

ESTIMATORS OF RELATIVE RISK FOR CASE-CONTROL STUDIES

CAROL J. R. HOGUE,¹ DAVID W. GAYLOR² AND KENNETH F. SCHULZ¹

Hogue, C. J. R. (CDC, Atlanta, GA 30333), D. W. Gaylor and K. F. Schulz. Estimators of relative risk for case-control studies. *Am J Epidemiol* 1983;118: 396-407.

The odds ratio from a case-control study of the "cumulative-incidence" type can be used as an estimate of the relative risk of a disease attributable to exposure to an agent only when the incidence of the disease is low. The odds ratio can be modified to obtain an accurate estimate of the relative risk, regardless of the incidence of the disease. This modification of the odds ratio can be performed with any one of four types of auxiliary information: overall probability of disease, probability of disease in the unexposed population, probability of disease in the exposed population, or overall probability of exposure. For "incidence density" case-control studies, the odds ratio equals the relative risk when the estimate of exposure in the comparison group can be considered to be an estimate of the overall probability of exposure in the population at risk. Under certain conditions, such "case-exposure" studies may be preferable to cohort studies and to cumulative-incidence case-control studies. The authors present an approach to hypothesis testing for crude and stratified data from a case-exposure study.

epidemiologic methods; reproduction; risk; statistics

Two types of case-control studies designed to estimate incidence ratios (relative risk) were identified by Miettinen (1) and further elucidated by Neutra and Drolette (2). The first kind is the "incidence-density" study, in which exposure histories of incident cases are compared with exposure histories of noncases who

are still at risk of becoming cases; the second type is the "cumulative-incidence" study, in which exposure histories of incident cases are compared with exposure histories of noncases who are no longer at risk of becoming cases.

Cumulative-incidence studies are often found in the literature of reproductive epidemiology (3, 4). In such studies, affected infants (cases) are compared with normal infants (controls) with respect to in utero exposures such as maternal smoking, or preconceptual exposures such as induced abortion of previous pregnancies. In the field of infectious disease epidemiology, cumulative-incidence studies have been employed to investigate the efficacy of vaccination in preventing infection during an epidemic (5, 6).

The odds ratio is not equivalent to the relative risk for the cumulative-incidence study (1, 2); the accuracy of the odds ratio

Received for publication August 5, 1982, and in final form March 3, 1983.

¹ Division of Reproductive Health, Centers for Disease Control, 1600 Clifton Road, Atlanta, GA 30333. (Send reprint requests to Dr. Hogue at this address.)

² Division of Biometry, National Center for Toxicological Research, Jefferson, AR 72079.

The authors thank the following persons for their contributions to this work: Phillip Stubblefield, the Boston Hospital for Women; George Carlo, Dow Chemical Company; Henry Arrighi, Aramco Oil Company, Saudi Arabia; Dr. Howard Ory, Martha Mayfield, Gail Carpenter, and Ann Agnor, the Centers for Disease Control; Susan Taylor, the National Center for Toxicological Research; and Teri Wallis, the University of Arkansas for Medical Sciences.

as an approximation to the relative risk depends on the rare disease assumption, as described by Cornfield (7). Unfortunately, rare disease cannot be assumed either for infectious epidemic investigation or for many reproductive outcomes such as spontaneous abortion, which often amounts to 10 per cent of all pregnancies that terminate in a hospital (4).

The purposes of this paper with respect to cumulative-incidence case-control studies are to explore the extent of inaccuracy in the odds ratio approximation of relative risk and to recommend approaches for eliminating or reducing this difference. Specifically, the goals are 1) to illustrate the direction and amount of difference, under different levels of disease incidence, and relative risk, and 2) with the use of one of four kinds of auxiliary information, to de-

velop estimators of relative risk for cumulative-incidence case-control studies which are accurate estimators of relative risk—regardless of the incidence of the disease or of the size of the relative risk.

With incidence-density case-control studies, the odds ratio is an accurate estimator of relative risk. It will be shown that, assuming that controls who are obtained concurrently with cases are representative of the exposure experience of the population from which the cases are drawn, the odds ratio is equivalent to the relative risk. Under this assumption, incidence-density case-control studies may be thought of as “case-exposure” studies. We will explore the usefulness of this approach vis-à-vis cohort studies and cumulative-incidence case-control studies.

NOTATION

Some notation is required to establish the relationship between the odds ratio and the relative risk. First, consider the true population at the end of the period of risk, when all persons in the population who are going to develop the disease have become ill. The various probabilities, P , when disease (D) and exposure (E) are present (+) or absent (–), are given in the following table:

Exposure	Disease		
	+	–	
+	$P(D,E)$	$P(\bar{D},E)$	$P(E)$
–	$P(D,\bar{E})$	$P(\bar{D},\bar{E})$	$P(\bar{E})$
Total	$P(D)$	$P(\bar{D})$	1.0

For example, $P(\bar{D},\bar{E})$ is the probability that an individual is both unexposed and free from the disease, i.e., the number of unexposed, nondiseased individuals in the population divided by the total population size. The proportion of nondiseased individuals in the population is $P(\bar{D},E) + P(\bar{D},\bar{E}) = P(\bar{D})$.

The relative risk, R , is the ratio of the proportion of the diseased individuals who were exposed relative to the proportion of diseased individuals who were not exposed:

$$R = \frac{P(D|E)}{P(D|\bar{E})} = \frac{P(D,E)/P(E)}{P(D,\bar{E})/P(\bar{E})} \quad (1)$$

In a cumulative-incidence case-control study, cases and controls are selected and the status of past exposure is determined. Thus, all proportions are conditional upon the disease state. Such a case-control study provides estimates of the following proportions.

Exposure	Disease	
	Case +	Control -
+	$P(E D)$	$P(E \bar{D})$
-	$P(\bar{E} D)$	$P(\bar{E} \bar{D})$
Total	1.0	1.0

The odds ratio, Θ , is a ratio derived directly from these estimates:

$$\Theta = \frac{P(E|D)/P(\bar{E}|D)}{P(E|\bar{D})/P(\bar{E}|\bar{D})} = \frac{P(D,E)/P(D,\bar{E})}{P(\bar{D},E)/P(\bar{D},\bar{E})} \tag{2}$$

When $P(D)$ is small, the odds ratio is nearly equivalent to the relative risk (7).

RELATIVE RISK ESTIMATORS FOR CUMULATIVE-INCIDENCE STUDIES

Relative risk versus the odds ratio. Substantial differences can occur when the odds ratio, one population parameter, is used to estimate relative risk, a different population parameter, in cumulative-incidence studies. The difference between the two parameters can be illustrated by examining the ratio of Θ to R . When the invariance property of the odds ratio (8) is used,

$$\Theta = \frac{P(E|D) \cdot P(\bar{E}|\bar{D})}{P(\bar{E}|D) \cdot P(E|\bar{D})} = \frac{P(D|E) \cdot P(\bar{D}|\bar{E})}{P(D|\bar{E}) \cdot P(\bar{D}|E)}$$

The relationship between the two parameters is the following,

$$\Theta = R \cdot \frac{P(\bar{D}|\bar{E})}{P(\bar{D}|E)} \tag{3}$$

When R equals 1, the two parameters are equivalent, but whenever R is greater than 1, Θ is greater than R . Whenever R is less than 1, Θ is less than R .

Inaccuracies that can occur when the odds ratio is used to estimate the relative risk are illustrated in table 1. For any value of proportion of unexposed individuals with the disease, $P(D|\bar{E})$, the odds ratio increasingly overestimates the relative risk as relative risk increases. In fact, for large R the difference can be sizable, even for small $P(D|\bar{E})$. For a relative risk greater than two, the odds ratio overestimate increases rapidly with $P(D|\bar{E})$. For values of $R \cdot P(D|\bar{E})$ less than 0.2, the difference never exceeds 25 per cent.

Equation 3 may be rewritten as

$$\Theta = R + (\Theta - 1) \cdot P(D|\bar{E}) \tag{4}$$

That is, the difference between the two parameters is $(\Theta - 1) \cdot P(D|\bar{E})$. As the odds ratio approaches one, the difference approaches zero, and the odds ratio approaches the relative risk regardless of the value of $P(D|\bar{E})$. As the probability of disease in the total population approaches zero, the difference approaches zero and the odds ratio approaches the relative risk regardless of the size of the relative risk. Furthermore, the relative difference $(\Theta - R)/\Theta$ approaches a maximum value of one as the odds ratio increases. Using

$$\frac{\Theta - R}{\Theta} = \frac{\Theta - 1}{\Theta} \cdot P(D|\bar{E}),$$

it also follows that the maximum value of the relative difference is the probability of disease among the exposed population, $P(D|E)$.

Orenstein et al. (9) considered the proportionate relationship of Θ to R in studies of vaccine efficacy. In efficacy studies, R is less than 1, and Θ is less than R . The authors have shown that the proportionate relationship in the efficacy ratio, i.e., $(R - \Theta)/(1 - R)$, is entirely a function of the attack rate in the vaccinated population.

Adjusting the odds ratio using auxiliary information. The relative risk may be written as a function of the odds ratio, the conditional probabilities from a cumulative-incidence case-control study, and the probability of disease, $P(D)$, derived from auxiliary information. When Bayes' theorem (2, 10) is used, equation 3 may be written as

$$R = \Theta \cdot \frac{P(E|\bar{D})}{P(\bar{E}|\bar{D})} \cdot \frac{P(\bar{E}|D)P(D) + P(\bar{E}|\bar{D})P(\bar{D})}{P(E|D)P(D) + P(E|\bar{D})P(\bar{D})}. \quad (5)$$

From equation 4,

$$R = \Theta - (\Theta - 1) \cdot P(D|E). \quad (6)$$

Thus, the relative risk can be expressed as a function of the odds ratio and the probability of disease among the exposed population, $P(D|E)$.

From equations 1 and 3,

$$R = \frac{\Theta}{1 - P(D|\bar{E}) + \Theta \cdot P(D|\bar{E})}. \quad (7)$$

Thus, the relative risk can be expressed as a function of the odds ratio and the probability of disease among the unexposed population, $P(D|\bar{E})$.

From equation 3,

$$\frac{\Theta}{R} = \frac{P(\bar{D}, \bar{E}) \cdot P(E)}{P(\bar{E}) \cdot P(\bar{D}, E)} = \frac{P(\bar{D}, \bar{E})/P(\bar{D})}{P(\bar{D}, E)/P(\bar{D})} \cdot \frac{P(E)}{P(\bar{E})} = \frac{P(\bar{E}|\bar{D})}{P(E|\bar{D})} \cdot \frac{P(E)}{P(\bar{E})}.$$

Solving for R gives

$$R = \Theta \cdot \frac{P(E|\bar{D})}{P(\bar{E}|\bar{D})} \cdot \frac{1 - P(E)}{P(E)}. \quad (8)$$

Thus, the relative risk can be expressed as a function of the odds ratio, $P(E|\bar{D})$, and the probability of exposure in the population, $P(E)$.

These four methods are available for estimating the relative risk from modifications of the odds ratio, as given by equations 5–8, with the use of auxiliary information on $P(D)$, $P(D|E)$, $P(D|\bar{E})$, or $P(E)$, respectively. Estimates of relative risk can be obtained from cumulative-incidence case-control studies without the requirement of a rare disease when an estimate of any one of the four auxiliary probabilities is available.

Methods for obtaining approximate confidence limits of these four estimators are given in the Appendix.

Theoretical example. Suppose the true population proportions are as follows:

Exposure	Disease		Total
	+	-	
+	0.24	0.16	0.40
-	0.09	0.51	0.60
Total	0.33	0.67	1.00

The relative risk is

$$R = \frac{0.24/0.40}{0.09/0.60} = 4.0.$$

From this population, the true case-control proportions are:

Exposure	Disease	
	+	-
+	0.727	0.239
-	0.273	0.761
Total	1.000	1.000

The true odds ratio is

$$\Theta = \frac{0.727/0.273}{0.239/0.761} = 8.5,$$

or 2.125 times the relative risk (cf. table 1).

When $P(D) = 0.33$ is used, the odds ratio can be modified as given in equation 5 to obtain the relative risk. When $P(D|E) = 0.24/0.40 = 0.60$ is used, the odds ratio can be modified as given in equation 6 to obtain the relative risk. When $P(D|\bar{E}) = 0.09/0.60 = 0.15$ is used, the odds ratio can be modified as given in equation 7 to obtain the relative risk. Finally, when $P(E) = 0.40$ is used, the odds ratio can be modified as given in equation 8 to obtain the relative risk.

Feasibility of using auxiliary information. The usefulness of any formula for correcting an odds ratio estimate of relative risk should be weighed within the context of four items: prevalence of disease, hypothesized relative risk, availability of auxiliary information, and accuracy of the auxiliary information. As table 1 illustrates, relative risk will be overestimated by at least 12 per cent when the true relative risk is 2.0 or greater and the attack rate among the unexposed is at least 5 per cent. One probably does not need to be concerned about using the odds ratio estimate for attack rates less than 5 per cent or for relative risks hypothesized to be less than 2.0, since the overestimate with the odds ratio approximation would probably not be greater than errors in measuring disease and exposure status. When very precise measures are desired, however, and accurate auxiliary information is available, one may wish to adjust for the difference, even in the range of difference less than 10 per cent. An example of such a situation could be an investigation of vaccine efficacy in a school outbreak of an infectious disease in which the overall attack rate can be easily and accurately determined.

TABLE 1
Values of the Ratio of θ to R

Probability of disease among the unexposed $P(D \bar{E})$	Relative risk (R)				
	1	2	3	4	5
0.01	1.00	1.01	1.02	1.03	1.04
0.05	1.00	1.06	1.12	1.19	1.27
0.10	1.00	1.12	1.29	1.50	1.80
0.15	1.00	1.21	1.55	2.12	3.40
0.20	1.00	1.33	2.00	4.00	∞

Once it has been established that a substantial difference could be present because of high disease prevalence and hypothesized relative risk, it is necessary to ascertain the availability of auxiliary information. When infectious diseases or reproductive outcomes are being studied, additional data are often obtainable; however, depending on the type of case-control study design, some types of auxiliary data may be more easily gathered than others.

Probability of disease may be calculated or estimated in a variety of common health problems. As mentioned above, the overall attack rate in an infectious disease outbreak may be measured easily in a defined population at risk, such as a school (11). Definition of the population at risk can be refined also, e.g., by excluding classrooms not affected or by calculating age-specific attack rates.

Within reproductive epidemiology, rates of low birth weight, premature delivery, and hospitalized spontaneous abortions may all be calculated for the total hospitalized population of pregnant women for the period during which the cases were collected. To illustrate the benefit of using such auxiliary information on overall disease rate, we estimated the difference in the odds ratio approximation of 2.6 for the effect of repeated induced abortion on subsequent spontaneous abortions, as reported by Levin et al. (4). Dr. P. G. Stubblefield, one of the co-authors of the Levin paper, informed us that 11,608 maternal discharges occurred during the study period (personal communication, 1981). Since 1238 women were discharged with the diagnosis of spontaneous abortion less than 20 weeks' gestation or premature delivery between 20 to 27 weeks' gestation (4), $\hat{P}(D) = 0.1067$, or 106.7 per 1000 discharges. When equation 5 and the Appendix were used, $\hat{R} = 2.24$ with the 95 per cent confidence interval of (1.47, 3.40). The odds ratio overestimated relative risk in this example by 16 per cent.

Calculating $P(D)$ in nonhospitalized populations or noninstitutionalized populations may not be as feasible as it was in the above examples. Morbidity statistics for the population at risk might be more difficult to obtain than morbidity rates for specific subpopulations, such as an exposed cohort or the unexposed group. Occupational studies of exposed workers can be used to estimate $P(D|E)$ (12, 13). Combined with a case-control strategy within the workplace, this information could provide for an estimate of relative risk that would be less subject to the problems inherent with studies that use standardized morbidity ratios. The probability of disease in an unexposed population could be estimated under certain circumstances by the use of general morbidity statistics. Situations in which this approach might be appropriate would include the investigation of the effects of a new drug on pregnancy outcome. Assuming that other factors affecting risks of adverse outcomes have not changed during the study period, data gathered immediately before the drug was introduced could serve as an estimate of $P(D|\bar{E})$.

The quality of auxiliary data is another important consideration. External data should be gathered, when possible, on the population from which the cases arose. To the extent that the data do adequately describe that population, they will be useful for correcting the odds ratio. Errors of measurement will reduce the accuracy of the estimate of relative risk and, concomitantly, lower the benefit derived from having auxiliary information. Even inaccurate auxiliary data, however, may result in an improved approximation of relative risk. For example, in the hypothetical example of the previous section, the odds ratio overestimated relative risk by a factor of 2.125. With perfect measures of auxiliary information, the odds ratio can be adjusted to equal the relative risk. If, on the other hand, $P(D)$ had been overestimated or underestimated by 10 per cent, R would have been underestimated or overestimated by 7

per cent. Similar errors with the other pieces of auxiliary information would have produced errors in relative risk in the range of 5 to 19 per cent.

INCIDENCE-DENSITY (CASE-EXPOSURE) STUDIES

In the previous section, methods to obtain an estimate of $P(E)$ were not discussed. When an adequate measure of the probability of exposure is obtainable, it is not necessary to collect a control series for a cumulative-incidence case-control study. This fact becomes obvious when equation 2 is substituted into equation 8, for

$$R = \frac{P(E|D)}{P(\bar{E}|D)} \cdot \frac{1 - P(E)}{P(E)} \tag{9}$$

Equation 9 consists of conditional probabilities of exposure status among diseased and the unconditional probability of exposure in the population at risk. This information may be displayed in a 2×2 table as follows:

Exposure	Cases	Sample of population at risk
+	$P(E D)$	$P(E)$
-	$P(\bar{E} D)$	$1 - P(E)$
Total	1.0	1.0

In incidence-density case-control studies, the control series remains at risk of developing the disease at some time in the future (1, 2, 14). If prevalent cases are included in the control series, the odds ratio is an unbiased estimator of relative risk (1) because the control series is actually a sample chosen to estimate the probability of exposure in the population from which the cases arose. It could be termed a case-exposure study rather than a case-control study. In such case-exposure studies, no rare disease assumption is required, since the odds ratio is equivalent to the relative risk, under the condition of constant exposure rates over the study period (15).

Examples of case-exposure methodology include population-based case-control studies of cancer (16-18) and historical prospective studies of occupational cohorts, in which a case series is "imbedded" in a follow-up study of a cohort of workers (14). The approach has not been applied in studies of infectious disease outbreaks or in reproductive epidemiology, although its theoretical advantages have been recognized (11, 19).

Extension to stratified data. The stratified 2×2 tables generated from a case-exposure study can easily be imputed into one of the Rothman and Boice (20) programs for case-control studies to obtain a weighted relative risk estimate. For example, with each stratum of case-exposure study being i and the 2×2 table for each stratum being

Exposure	Cases	Population at risk	Totals
+	a_i	b_i	N_{1i}
-	c_i	d_i	N_{0i}
Totals	M_{1i}	M_{0i}	T_i

a useful approximation to the asymptotic maximum likelihood estimate of relative risk over all strata is

$$\hat{R}_{M-H} = \frac{\sum_i a_i d_i / t_i}{\sum_i b_i c_i / t_i}.$$

Referred to as the Mantel-Haenszel pooled point estimate (21), this is the same formulation as given in Rothman and Boice (20) except that the cells are defined by the case-exposure study. The net result will be an asymptotically unbiased estimate of relative risk.

Statistical inference. One approach to estimating variance and confidence limits for the case-exposure estimate of relative risk allows for stratification to control for confounding by a combination of the Mantel-Haenszel pooled point estimate, obtained as described above for stratified data, with a Mantel-Haenszel chi derived from the pooled data. Then test-based lower (\underline{R}) and upper (\overline{R}) confidence limits are

$$\underline{R} = \hat{R}_{M-H} (1 - Z/\chi_{M-H}) \text{ and } \overline{R} = \hat{R}_{M-H} (1 + Z/\chi_{M-H})$$

where \hat{R}_{M-H} is the unbiased pooled point estimate, χ_{M-H} is the value of chi from the pooled data, and Z is the value of a standard normal deviate corresponding to the desired level of confidence.

It is possible to reduce the length of the confidence interval and thus to increase the power of a case-exposure study by "decontaminating" the exposure sample before the Mantel-Haenszel chi is calculated; this removes persons who ultimately will become cases or who are prevalent cases from the exposure series. These cases may be shifted to the case series, if they are not already represented there. The rationale is that the test of hypothesis is a test of the degree of association between being a case and being exposed. Although the odds ratio approximation of relative risk from "pure" samples of noncases overestimates R , as has been shown for cumulative-incidence case-control studies, pure samples of noncases are better for testing the degree of association between case status and level of exposure. This procedure is particularly valuable when the incidence of disease is high.

The realigned table would be

Exposure	Cases	Noncases	Totals
+	a'_i	b'_i	N_{1i}
-	c'_i	d'_i	N_{0i}
Totals	M_{1i}	M_{0i}	T_i

where a'_i is the original number of exposed cases plus additions, if any, of cases from the exposure sample. Note that b_i is always greater than b'_i , but a_i may not be less than a'_i , since the case series may already contain those individuals in the population at risk sample who became cases.

Once the data from the case-exposure study have been realigned to conform to the concept of a cumulative-incidence study, a Mantel-Haenszel chi may be derived from the pooled, realigned data. Test-based lower (\underline{R}) and upper (\overline{R}) confidence limits would then be

$$\underline{R} = \hat{R}_{M-H} (1 - Z/\chi'_{M-H}) \text{ and } \overline{R} = \hat{R}_{M-H} (1 + Z/\chi'_{M-H})$$

where \hat{R}_{M-H} and Z are as previously defined and χ'_{M-H} is the value of chi from the pooled, realigned data.

Although this realignment approach is more powerful than testing for association with nonrealigned data, it is not always possible to identify those individuals in the population at risk sample who will become cases at some future date after the study has been completed. Without knowledge of whom to move, the investigator will be limited to the original Mantel-Haenszel chi, with a resultant wider confidence limit around the unbiased estimate of R .

Hypothetical example. To illustrate the case-exposure method, we have chosen a hypothetical example of an investigation of spontaneous abortion, with a population at risk of 1000 pregnant women, of whom 100 will miscarry. The probability of exposure is 0.40, an exposure level comparable to smoking among pregnant women (22). For exposed women, relative risk = 1.5.

In a case-exposure study of reproductive outcomes, a sample of women would be questioned at their first prenatal visit, e.g., regarding first-trimester smoking. This is the population at risk sample. As the members of the population terminate their pregnancies, a sample of women who spontaneously abort, i.e., the case series, would also be questioned. For this example, we are assuming that all cases are included in the study, and a random sample of 180 pregnant women is included in the exposure series. Table 2 gives the expected results of the study.

Since all cases are determined in the study, there are an expected 18 women of the exposure group who ultimately became cases. For purposes of statistical inference, these may be removed from the exposure series. The expected point estimate is $\hat{R} = 1.5$, with a 90 per cent confidence interval of 1.02 to 2.21.

A comparable case-control study with 100 cases and 180 controls, of the cumulative-incidence type, would produce an expected estimate of $\hat{\Theta} = 1.57$, a 5 per cent overestimate of R . Case-control studies of this type may be particularly vulnerable to selective recall, with mothers of affected offspring more likely to recall past exposures accurately than are mothers of normal children. If sensitivity of the exposure measure for controls were reduced to 90 per cent, while the remaining sensitivity and specificities were 1.0, the expected estimate would be $\hat{\Theta} = 1.9$, compared with the relative risk of 1.5.

If, instead of a case-control design, a cohort approach were chosen, with 280 women in the cohort—112 of them smokers—the expected estimate would be $\hat{R} = 1.5$, but the 90 per cent confidence interval would include 1.0 (0.83, 2.7) because of the smaller number of cases. Furthermore, when a case-exposure design is compared with a cohort design, case status may be determined independently, e.g., from a case register, obviating the need to follow the cohorts to discover ultimate disease status. Also, a cohort study often is preceded by a prevalence study to eliminate prevalent cases and to identify a sufficient number of exposed and unexposed individuals for division into two distinct cohorts. That prevalence study alone is enough to establish the estimate of probability of exposure. Hence, a case-exposure study should be less expensive than

TABLE 2

Hypothetical data illustrating case-exposure study of spontaneous abortion and maternal smoking

Smoking in first trimester	Cases	Exposure series from population at risk
Yes	50	72
No	50	108
Total	100	180

such a cohort study, since identification and measurement of an incident case series may well cost less than identification and follow-up of two cohorts.

CONCLUSION

The odds ratio approximation to relative risk in cumulative-incidence case-control studies is different from the direct estimate of relative risk in incidence-density and case-exposure studies. Understanding this distinction, one may explain why the rare disease assumption is unnecessary for incidence-density studies. This explanation is not based on distinguishing between incidence-density rates and cumulative-incidence rates but rather on the fact that the "control" series in an incidence-density study is actually a sample of the population at risk for the purpose of estimating probability of exposure.

Case-exposure studies may be undertaken in areas of epidemiologic research normally reserved for case-control, cumulative-incidence studies. They may be useful for exploring vaccine efficacy (11) and for investigating risks associated with adverse pregnancy outcomes (19). When ultimate disease status can be determined for such studies, the case-exposure approach is nearly as powerful as the traditional case-control approach for testing differences between cases and controls and is considerably more accurate in estimating relative risk, especially when the probability of disease is greater than 0.05 or when true relative risk is greater than 2.0.

REFERENCES

- Miettinen OS. Estimability and estimation in case-referent studies. *Am J Epidemiol* 1976;103:226-35.
- Neutra RR, Drolette ME. Estimating exposure-specific disease rates from case-control studies using Bayes' theorem. *Am J Epidemiol* 1978;108:214-22.
- Fabia J, Drolette M. Twin pairs, smoking in pregnancy and perinatal mortality. *Am J Epidemiol* 1980;112:404-8.
- Levin AA, Schoenbaum SC, Monson RR, et al. Association of induced abortion with subsequent pregnancy loss. *JAMA* 1980;243:2495-9.
- Shelton JD, Jacobson JE, Orenstein WA, et al. Measles vaccine efficacy: influence of age at vaccination vs. duration of time since vaccination. *Pediatrics* 1978;62:961-4.
- Sayvetz TA. Study of measles vaccine efficacy in a private school. Presented at the 30th Annual Epidemic Intelligence Service Conference, Centers for Disease Control, Atlanta, GA, April 21, 1981.
- Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *JNCI* 1951;11:1269-75.
- Fleiss JL. *Statistical methods for rates and proportions*. Second edition. New York: John Wiley and Sons, 1981.
- Orenstein WA, Sirotkin B, Bernier R, et al. Evaluation of vaccine efficacy in case-control studies. Presented at the 109th Annual Meeting of the American Public Health Association, Los Angeles, CA, November 1-5, 1981.
- MacMahon B, Pugh TF. *Epidemiology: principles and methods*. Boston: Little, Brown and Company, 1970:268-73.
- Orenstein W, Marks J, Hogue C, et al. Vaccine efficacy—a new application of case-control and case-exposure methods. Abstract. *Am J Epidemiol* 1982;116:557.
- Edling C. Anesthetic gases as an occupational hazard—a review. *Scand J Work Environ Health* 1980;6:85-93.
- Binkin N, Okun AH, Cates W Jr. Spontaneous abortion clusters: when should they be investigated? Presented at the Epidemic Intelligence Service Conference, Centers for Disease Control, Atlanta, GA, April 20, 1981.
- Kupper LL, McMichael AJ, Spirtas R. Hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assn* 1975;70:524-8.
- Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol* 1982;116:547-53.
- Silverman DT, Hoover RN, Swanson MW. Artificial sweeteners and lower urinary tract cancer: hospital vs. population controls. Abstract. *Am J Epidemiol* 1981;114:440-1.
- Layde P, Webster L, Wingo P, et al. Breast cancer and oral contraceptives. Abstract. *Am J Epidemiol* 1982;116:562.
- Cole P, Monson RR, Haning H, et al. Smoking and cancer of the lower urinary tract. *N Engl J Med* 1971;284:129-34.
- Hogue C, Gaylor D. Case-exposure studies: a new, simplified approach to relative risk. Abstract. *Am J Epidemiol* 1981;114:427.
- Rothman KJ, Boice JD Jr. *Epidemiologic analysis with a programmable calculator*. Washington, DC: US GPO, 1979. (NIH Publication No. 79-1649).

- 21. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI* 1959;22:719-48.
- 22. Abel EL. Smoking during pregnancy: a review of effects on growth and development of offspring. *Hum Biol* 1980;52:593-625.
- 23. Bishop YMM, Fienberg SE, Holland PW. *Discrete multivariate analysis: theory and practice*. Boston: MIT Press, 1975.
- 24. Gart JJ, Zweifel JR. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* 1967;54:181-7.

APPENDIX: VARIANCE AND CONFIDENCE LIMITS

The approximate variances for the estimators of relative risk are obtained from the first terms of the Taylor series expansions of the relative risk estimators (23). This is the approach used by Neutra and Drolette (2).

The estimator given by equation 5, \hat{R}_1 , may be written in the form

$$\hat{R}_1 = \frac{\hat{p}_1}{1 - \hat{p}_1} \cdot \frac{1 - \hat{P}}{\hat{P}}$$

where $\hat{p}_1 = P(E|D)$ and $\hat{P} = \hat{P}(E|D) \cdot \hat{P}(D) + \hat{P}(E|\bar{D}) \cdot \hat{P}(\bar{D})$ is an estimate of the probability of exposure, $P(E)$. Taking logarithms and using the first terms of the Taylor's series expansion (23) gives

$$\hat{V}[\ln \hat{R}_1] \approx \frac{1}{n_D \hat{p}_1 + 0.5} + \frac{1}{n_{\bar{D}}(1 - \hat{p}_1) + 0.5} + \frac{1}{n_p \hat{P} + 0.5} + \frac{1}{n_p(1 - \hat{P}) + 0.5} - \frac{2\hat{P}(D)}{n_D \hat{P} + 0.5} - \frac{2\hat{P}(D)}{n_{\bar{D}}(1 - \hat{P}) + 0.5}$$

where n_D is the number of disease cases, $n_{\bar{D}}$ is the number of nondisease controls, $n_p = \hat{P}(1 - \hat{P})/\hat{V}[\hat{P}]$ and

$$\hat{V}[\hat{P}] = \frac{[\hat{P}(D)]^2 \hat{P}(E|D) \hat{P}(\bar{E}|D)}{n_D} + \frac{[\hat{P}(\bar{D})]^2 \hat{P}(E|\bar{D}) \hat{P}(\bar{E}|\bar{D})}{n_{\bar{D}}} + \frac{[\hat{P}(E|D) - \hat{P}(E|\bar{D})]^2 \hat{P}(D) \hat{P}(\bar{D})}{n}$$

where n is the size of the auxiliary sample used to estimate $P(D)$.

For the estimator, \hat{R}_2 , given by equation 6, we use the following as an estimator for the variance of the logarithm of the odds ratio, as proposed by Gart and Zweifel (24):

$$\hat{V}[\ln \hat{\Theta}] = \frac{1}{n_{11} + 0.5} + \frac{1}{n_{12} + 0.5} + \frac{1}{n_{21} + 0.5} + \frac{1}{n_{22} + 0.5}$$

where n_{11} is the number of exposed cases, n_{12} is the number of nonexposed cases, n_{21} is the number of exposed controls, and n_{22} is the number of nonexposed controls in a case-control study. Using the Taylor series expansion method,

$$\hat{V}[\ln \hat{R}_2] \approx \frac{1}{\hat{R}_2^2} \{ \hat{\Theta}^2 [\hat{P}(\bar{D}|E)]^2 \hat{V}[\ln \hat{\Theta}] + (\hat{\Theta} - 1)^2 \hat{V}[\hat{P}(D|E)] \}$$

When a sample size n_E exposed individuals is used to estimate the auxiliary probability of disease, $P(D|E)$,

$$\hat{V}[\hat{P}(D|E)] = \frac{\hat{P}(D|E) \cdot \hat{P}(\bar{D}|E)}{n_E}$$

For the estimator, \hat{R}_3 , given by equation 7, the approximate variance is

$$\hat{V}[\ln \hat{R}_3] \approx \left[\frac{\hat{R}_3}{\hat{\theta}} \right]^2 \{ [\hat{P}(\bar{D}|\bar{E})]^2 \hat{V}[\ln \hat{\theta}] + (\hat{\theta} - 1)^2 \hat{V}[\hat{P}(D|\bar{E})] \},$$

where $\hat{V}[\ln \hat{\theta}]$ is given above. If a sample of size $n_{\bar{E}}$ nonexposed individuals is used to estimate the auxiliary probability of disease, $P(D|\bar{E})$, then

$$\hat{V}[\hat{P}(D|\bar{E})] = \frac{\hat{P}(D|\bar{E}) \cdot \hat{P}(\bar{D}|\bar{E})}{n_{\bar{E}}}.$$

Approximate confidence limits for the logarithm of each estimator— $\ln \hat{R}_1$, $\ln \hat{R}_2$, $\ln \hat{R}_3$ —are similar. For example, for $\ln \hat{R}_3$ they are given by $\ln \hat{R}_3 \pm Z\sqrt{\hat{V}[\ln \hat{R}_3]}$, where Z is the standard normal deviate corresponding to the desired confidence level. Approximate confidence limits for R_3 are then obtained by taking antilogarithms.