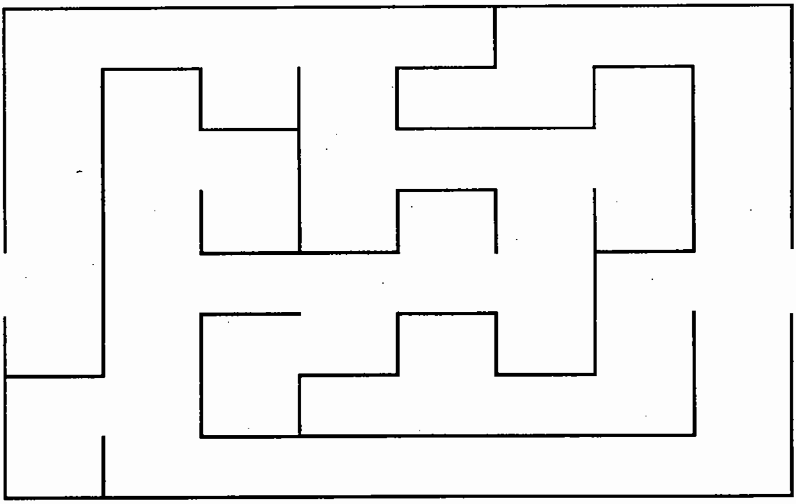
9

Confounding



Confounding results from a mixture of effects in a single estimate. An uneven distribution of baseline risks across comparison groups results in a confounded estimate of the differences between the groups. The stratified analyses that were introduced in the last chapter to deal with confounding assumed that the information necessary for segregating study subjects into strata would be available, and a similar requirement holds for mathematical modeling techniques. Often the crucial data are absent. This chapter addresses the strength of bias introduced by confounding when confounding factors are ignored. The approach is entirely theoretical; readers with low tolerance for the abstract may wish to pass directly to the section on Implications at the end.

The Apparent Relative Risk

Let E stand for an exposure that is either present or absent, and let $Pr(E)$ be the prevalence of one level of E ("exposed") in a population. Let C stand for a covariate characteristic or exposure, also with only two levels, and $Pr(C)$ be the prevalence of one level of C ("possessing the covariate") in the same population. The population prevalences e, f, g, and h displayed in Table 9.1 describe the joint distribution of the population over E and C.

Table 9.1 Prevalences of exposure and covariate

	Exposure		
Covariate	Present	Absent	Total
Present	*e*	*f*	*Pr(C)*
Absent	*g*	*h*	*1-Pr(C)*
Total	*Pr(E)*	*1-Pr(E)*	*1*

Let the population described in Table 9.1 be observed without loss to follow-up for sufficient time to allow the appearance of disease. Assume that the probability of disease is unrelated to exposure, but is a function of covariate status. For population subgroups in which the covariate is present or absent, assume that the probabilities of disease acquisition are R_C or $R_{\bar{C}}$, respectively.

A development exactly analogous to the one that follows can be laid out for cumulative incidences, incidence rates, or hazards, any of which could therefore be substituted for the word "probability" below.

The total probability of disease in exposed persons is

$$R_E = \frac{R_C e + R_{\bar{C}} g}{e + g}$$

That in persons not exposed is

$$R_{\bar{E}} = \frac{R_C f + R_{\bar{C}} h}{f + h}$$

The apparent relative risk, comparing exposed to unexposed segments of the population is

Apparent $RR(E) =$

$$\frac{R_E}{R_{\bar{E}}} = \frac{R_C e + R_{\bar{C}} g}{R_C f + R_{\bar{C}} h} \; \frac{1 - Pr(E)}{Pr(E)} \qquad [1]$$

Since exposure by definition is not associated with any excess probability of disease, the apparent relative risk is a direct measure of confounding. The first term in the expression above is the exposure odds in cases, the second is the reciprocal of exposure odds in the study population. Since the exposure odds in the control series of a case-control study is a consistent estimate of that in the population in which the cases occurred, the arguments developed here apply to case-control studies as well as to cohort studies.

Define the relative risk comparing persons in whom the covariate is present to those in whom it is absent as

$$RR(C) = \frac{R_C}{R_{\bar{C}}}$$

For fixed overall prevalences of exposure and covariate in the study population, the quantities f, g, and h can be rewritten as functions of e, $Pr(E)$, and $Pr(C)$.

$$f = Pr(C) - e \qquad g = Pr(E) - e$$

$$h = 1 - Pr(C) - Pr(E) + e$$

On the basis of the preceding expressions, the apparent relative risk for exposure can be rewritten as

Apparent $RR(E) =$

$$\frac{e(RR(C) - 1) + Pr(E)}{(Pr(C) - e)(RR(C) - 1) - Pr(E) + 1} \; \frac{1 - Pr(E)}{Pr(E)} \qquad [2]$$

The import of Equation 2 is that the degree of confounding introduced into a 2x2 table by the presence of a third, disease-causing factor depends upon the strength of the association between the third factor and disease ($RR(C)$), on the prevalence of exposure ($Pr(E)$), on the prevalence of the third factor ($Pr(C)$), and on the proportion of the study population in which the confounding characteristic and exposure occur together in the same people (e).

Equation 1 can also be written as

Apparent $RR(E)=$

$$\frac{RR(C)e+g}{RR(C)f+h} \; \frac{1-Pr(E)}{Pr(E)} \quad [3]$$

The prevalences of the confounding variable in the exposed and nonexposed segments of the population are

$$Pr(C \mid E)=\frac{e}{Pr(E)} \qquad Pr(C \mid \bar{E})=\frac{f}{1-Pr(E)}$$

Furthermore

$$Pr(\bar{C} \mid E)=\frac{g}{Pr(E)} \qquad Pr(\bar{C} \mid \bar{E})=\frac{h}{1-Pr(E)}$$

By substitution into Equation 3,

Apparent $RR(E)=$

$$\frac{Pr(C \mid E)RR(C)+Pr(\bar{C} \mid E)}{Pr(C \mid \bar{E})RR(C)+Pr(\bar{C} \mid \bar{E})} \quad [4]$$

Special Cases and Limits for Apparent RR(E)

The right hand side of Equation 4 is the ratio of two weighted averages of the quantities $RR(C)$ and 1. (The weights are different, being $P(C \mid E)$ and $P(\bar{C} \mid E)$ in the numerator and $P(C \mid \bar{E})$ and $P(\bar{C} \mid \bar{E})$ in the denominator.) When $RR(C)=1$, then the Apparent $RR(E)=1$. When $RR(C)>1$, the numerator quantity in the ratio is less than $RR(C)$, and the denominator is greater than 1. It follows that

$$\text{Apparent } RR(E) < RR(C) \quad [5]$$

That the relative risk that could be ascribed to confounding must be less than the relative risk associated with the confounding factor was first observed by Cornfield, Haenszel, and others.[67]

67. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: Recent evidence and a discussion of some questions. J Nat Cancer Inst 1959;22:173-203 (Appendices A and B)

If E and C are independent, in the sense that

$$e=Pr(E)Pr(C) \quad [6]$$

then from Equation 2 the Apparent $RR(E)=1$. That is to say, there is no confounding.

From Equation 4 it follows when $RR(C)>1$ and $P(C \mid E) > P(C \mid \bar{E})$ that

$$\text{Apparent } RR(E) < \frac{Pr(C \mid E)}{Pr(C \mid \bar{E})} \quad [7]$$

I will refer to Equation 7 as "the prevalence inequality." The prevalence inequality holds that the apparent relative risk associated with exposure is less than the ratio of the prevalence of the confounding variable among the exposed to the prevalence of the confounding variable among the nonexposed. When the prevalence of the confounding variable in the study population approaches 100 percent, the degree of confounding approaches nil (Apparent $RR(E) = 1$). When the prevalence of the confounding variable is zero, the right hand side of Equation 7 is undefined, but it follows from Equation 4 that the Apparent $RR(E) = 1$ for all values of $RR(C)$ and of $Pr(E)>0$.

Using the notation of Table 9.1, we can write the Exposure-Covariate Odds Ratio ($ECOR$) as

$$ECOR=\frac{eh}{fg}$$

Independence between E and C, in the sense defined by Equation 6, is equivalent to $ECOR=1$. When $ECOR$ is greater than 1, it always exceeds the corresponding ratio of confounder prevalences in the exposed and nonexposed, which in turn exceeds one. That is

$$1<\frac{Pr(C \mid E)}{Pr(C \mid \bar{E})}<ECOR$$

which in combination with Equation 7 implies that

$$1<\text{Apparent } RR(E)<ECOR \quad [8]$$

Determinants of the Apparent RR(E)

The results of the preceding section place limits on the degree of confounding that can result from values of *Pr(E)*, *Pr(C)*, *RR(C)*, and the *ECOR*, but they do not provide information on the manner in which the confounding varies as a function of these together. We will approach this question graphically.

Conditionally upon the margins of Table 9.1 and the *ECOR*, *e* can be found as the solution of a quadratic equation. Substitution of the derived value for *e* into Equation 2 yields an expression relating *RR(E)* to *ECOR*, *Pr(C)*, *Pr(E)*, and *RR(C)*. The dependence of the Apparent *RR(E)* on these terms is graphed in Figures 9.1 - 9.3.

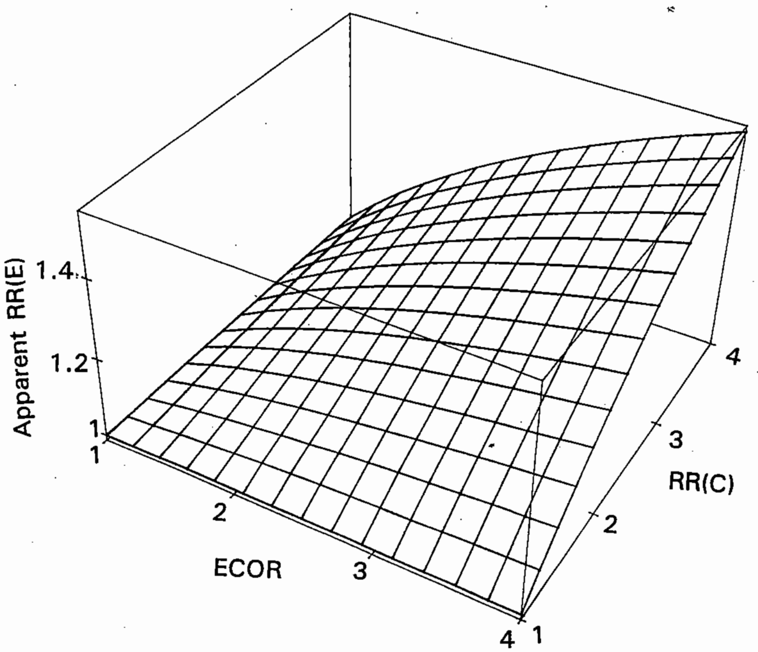


Figure 9.1 Dependence of the apparent relative risk associated with exposure on the relative risk associated with the covariate and on the exposure-covariate odds-ratio for *Pr(E)*=0.2 and *Pr(C)*=0.2

Figure 9.1 displays the Apparent *RR(E)* as a function of *RR(C)* and the *ECOR* for *Pr(E)=Pr(C)*=0.2. *ECOR* and *RR(C)* both range in the figure from 1 to 4. Where *ECOR=RR(C)*=1, the Apparent *RR(E)*=1. So long as *ECOR* equals unity, the Apparent *RR(E)* remains unity. The same holds true so long as *RR(C)* equals unity. For values of *ECOR* and *RR(C)* greater than their respective baselines, the Apparent *RR(E)* rises with increasing association of exposure and covariate and does so with increasing slope as *RR(C)* increases. Not shown in the figure is the situation when either *ECOR* or *RR(C)* falls below one. Then the confounding is in a negative direction, so that the Apparent *RR(E)* is less than one. When both *ECOR* and *RR(C)* are less than one, the confounding is again positive.

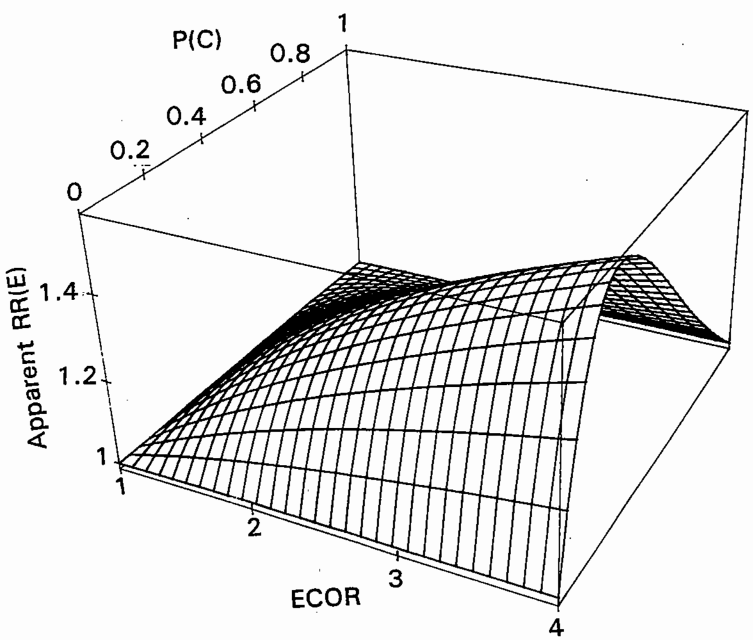


Figure 9.2 Dependence of the apparent relative risk associated with exposure on the overall prevalence of the covariate and on the exposure-covariate odds-ratio for *Pr(E)*=0.2 and *RR(C)*=4

Figure 9.2 presents the surface described by *Pr(C)*, *ECOR*, and *RR(E)* for *Pr(E)*=0.2 and *RR(C)*=4. The values of *Pr(C)* in Figure 9.2 range from 0 to 1, and those of the *ECOR* from 1 to 4. Here as in Figure 9.1, the Apparent *RR(E)* equals one when there is no association between *E* and *C*, that is, when *ECOR*=1. For larger values of *ECOR*, the degree of confounding (as measured by the distance between *RR(E)* and its null value of unity) rises from none when *Pr(C)*=0 to a maximum (at *Pr(C)*=0.2) and then falls back to none at *Pr(C)*=1. For values of *ECOR* less than one (not shown in Figure 9.2), the Apparent *RR(E)* falls below unity, the more so as *ECOR* approaches 0.

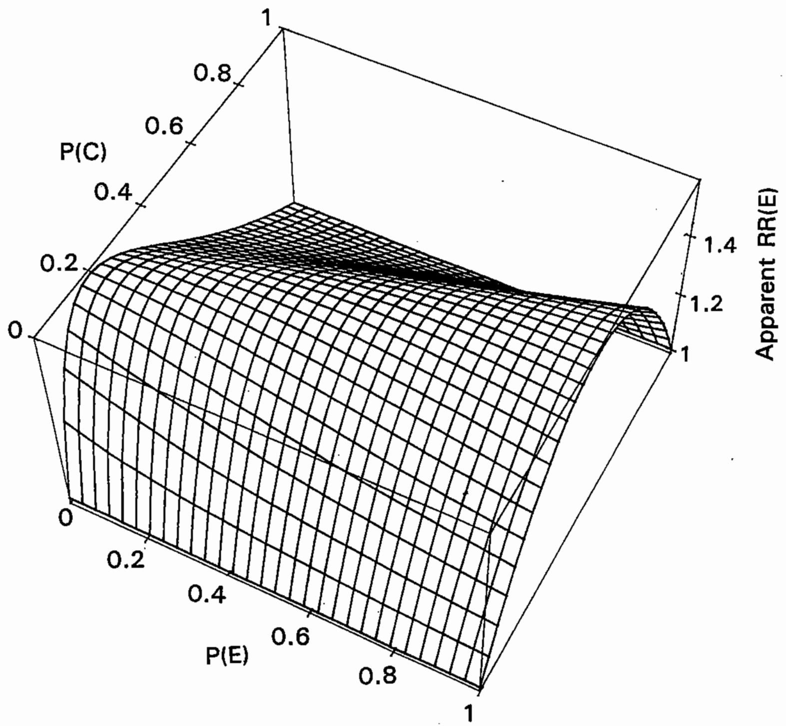


Figure 9.3 Dependence of the apparent relative risk associated with exposure on the overall prevalences of exposure and covariate for *ECOR*=4 and *RR(C)*=4

The relation between the prevalence inequality and Figure 9.2 can be seen most clearly at the upper limit values for *Pr(C)*. When the prevalence of the confounding variable approaches 1, then the ratio of confounding variable prevalence in exposed to that in nonexposed persons must also approach 1, and the Apparent *RR(E)* approaches 1 as well. As *Pr(C)* nears zero, the confounding effect approaches the null in what appears to be a smooth manner. At *ECOR*=4, the prevalence of the covariate characteristic that results in the greatest distortion of the apparent relative risk associated with exposure is 0.2.

Figure 9.3 presents the dependence of the Apparent *RR(E)* on *Pr(E)* and *Pr(C)* for the case of *ECOR*=5 and *RR(C)*=4. This figure comprises an expansion of the foremost slice of Figure 9.2, for which *Pr(E)* was set equal to 0.2. The prevalence of the covariate characteristic associated with the maximum degree of confounding can be seen to increase as the prevalence of exposure rises. It does so in a linear fashion, increasing from *Pr(C)*=0.2 when *Pr(E)* is near zero to *Pr(C)*=0.5 when *Pr(E)* is near one.

Implications

The degree of confounding is a joint function of the prevalence of exposure (*Pr(E)*), the prevalence of the covariate (*Pr(C)*), the relative risk for disease associated with the covariate (*RR(C)*), and the exposure-covariate odds ratio (*ECOR*). None of these terms can be considered in isolation from the others if a quantitative interpretation is the goal. Qualitatively,

1. Confounding increases as the strength of the association between disease and the covariate increases (Figure 9.1), but is always less than the relative risk for disease associated with the covariate (Equation 5).

2. Confounding increases as the strength of the association between exposure and the covariate increases (Figures 9.1 and 9.2) but is always less than the ratio of the prevalence of covariate in the exposed to that in the nonexposed group (Equation 7). Equivalently, confounding is always less than the odds ratio between exposure and the covariate (Equation 8).

3. When the covariate is associated positively with exposure and with probability of disease, then the degree of confounding rises smoothly as the prevalence of the covariate increases from zero, reaches a maximum, and then declines toward nil as the covariate prevalence increases further towards one (Figures 9.2 and 9.3).