

COMPONENTS OF THE CRUDE RISK RATIO

OLLI S. MIETTINEN¹

(Received for publication December 17, 1971)

Miettinen, O. S. (Harvard School of Public Health, Boston, Mass. 02115). Components of the crude risk ratio. *Am J Epidemiol* 96: 168-172, 1972.—Both in cohort and case-control studies the estimate of the crude risk ratio factors into two important and easily derived components. One is a measure of the strength of confounding, and the other is an estimate of the residual risk ratio in terms of the standardized morbidity (mortality) ratio. For case-control studies the latter seems to be a new and useful “summary relative risk.”

epidemiologic methods; biometry

Risk ratio (“relative risk” (RR)) is a parameter of central interest in epidemiology. It is defined as the ratio of the risk among the exposed to that among the non-exposed, with “risk” referring to some measure of morbidity or mortality and “exposure” and “nonexposure” distinguishing between a pair of alternative experiences or characteristics. The data-layouts for its estimation in cohort and case-control studies, respectively, are presented in table 1. With a cohort study the “crude” RR is estimated naturally by the ratio of the two observed rates, i.e., by $\hat{p}_c = (e/F)/(g/H) = eH/gF$. The equivalent formula for case-control studies, introduced by Cornfield (1), is the “odds ratio” $\hat{p}_c = ad/bc$ as long as the compared series are independent (unmatched) and the disease is “rare” among exposed as well as nonexposed individuals.

These estimates of the crude RR are useful for predictive purposes, but in etiologic research it is necessary to make a distinction between two components of the parameter—one resulting from recogniza-

ble confounding and the other possibly due to the effect of exposure. Customarily, analyses in such studies have been focused on the estimation of the latter component with no explicit assessment of the strength of confounding. However, the quantitation of confounding may contribute importantly even to the evaluation of the effect of exposure. For example, it may show that the control of certain factors is unnecessary in the analysis—a welcome finding in the face of the common problem of vanishing numbers when the subjects are stratified by several potentially confounding factors. Moreover, routine assessment of the amount of confounding in the crude RR would help accumulate valuable experience for use in the planning and evaluation of other studies.

The present paper describes methods of breaking the crude RR down to the two components. Secondly, it may give some new insight into the interrelationship between cohort and case-control studies.

COHORT STUDIES

Suppose that the exposed and nonexposed series have been divided into strata of the confounding factor(s), and that in the j^{th} stratum the data layout analogous to the one in table 1 has entries e_j , F_j , g_j and H_j . With $\sum e_j = e$, $\sum F_j = F$, $\sum g_j$ and g

Abbreviations: ML, maximum likelihood; RR, risk ratio; SMR, standard morbidity (mortality) ratio.

¹Departments of Epidemiology and Biostatistics, Harvard School of Public Health, and Department of Cardiology, Children's Hospital Medical Center, Boston, Massachusetts.

Supported by grants 5 P01 CA 06373 and HE 10436 from the National Institutes of Health.

and $\sum H_j = H$ the estimate of the crude RR is, as before,

$$\hat{\rho}_c = eH/gF. \tag{1}$$

The component attributable to confounding by the stratification factor(s) may be estimated by simulating the removal of the effect of the exposure within each stratum. This removal would have no bearing on the nonexposed series, and generally only a negligible effect, if any, on the distribution of exposed subjects among the strata. Thus, for considering this hypothetical case of no effect of the exposure we may keep the values of g_j , H_j and F_j unchanged. On the other hand, each e_j is replaced by the estimated "null" value of $e_j^* = g_jF_j/H_j$ corresponding to $\hat{\rho}_j = 1$. The estimate of the RR component, ρ^* , attributable to confounding by the stratification factors may then be taken as

$$\hat{\rho}^* = e^*H/gF, \tag{2}$$

where $e^* = \sum e_j^*$.

The estimation of the residual RR, ρ_r , associated with the exposure conditionally on the stratification factors (i.e., with their effects removed) is a familiar problem. Among the exposed, a total of e events were observed as against an estimated "expected" number of e^* based on the rates in the nonexposed series, and the usual estimate of the residual RR is the standardized morbidity (mortality) ratio (SMR):

$$\hat{\rho}_r = e/e^*. \tag{3}$$

Even though this ratio is the core element in "indirect standardization" with the nonexposed group as the "standard" (in the sense of stratum-specific rates), it should be regarded as the ratio of "directly" standardized rates for the exposed and nonexposed—i.e., an estimate of *standardized RR*—with the *exposed* group as the standard (in the sense of distribution over the strata). If the RR is the same for all the strata, and if the events are referred to person-years of follow-up (so that e_j and g_j can be regarded as realizations for independent Poisson variates), then e/e^* is the maximum likelihood

TABLE 1
Data-layouts in cohort and case-control studies, respectively

Cohort study			Case-control study		
	Compared series			Compared series	
	Exposed	Nonexposed		Cases	Controls
Events*	e	g	Exposed	a	c
Denominator†	F	H	Nonexp.	b	d

* No. of cases of disease or death.

† No. of individuals studied or person-years of follow-up.

(ML) estimate of the common RR. With count denominators the above statistic is the ML estimate of a common RR only in the limit when the rates (of attack or prevalence) are very low, but it is a reasonable estimate even in less extreme situations.

It is apparent from the above formulas that the estimates of the components of RR are multiplicative, i.e., that

$$\hat{\rho}_c = \hat{\rho}^* \hat{\rho}_r. \tag{4}$$

CASE-CONTROL STUDIES

In case-control studies with substantial numbers of subjects at each stratum an argument analogous to the one above can be used to estimate ρ^* , and the procedure suggests a simple estimate for ρ_r as well. Under the rare disease assumption the hypothetical removal of the effect of exposure induces no appreciable change in the control series nor in the distribution of nonexposed cases over the strata (even though the proportion of nonexposed individuals among all cases would change if the exposure in fact had an effect). For estimating ρ^* we may, therefore, keep the values of c_j , d_j and b_j unchanged, whereas each a_j is replaced by the "null" value of $a_j^* = b_jc_j/d_j$ which makes $\hat{\rho}_j = 1$. Thus the estimate becomes

$$\hat{\rho}^* = a^*d/bc, \tag{5}$$

where $a^* = \sum a_j^*$.

TABLE 2

Drug-attributed rash in relation to allopurinol exposure among recipients of ampicillin. The Boston Collaborative Drug Surveillance Program (3)

Rash	Males			Females			Total		
	Allopurinol			Allopurinol			Allopurinol		
	+	-	Total	+	-	Total	+	-	Total
+	5	36	41	10	58	68	15	94	109
-	33	645	678	19	518	537	52	1163	1215
Total	38	681	719	29	576	605	67	1257	1324
$\hat{\rho}$	2.49			3.42			2.99		

For estimating ρ_s we first observe that the j^{th} stratum provides the estimate of $a_j d_j / b_j c_j = a_j / a_j^*$, the ratio of the observed number of exposed cases to the respective estimated expectation— analogously to e_j / e_j^* in the case of cohort studies. This analogy with cohort studies suggests that a reasonable overall estimate of the residual component in the large-sample case is

$$\hat{\rho}_s = a/a^*, \tag{6}$$

which—like e/e^* in cohort studies—is the ratio of the observed total number of exposed cases to the estimate of its “expected” value under the assumption that $\rho_j \equiv 1$. Moreover, a/a^* must have the interpretation of being the sample value for the *standardized RR, with the exposed individuals in the source population as the standard* (in the sense of distribution over the strata). This inference can be directly verified: As noted above, the sample standardized RR is $\hat{\rho}_s = \sum w_j \hat{R}_{1j} / \sum w_j \hat{R}_{0j}$, where the w_j are the weights of (“direct”) standardization, and the \hat{R}_{1j} ’s and \hat{R}_{0j} ’s are rates in the exposed and nonexposed series, respectively. By definition, then, $\hat{\rho}_s = \sum w_j \hat{R}_{0j} \hat{\rho}_j / \sum w_j \hat{R}_{0j}$. With the exposed individuals in the source population as the standard we may set w_j proportional to c_j / f_j , where f_j is the sampling fraction of noncases in the j^{th} stratum. The values of \hat{R}_{0j} may be taken proportional to $b_j / (d_j / f_j)$ as long as the

sampling fraction for cases is uniform over the strata. These substitutions, together with $\hat{\rho}_j = a_j d_j / b_j c_j$, yield

$$\hat{\rho}_s = \sum a_j / \sum (b_j c_j / d_j) = a/a^*.$$

This proof also underscores the fact that the above interpretation of a/a^* as a standardized RR presupposes that the *cases be representative* of all cases in the source population, whereas the control series may be either representative of noncases or matched.

As with cohort studies, it is seen that $\hat{\rho}_c = \hat{\rho}^* \hat{\rho}_s$.

The above method of estimating the components of risk ratio should not be used in the small-sample case, particularly if d_j takes on small values. In the extreme, a single $d_j = 0$ can make $a^* = \infty$ and, consequently, $\hat{\rho}^* = \infty$ and $\hat{\rho}_s = 0$. When the values of c_j tend to be larger than those of d_j , a more stable measure of residual RR is obtained by using the *nonexposed as the standard* (in the sense of distribution over the strata) and computing

$$\hat{\rho}_s = b^*/b, \tag{7}$$

where $b^* = \sum (a_j d_j / c_j)$. In the case of uniform RR over the strata a generally applicable, though computationally often somewhat cumbersome, approach is one where the ML estimate is first derived for $\hat{\rho}_r$, and $\hat{\rho}^*$ is then obtained from the relationship $\hat{\rho}^* = \hat{\rho}_c / \hat{\rho}_r$. A description of the ML estimation has been given recently by Gart (2).

EXAMPLES

Example 1. Table 2 presents some drug surveillance data analyzed in the spirit of *cohort* studies. Specifically, the frequency of drug-attributed rash is related to Allopurinol exposure among recipients of Ampicillin, and sex is treated as a potential confounding factor. The estimate of the crude risk ratio is $\hat{\rho}_c = 15(1257)/94(67) = 2.99$. Simulating the removal of the Allopurinol effect we have $e_1^* = 36(38)/681 = 2.01$ and $e_2^* = 58(29)/576 = 2.92$, and the risk

ratio related to confounding by sex therefore has the estimate $\hat{\rho}^* = (2.01 + 2.92) (1257)/94(67) = 0.98$. This indicates that there is no material confounding by sex even though the rash rate has a rather strong relationship to sex among the nonexposed, being $36/681 = 5.3$ per cent in males and $58/576 = 10.1$ per cent in females; the reason for the absence of confounding by sex is that the exposure rates in males and females are similar, namely $38/719 = 5.3$ per cent and $29/605 = 4.8$ per cent, respectively. The estimate of possibly Allopurinol-attributable, standardized risk ratio is $\hat{\rho}_s = 15/(2.01 + 2.92) = 3.04$. The interrelationship of $\hat{\rho}_c = \hat{\rho}^*\hat{\rho}_s$ is found to hold, as $2.99 = 0.98 (3.04)$ apart from round-off inaccuracy. If in this example there was a need to stratify by other factors, the lack of sex confounding would permit omission of stratification by sex.

Example 2. Table 3 gives data from a case-control study of the relationship of oral contraceptive use to venous thrombosis with age as a confounding factor. The estimate of the crude risk ratio is $\hat{\rho}_c = 12(347)/53(30) = 2.62$. The component of this attributable to confounding by age is not fully estimable from the data presented, inasmuch as there is residual age confounding within the 10-year categories considered. Ignoring this problem, we may compute $a_1^* = 39(18)/158 = 4.44$ and $a_2^* = 14(12)/189 = 0.89$, and these give as the estimate of the risk ratio component attributable to confounding by age in the 20-39-year range $\hat{\rho}^* = (4.44 + 0.89)(347)/53(30) = 1.16$. The estimate of the standardized component is $\hat{\rho}_s = 12/(4.44 + 0.89) = 2.25$. The product of these two components is $1.16 (2.25) = 2.62$, the value of the crude risk ratio from the total series without stratification. As the age-specific estimates of risk ratio are identical, it is obvious that the ML procedure also gives $\hat{\rho}_r = 2.25$ and, therefore $\hat{\rho}^* = 2.62/2.25 = 1.16$. Considering the existence of confounding within the two 10-year categories of age, the total

TABLE 3
Venous thrombosis in relation to oral contraceptives (O.C.) use among hospitalized women (ref. 4)

O.C.	Age category								
	20-29 years			30-39 years			Total		
	Thrombosis			Thrombosis			Thrombosis		
	+	-	Total	+	-	Total	+	-	Total
+	10	39	49	2	14	16	12	53	65
-	18	158	176	12	189	201	30	347	377
Total	28	197	225	14	203	217	42	400	442
$\hat{\rho}$	2.25			2.25			2.62		

confounding by age might correspond to $\hat{\rho}^* = 1.3$ with the corresponding $\hat{\rho}_s = 2.62/1.3 = 2.0$.

DISCUSSION

Confounding and residual components in risk difference ("attributable risk")—the risk in the exposed minus that in the nonexposed—in cohort studies have been discussed quite thoroughly by Kitagawa (5).

The present paper presents simple risk ratio measures of confounding and of the residual association between exposure and disease, both for cohort and case-control studies. The measure of confounding ($\hat{\rho}^*$) follows immediately from the definition of the confounding effect, provided that the "null" situation is not approached in the usual manner (where the marginal frequencies are "fixed") but by "fixing" those frequencies which generally do not materially depend on the effect of exposure (i.e., on the number of cases among the exposed). Consideration of cohort studies suggests that the most natural RR measure of the residual association might be the $\widehat{\text{SMR}}$, i.e., the sample standardized RR ($\hat{\rho}_s$) with the exposed series as the standard. This has conceptual appeal and makes the estimate of crude RR ($\hat{\rho}_c$) to have a simple multiplicative partitioning: $\hat{\rho}_c = \hat{\rho}^*\hat{\rho}_s$. And quite interestingly, it turns out that this residual parameter (SMR or ρ_s) can be estimated in a simple

manner from case-control studies as well. The procedure adds a "summary relative risk" with a clear-cut interpretation to the several measures previously discussed by Mantel and Haenszel (6).

REFERENCES

1. Cornfield J: A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J Natl Cancer Inst* 11: 1269-1275, 1951
2. Gart JJ: Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals. *Biometrika* 57: 471-475, 1970
3. The Boston Collaborative Drug Surveillance Program: Excess of ampicillin rashes associated with allopurinol or hyperuricemia. *N Eng J Med* 286: 505-507, 1972
4. Miettinen OS: Comprehensive monitoring of hospitalized patients in the evaluation of side-effects of oral contraceptives. Presented to the International Symposium on Statistical Problems in Population Research, the East-West Population Institute, Honolulu, August 1971
5. Kitagawa EM: Components of a difference between two rates. *J Am Stat Assn* 50: 1168-1194, 1955
6. Mantel N, Haenszel W: Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 22: 719-748, 1959