Hanley.

OBSERVATION AND INFERENCE
An Introduction to the Methods of Epidemiology

Alexander M. Walker, MD, DrPH

Epidemiology Resources Inc.

1991

Contents

estimate, but its proper use presupposes that the only source of discrepancy between the stratum-specific cumulative incidence differences is chance.

Confounding. If the interest in Table 5.1 had focussed on the cumulative incidence difference associated with attendance at the dance, the investigators could have calculated estimates of (47/86) - (8/23) or 20 percent in those who attended the luncheon and (11/77) - (1/40) or 12 percent in those who did not. Any standardized estimate of an overall effect would lie between these two values. If luncheon attendance were ignored, a crude cumulative incidence difference might also have been calculated as

$$CID = \frac{47+11}{86+77} - \frac{8+1}{23+40}$$

$$= 0.213$$

or 21 percent. This value lies outside of the range of stratum-specific estimates. Because luncheon attendance was more common among dancers than among those who did not dance, the crude cumulative incidence difference reflects a part of the cumulative incidence associated with luncheon attendance, in addition to the effect of dance attendance on risk. The crude cumulative incidence difference therefore provides a biased estimate of the increase in probability of pharyngitis associated with attendance at the dance.

Analysis of Open Cohort Studies

Example 5.2 will be used to illustrate the techniques presented here.

Error estimates and comparisons of incidence rates. Just as the observed proportion of the disease in a closed cohort study is an estimate of the underlying probability of developing disease, so the estimate of the incidence rate, provides an estimate ratio of cases to person time, the incidence rate, provides an estimate of the underlying hazard of disease. The most straightforward technique for assessing the variability of incidence rates in open cohort studies is based on a treatment of the incidence rate calculation as if the numerator (the number of cases) were variable and the denominator (the amount of person time) were fixed. If x is the number of observed events and P is the person time at risk, then x is the realization of what is called a Poisson process. The probability distribution from which x is drawn is the Poisson distribution.[58]

Poisson distribution *is the probability distribution that describes the number of events observed in a block of person time when the expected number of events is directly proportional to the total person time of observation. Let θ be the expected number of events per unit of person time and $\lambda = \theta P$ be the number of events expected in a block of person time of size P.*

$$\Pr(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$E(X) = \lambda$$

$$\mathrm{Var}(X) = \lambda$$

The range of possible values for x is $[0, \infty)$. λ is the Poisson parameter. If P is imagined as being composed of a very large number of discrete units of person time, so that the probability of an event in any person time unit is very small, then the probability distribution of the number of events in P may also be considered to be binomial, with N taken as the number of discrete person time units. All the formulas above are derivable from their binomial counterparts in the limiting case in which N approaches infinity, with P and λ constant.

The number of observed events x is an estimate of a Poisson parameter λ. The incidence rate estimate IR is given by x/P, with variance x/P^2.[59] The mortality rate estimate and its variance for the period from 30 through 34 years since first exposure (Table 5.2) are given by

[58] The development here presumes that the expected number of cases is directly proportional to the amount of person time of observation. Put another way, we presume that there is no element of contagion, in which the probability of a case occurring is a function of the number of other cases that have occurred.

[59] c.f. Table 1.2 and Chapter 13. Note that

$$IR = \frac{x}{P}$$

$$\frac{IR}{P} = \frac{x}{P^2}$$

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

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

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

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

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

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

$$= 4.831$$

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]

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

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

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

[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.