$$IR = \frac{103}{11,598}$$

$$= 0.008881 \text{ cases per person year}$$

$$\mathrm{Var}(IR) = \frac{103}{(11,598)^2}$$

$$= 7.657 \times 10^{-7}$$

The 95 percent confidence bounds are

$$\text{lower} = 0.00881 - 1.96\sqrt{7.657 \times 10^{-7}}$$

$$= 0.00717 \text{ cases per person year}$$

$$\text{upper} = 0.00881 + 1.96\sqrt{7.657 \times 10^{-7}}$$

$$= 0.01060 \text{ cases per person year}$$

All the techniques for estimating incidence rate differences and summary incidence rate changes over strata are precisely analogous to those presented earlier for risks in closed cohort studies. The sole differences are to introduce incidence rate estimates (x/P) in the place of cumulative incidence estimates (x/N) and variance estimates for incidence rates (x/P^2) in the place of variance estimates for cumulative incidences $(x(N-x)/N^3)$ in all the formulae.

It is common practice to examine the ratios of incidence rates in open cohort studies; this is the result of an empirical observation in chronic disease research, that incidence rate ratios tend to be more constant from study to study or from stratum to stratum of a single study than are rate differences. The easiest way to account for variability in incidence ratio estimates is on a logarithmic scale, in which the ratio estimate can be examined as a difference between the logarithms of the component incidence rate estimates. All of the foregoing procedures can then be adapted to confidence interval estimation on the log scale. Estimates, once obtained, are transformed back to the natural scale by exponentiation.

Denote the natural logarithm of the incidence rate estimate as $\ln(x/P)$. The variance of this quantity is approximately $1/x$. The variance of the logarithm of the incidence rate ratio is the sum of the variances of the logarithms of the component incidence rates.

Thus, to compare the lung cancer rate at 30-34 years after first exposure to that 20-24 years after first exposure, the procedure would be as follows:

$$RR = \left(\frac{103}{11,598}\right) / \left(\frac{57}{31,268}\right)$$

$$= 4.87$$

$$\ln(RR) = \ln(4.87)$$

$$= 1.5834$$

$$\mathrm{Var}[\ln(RR)] = \frac{1}{103} + \frac{1}{57}$$

$$= 0.02725$$

The 95 percent confidence bounds for the logarithm of the ratio are

$$\text{lower} = 1.5834 - 1.96\sqrt{0.02725}$$

$$= 1.260$$

$$\text{upper} = 1.5834 + 1.96\sqrt{0.02725}$$

$$= 1.907$$

The 95 percent confidence bounds for the ratio are then

$$\text{lower} = \exp(1.260)$$

$$= 3.52$$

$$\text{upper} = \exp(1.907)$$

$$= 6.73$$

The ratio of lung cancer mortality rates for insulation workers 30-34 years from first exposure to asbestos to that 20-24 years from first exposure was approximately 4.9, with 95 percent confidence bounds of 3.5 and 6.7.

Stratified analysis. Two techniques are commonly used for summarizing incidence rate ratios across strata. Consider the hypothetical data in Table 8.1. The first subscript on the symbols displayed indicates the presence (1) or absence (0) of exposure, and the second subscript indicates the age group: 50-54 (1) or 55-59 (2).

Table 8.1 Lung cancer mortality in men exposed and unexposed to asbestos (hypothetical data)

	Age Group			
	50 - 54		55 - 59	
	Quantity	Symbol	Quantity	Symbol
Exposed				
Person Years	1,000	P_{11}	500	P_{12}
Cases	40	x_{11}	40	x_{12}
Unexposed				
Person Years	10,000	P_{01}	15,000	P_{02}
Cases	100	x_{01}	200	x_{02}

The summary technique most used in occupational health studies is to compare the number of cases of disease in the exposed group to that which would have been expected among the exposed, had the incidence rates observed in unexposed persons applied to those exposed. This expectation is obtained by multiplying the person years at risk in each stratum of the exposed group by the incidence rates observed in the unexposed group, and summing over all strata. Thus, in exposed workers,

$$\begin{aligned}\text{Observed} &= x_{11} + x_{12}\\ &= 40 + 40\\ &= 80\end{aligned}$$

$$\begin{aligned}\text{Expected} &= P_{11}\frac{x_{01}}{P_{01}} + P_{12}\frac{x_{02}}{P_{02}}\\ &= 1,000\left(\frac{100}{10,000}\right) + 500\left(\frac{200}{15,000}\right)\\ &= 16.67\end{aligned}$$

The ratio of observed to expected cases is designated (for historical reasons) as "the" *standardized mortality (or morbidity) ratio* (*SMR*). The ratio is standardized because it is algebraically identical to the ratio of age-standardized incidence rates in exposed and unexposed study subjects, taking for each the age distribution among exposed as the standard. In the present case

$$\begin{aligned}SMR &= \frac{\text{Obs}}{\text{Exp}} = \frac{80}{16.67}\\ &= 4.80\end{aligned}$$

In practice, the *SMR* is rarely used except when the unexposed population is very large (most commonly a geographically defined population that encompasses the exposed persons). When the number of events is large in every stratum of the comparison population, the variance of the *SMR* is approximately Obs/Exp^2. In the present example

$$\begin{aligned}\text{Var}(SMR) &= \frac{\text{Obs}}{\text{Exp}^2} = \frac{80}{(16.67)^2}\\ &= 0.2880\end{aligned}$$

The 95 percent confidence bounds can be obtained therefore as

$$\begin{aligned}\text{lower} &= 4.800 - 1.96\sqrt{0.2880}\\ &= 3.75\end{aligned}$$

$$\begin{aligned}\text{lower} &= 4.800 + 1.96\sqrt{0.2880}\\ &= 5.85\end{aligned}$$

When the sole source of stratum to stratum variation is thought to be random error, an incidence rate ratio estimate whose form is due to Mantel and Haenszel[60] is obtainable by summing the quantities

$$A_i = \frac{x_{1i}P_{0i}}{P_{1i} + P_{0i}} \qquad B_i = \frac{x_{0i}P_{1i}}{P_{1i} + P_{0i}}$$

over the strata, indexed here by i, and dividing the sums. In the present example,

60. The use of the procedure in open cohort studies was first proposed by Kenneth Rothman and John Boice. (Rothman KJ, Boice JR. Epidemiologic Analysis with a Programmable Calculator, NIH Publication No. 79-1649, Washington, 1979) The rationale was developed by David Clayton. (Clayton DG. The analysis of prospective studies of disease etiology. Commun Statist 1982;A11:2129-2155)

$$A = \sum_i A_i = \frac{(40)(10{,}000)}{10{,}000 + 1{,}000} + \frac{(40)(15{,}000)}{15{,}000 + 500}$$

$$= 75.07$$

$$B = \sum_i B_i = \frac{(100)(1{,}000)}{10{,}000 + 1{,}000} + \frac{(200)(500)}{15{,}000 + 500}$$

$$= 15.54$$

(When a variable, here i, appears below a sigma without any indication of the range of summation, the summation is taken over all possible values of the variable. In the present example, the possible values for i are 1 and 2.) The summary estimate, known as the *Mantel-Haenszel* estimate of the ratio is

$$RR_{MH} = \frac{A}{B}$$

$$= 4.831$$

The variance of the logarithm of the Mantel-Haenszel estimator is obtained by taking a further sum,

$$C = \sum_i (x_{1i} + x_{0i}) P_{1i} P_{0i} / (P_{1i} + P_{0i})^2$$

The variance estimate is then[61]

$$\mathrm{Var}[\ln(RR_{MH})] \doteq \frac{C}{AB}$$

Here,

$$C = (40 + 100)(1{,}000)(10{,}000)/(1{,}000 + 10{,}000)^2$$

$$+ (40 + 200)(500)(15{,}000)/(500 + 15{,}000)^2$$

$$= 19.06$$

and

$$\mathrm{Var}[\ln(RR_{MH})] \doteq \frac{19.06}{(75.07)(15.54)} = 0.01634$$

61. Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data. Biometrics 1985;41:55-68

The natural logarithm of the hazard ratio estimate is ln(4.831) = 1.575. Proceeding as before, the 95 percent confidence interval to the logarithm of the incidence rate ratio can be found to be 1.325 to 1.826, yielding a corresponding interval on the ratio scale of 3.8 to 6.2.

When the ratios observed in the strata being summarized are not very disparate, when the amounts of person time under study in each exposure group do not vary greatly across strata, or when the person time of the unexposed group is vastly larger than that of the exposed in each stratum, the *SMR* and the Mantel-Haenszel estimate of the incidence rate ratio will be very close to one another, and there is little practical distinction to be made between the two. In the last situation, the closeness of the Mantel-Haenszel estimator to the SMR arises from the fact that both procedures give weight in approximate proportion to the information contained in the exposed half of each stratum.[62] The theory underlying their respective derivations leads to a choice of the *SMR* whenever the stratum-specific hazard ratios are inconstant, and to the Mantel-Haenszel estimator when they do not vary greatly.

Case-Control Studies

Random Error. Analysis of the variability of odds ratios and of more complex functions involving odds ratios is almost always carried out on a logarithmic scale. Expressed as a logarithm, the odds ratio has a simple additive structure:

$$\ln(RR) = \ln\left(\frac{x_1 y_0}{y_1 x_0}\right)$$

$$= \ln(x_1) + \ln(y_0) - \ln(y_1) - \ln(x_0)$$

Here as before "ln(x)" stands for the natural logarithm of x.

An estimate of the variance of the logarithm of a count is given by[63]

62. Walker AM. Small sample properties of some estimators of a common hazard ratio. Appl Statistics 1985;34:42-8

63. The capital X in the formula is the random variable, of which the value x is the observed value.

$$\mathrm{Var}[\ln(X)] \doteq \frac{1}{x}$$

Let O stand for the parameter of which the odds ratio is an estimate. Since the variance of the sum or difference of terms equals the sum of the variances of the terms, we have an estimate of the variance of the logarithm of an odds ratio.

$$\mathrm{Var}[\ln(O)] = \mathrm{Var}[\ln(X_1)] + \mathrm{Var}[\ln(X_0)] + \mathrm{Var}[\ln(Y_1)] + \mathrm{Var}[\ln(Y_0)]$$

$$\doteq \frac{1}{x_1} + \frac{1}{x_0} + \frac{1}{y_1} + \frac{1}{y_0}$$

Approximate 95 percent confidence limits on the log scale are derived by subtracting 1.96 standard errors from $\ln(RR)$ to derive the lower limit, and adding ($1.96 \cdot \mathrm{SE}$) to $\ln(RR)$ to derive the upper limit. Finally, the confidence interval is expressed on the untransformed scale of the RR by exponentiating the limits just obtained.

Using the data from Example 6.1, the variance of the logarithm of the rate ratio is

$$\mathrm{Var}[\ln(RR)] = \frac{1}{17} + \frac{1}{2} + \frac{1}{579} + \frac{1}{600}$$

$$= 0.5622$$

The 95 percent confidence interval is

$$\text{lower} = \exp[\ln(8.808) - 1.96\sqrt{0.5622}]$$

$$= 2.026$$

$$\text{upper} = \exp[\ln(8.808) + 1.96\sqrt{0.5622}]$$

$$= 38.30$$

To two significant digits, the rate ratio is 8.8 with a 95 percent confidence interval of 2.0 to 38.

The dominant term in the variance estimate given above is due to the two controls with a history of undescended testis. Their contribution to the total estimated variance is so great that despite a relatively large number of exposed cases, the overall estimate of effect remains very uncertain, as evidenced by the wide confidence interval. This example illustrates one of the limitations of case-control research: when the exposure under study is rare, estimates are likely to be highly unstable.

Analysis of Stratified Data. The most widely used estimate of a summary odds ratio over strata in a case-control study is that of Mantel and Haenszel.[64] The Mantel-Haenszel estimator provides a central value for the odds ratio to which each of the stratum-specific estimates contributes in approximate proportion to its own precision. It is calculated as follows. For each stratum i define the values x_{1i}, x_{0i}, y_{1i}, and y_{0i}, as above, and calculate their sum, T_i.

$$T_i = x_{1i} + x_{0i} + y_{1i} + y_{0i}$$

Now calculate two more derived quantities for each stratum, A_i and B_i.

$$A_i = \frac{x_{1i} y_{0i}}{T_i}$$

$$B_i = \frac{y_{1i} x_{0i}}{T_i}$$

Sum the values of A and B over the strata.

$$A = \sum_i A_i$$

$$B = \sum_i B_i$$

The Mantel-Haenszel summary estimate of the relative rate of disease over strata is

$$RR_{\mathrm{MH}} = \frac{A}{B}$$

The parameter estimated by RR_{MH} is a postulated odds ratio that is common to all the strata. Under proper study design, this parameter is identical to the hazard ratio in the source population giving rise to cases and controls. As in the analysis of a single stratum, a confidence interval for the hazard ratio is best calculated on the

64. **Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719-48**

logarithmic scale and then transformed back to the natural scale by exponentiation. In order to obtain an estimate of the variance of the estimator $\ln(RR_{MH})$, it is necessary to calculate several further quantities for each stratum.[65]

$$C_i = \frac{x_{1i} + y_{0i}}{T_i}$$

$$D_i = \frac{y_{1i} + x_{0i}}{T_i}$$

The four derived quantities are combined and summed, stratum by stratum, as follows:

$$(AC) = \sum_i A_i C_i$$

$$(AD) = \sum_i A_i D_i$$

$$(BC) = \sum_i B_i C_i$$

$$(BD) = \sum_i B_i D_i$$

The variance of $\ln(RR_{MH})$ is, approximately,

$$\mathrm{Var}[\ln(RR_{MH})] \doteq \frac{1}{2}\left[\frac{(AC)}{A^2} + \frac{(AD)+(BC)}{AB} + \frac{(BD)}{B^2}\right]$$

As before, the standard error is calculated as the square root of the variance, and 95 percent confidence intervals are obtained on the logarithmic scale by adding and subtracting 1.96 times the standard error to $\ln(RR_{MH})$, after which all of these are transformed back to the original rate ratio scale.

At first glance, the calculation of the 95 percent confidence interval seems a burdensome job, and it does entail a good deal of arithmetic when carried out by hand. An important feature of the formulas is that the data from each stratum need to be processed only once and added to the various accumulating terms. Repeated stratum-by-stratum accumulations are readily accommodated in spreadsheet programs and programmable calculators. When there is only one stratum, all the above formulas simplify to those given previously for the unstratified case.

65. Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. Am J Epidemiol 1986;124:719-23

Example 8.1. *Body mass index and the relative incidence of breast cancer.*[66]

Seventy-two premenopausal women with breast cancer newly diagnosed at the Group Health Cooperative of Puget Sound from July 1975 through June 1978 were interviewed, along with 80 premenopausal women from the same HMO who were hospitalized with a variety of acute conditions. For each subject, the body mass index was ascertained. Women with a body mass index (weight in kilograms divided by the square of height in meters) greater than 28 were classified as "heavy." The first two panels on the left side of Table 8.2 give the distribution of study subjects over categories of disease status, age, and body mass index, together with the age-specific estimates of the rate ratio and the 95 percent confidence intervals.

Calculated values for each of the quantities necessary for the summary estimate of the rate ratio and for the corresponding 95 percent confidence intervals are shown in the right hand panels of Table 8.2, and the resulting estimates are shown in the lower left. Heavy women appear to be at lower risk of breast cancer in both age groups than are other women. Both age-specific estimates are very unstable, because of the small number of heavy women in the study, and particularly so because of the paucity of heavy cases. The common estimate, which accumulates the information available from both strata, is more precise than either of the component values. Note that the common estimate lies within the range of the stratum-specific estimates, as will always be the case.

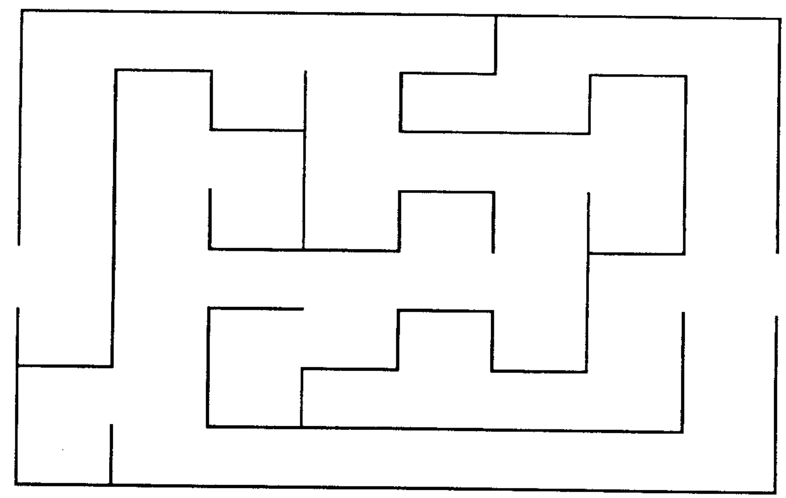
66. Jick H, Walker AM, Watkins RN, D'Ewart D, et al. Oral contraceptives and breast cancer. Am J Epidemiol 1980;112:577-85

Table 8.2 Body mass index and the relative incidence of breast cancer among premenopausal women *Calculation of table-specific and summary measures for a case-control study*

Age ≤ 45		Not		
	Heavy	Heavy	$T_1 = 78$	
			$A_1 = 0.7692$	$A_1C_1 = 0.3156$
Breast Ca	2	35	$B_1 = 0.4103$	$A_1D_1 = 0.4536$
			$C_1 = 4.9359$	$B_1C_1 = 2.0250$
Controls	11	30	$D_1 = 0.5897$	$B_1D_1 = 2.9109$
$RR = 0.16$			Variance of $\ln(RR) = 0.6528$	
95% CI = 0.032 - 0.76				
Age > 45		Not		
	Heavy	Heavy	$T_2 = 73$	
			$A_2 = 0.8767$	$A_2C_2 = 0.4083$
Breast Ca	2	33	$B_2 = 2.7123$	$A_2D_2 = 0.4684$
			$C_2 = 0.4658$	$B_2C_2 = 1.2633$
Controls	6	32	$D_2 = 0.5342$	$B_2D_2 = 1.4491$
$RR = 0.32$			Variance of $\ln(RR) = 0.7282$	
95% CI = 0.061 - 1.7				
Summary			$A = 1.6459$	$AC = 0.7239$
			$B = 7.6482$	$AD = 0.9220$
$RR_{MH} = 0.22$				$BC = 3.2883$
95% CI = 0.069 - 0.67				$BD = 4.3600$
			Variance of $\ln(RR) = 0.3381$	

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.