously, if a case-cohort study is conducted, the rarity assumption is not needed (Section 3.4.1), as both the relative risk and the exposure prevalence can be directly estimated.

As shown by Levin and Bertell,¹⁷ if the odds ratio is used as the relative risk estimate, Equation 3.16 reduces to a simpler equation:

$$\% \text{ Pop AR} = \frac{p_{e/case} - p_{e/control}}{1.0 - p_{e/control}} \times 100$$

where $p_{e/case}$ represents the prevalence of exposure among cases—that is, $a/(a + c)$ in Table 3–9—and $p_{e/control}$ represents the prevalence of exposure among controls—that is, $b/(b + d)$ in Table 3–9.

3.5 ASSESSING THE STRENGTH OF ASSOCIATIONS

The values of the measures of association discussed in this chapter are often used to rank the relative importance of risk factors. However, because risk factors vary in terms of their physiologic modus operandi as well as their exposure levels and units, such comparisons are often unwarranted. Consider, for example, the absurdity of saying that systolic blood pressure is a more important risk factor for myocardial infarction than total cholesterol, based on comparing the odds ratio associated with a 50-mmHg increase in systolic blood pressure with that associated with a 1-mg/dL increase in total serum cholesterol. In addition, regardless of the size of the units used, it is hard to compare association strengths, given the unique nature of different risk factors.

An alternative way to assess the strength of the association of a given risk factor with an outcome is to estimate the *exposure intensity* necessary for that factor to produce an association of the same magnitude as that of well-established risk factors or vice-versa. For example, Tverdal et al.¹⁸ evaluated the level of exposure of four well-known risk factors for coronary heart disease mortality necessary to replicate the relative risk of 2.2 associated with a coffee intake of nine or more cups per day. As seen in Exhibit 3–1, a relative risk of 2.2 corresponds to smoking about 4.3 cigarettes per day or having an increase in systolic blood pressure of about 6.9 mm Hg, and so on.

Exhibit 3–1 A Possible Way to Describe the Strength of an Association Between a Risk Factor and an Outcome

A relative risk of 2.2 for coronary heart disease mortality comparing men drinking 9+ or more cups of coffee per day versus < one cup per day corresponds to 18

Smoking:	4.3 cigarettes/day
Systolic blood pressure:	6.9 mm/Hg
Total serum cholesterol:	0.47 mmol/L
Serum high-density lipoprotein:	–0.24 mmol/L

Source: Data from A Tverdal et al, Coffee Consumption and Death from Coronary Heart Disease in Middle-Aged Norwegian Men and Women, *British Medical Journal*, Vol 300, pp 566–569, © 1990.

Exhibit 3–2 Cross-Sectionally Determined Mean Intima-Media Thickness (IMT) of the Carotid Arteries (mm) by Passive Smoking Status in Never-Active Smokers, the Atherosclerosis Risk in Communities Study, 1987–1989

	Passive Smoking Status in Never-Active Smokers		Estimated Increase by Year of Age
	Absent (n = 1,774)	Present (n = 3,358)	
Mean IMT (mm)→	0.700	0.711	0.011
Age-equivalent excess attributable to passive smoking: (0.711 – 0.700)/0.011 = 1 year			
<i>Source:</i> Data from G. Howard et al, Active and Passive Smoking Are Associated with Increased Carotid Wall Thickness. The Atherosclerosis Risk in Communities Study, Archives of Internal Medicine, Vol 154, pp 1277–1282, © 1994, American Medical Association.			

Another example comes from a study by Howard et al.,¹⁹ who evaluated the cross-sectional association between passive smoking and subclinical atherosclerosis measured by B-mode ultrasound-determined intimal-medial thickness of the carotid artery walls. Because passive smoking had not been studied previously in connection with directly visualized atherosclerosis, its importance as a risk factor was contrasted with that of a known atherosclerosis determinant, age (Exhibit 3–2). As seen in the exhibit, the cross-sectional association between passive smoking and atherosclerosis is equivalent to an age difference of 1 year. That is, assuming that the cross-sectional association adequately represents the prospective relationship between age and atherosclerosis and that the data are valid, precise, and free of confounding, the average thickness of the carotid arteries of passive smokers looks like that of never smokers who are 1 year older. This inference was extended by Kawachi and Colditz,²⁰ who on the basis of data from Howard et al.'s study, estimated that the change in intimal-medial thickness related to passive smoking would result in an increase in the risk of clinical cardiovascular events equivalent to an increment of 7 mm Hg of systolic blood pressure, or 0.7 mmol/L of total cholesterol—thus, not negligible.

REFERENCES

1. Lilienfeld DE, Stolley, PD. *Foundations of Epidemiology*, 3rd ed. New York: Oxford University Press; 1994.
2. Gordis L. *Epidemiology*. Philadelphia: Elsevier Saunders; 2004.
3. Seltser R, Sartwell PE, Bell JA. A controlled test of Asian influenza vaccine in a population of families. *Am J Hygiene*. 1962;75:112–135.
4. Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol*. 1988;128:1185–1197.
5. Rothman K, Greenland S. *Modern Epidemiology*, 2nd ed. Philadelphia: Lippincott-Raven; 1998.
6. Levin ML. The occurrence of lung cancer in man. *Acta Unio Int Contra Cancrum*. 1953;9:531–541.
7. Walter SD. The estimation and interpretation of attributable risk in health research. *Biometrics*. 1976;32:829–849.
8. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health*. 1998;88:15–19.

9. Walter SD. Effects of interaction, confounding and observational error on attributable risk estimation. *Am J Epidemiol.* 1983;117:598–604.
10. Cornfield J. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst.* 1951;11:1269–1275.
11. Vilar-Compte D, Jacquemin B, Robles-Vidal C, Volkow P. Surgical site infections in breast surgery: Case-control study. *World J Surg.* 2004;28:242–246.
12. Valeri A, Fromont G, Sakr W, et al. High frequency of allelic losses in high-grade prostate cancer is associated with biochemical progression after radical prostatectomy. *Urol Oncol.* 2005;23:87–92.
13. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika.* 1986;73:1–11.
14. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: A critical comparison. *Am J Epidemiol.* 1990;131:169–176.
15. Alderman BW, Lammer EJ, Joshua SC, et al. An epidemiologic study of craniosynostosis: Risk indicators for the occurrence of craniosynostosis in Colorado. *Am J Epidemiol.* 1988;128:431–438.
16. Coughlin SS, Benichou J, Weed DL. Attributable risk estimation in case-control studies. *Epidemiol Rev.* 1994;16:51–64.
17. Levin ML, Bertell R. RE: “simple estimation of population attributable risk from case-control studies.” *Am J Epidemiol.* 1978;108:78–79.
18. Tverdal A, Stensvold I, Solvoll K, Foss OP, Lund-Larsen P, Bjartveit K. Coffee consumption and death from coronary heart disease in middle aged Norwegian men and women. *Br Med J.* 1990;300:566–569.
19. Howard G, Burke GL, Szklo M, et al. Active and passive smoking are associated with increased carotid wall thickness: The Atherosclerosis Risk in Communities Study. *Arch Intern Med.* 1994;154:1277–1282.
20. Kawachi I, Colditz GA. Invited commentary: confounding, measurement error, and publication bias in studies of passive smoking. *Am J Epidemiol.* 1996;144:909–915.

EXERCISES

1. The following results were obtained in an occupational cohort study of the risk of cancer associated with exposure to radiation.

<i>Radiation Dose (rem)</i>	<i>Total Size of the Cohort at Baseline</i>	<i>New Cancer Cases</i>	<i>Cumulative Cancer Incidence*</i>	<i>Relative Risk</i>	<i>Odds Ratio (Comparing Cases to Noncases)</i>	<i>Odds Ratio (Comparing Cases to Total Population)</i>
0–0.99	3642	390				
1–4.99	1504	181				
5+	1320	222				

*Assume no losses to follow-up.

- Fill in the empty cells in this table. For the calculation of relative risks and odds ratios, use the lowest radiation category as the reference category (show your calculations).
 - How do you explain the difference (or the similarity) between each of the two types of odds ratios calculated here and the corresponding relative risks?
 - Assuming that an association between exposure and disease is actually present, which of the following statements is true? Why?
 - The odds ratio is always smaller in absolute value than the relative risk.
 - The odds ratio is always bigger in absolute value than the relative risk.
 - The odds ratio is always closer to 1 than the relative risk.
 - The odds ratio is always farther away from 1 than the relative risk.
 - In looking at the progressively higher relative risks (or odds ratios) with increasing radiation dose, which traditional causality criterion is being assessed?
2. A cohort study to examine the relationship of inflammatory markers (such as interleukin-6 and C-reactive protein) to incident dementia was conducted within the Rotterdam Study cohort ($n = 6713$).¹ A random cohort sample of the total cohort at baseline ($n = 727$) and the 188 individuals who developed dementia on follow-up were compared. Serum inflammatory markers were measured in cases and in the random sample.
- Which type of study have the authors conducted?
 - By dividing the odds of exposure in cases to that in controls, which measure of association is obtained?
 - If the authors wished to study the relationship of inflammatory markers to stroke, could they use the same control group? Why or why not?
 - The relative risk of dementia associated with an IL-6 value in the highest quintile compared with that in the lowest quintile was found to be about 1.9. Assuming no random variability and that the relative risks of the second, third, and fourth quintiles compared with the lowest quintile were close to 1.0, calculate the proportion

¹Engelhart MJ, Geerlings MJ, Meijer J, et al. Inflammatory proteins in plasma and risk of dementia: The Rotterdam Study. *Arch Neurol*. 2004;61:668–672.

of dementia incidence in the population that might be explained by values in the highest quintile.

3. A recent case-control study² assessed the relationship of hepatitis C virus (HCV) infection to B-cell non-Hodgkin lymphomas (B-NHL). Cases were identified in the hematology department wards of 10 hospitals located in different cities throughout Italy. The control group consisted of patients admitted to other departments of the same hospitals (e.g., ophthalmology, general surgery, internal medicine). For both cases and controls, only patients with newly diagnosed diseases were included in the study.
- a. What type of study have the authors conducted?

The numbers of cases and controls as well as the numbers who tested positive for HCV by age (55 or less, and more than 55 years old) are seen in this table:

Age (Years)	Cases		Controls	
	Number	HCV Positive	Number	HCV Positive
≤ 55	163	18	231	6
> 55	237	52	165	16

- b. Calculate the exposure odds ratio reflecting the HCV/B-NHL association for each age group.
- c. Describe in words the meaning of the odds ratio for age group > 55 years.

²Mele A, Pulsoni A, Bianco E, et al. Hepatitis C virus and B-cell non-Hodgkin lymphomas: An Italian multi-center case-control study. *Blood*. 2003;102:996–999.

