

EXCLUDING CONTROLS: MISAPPLICATIONS IN CASE-CONTROL STUDIES¹

JAY H. LUBIN AND PATRICIA HARTGE

In many hospital-based case-control studies, investigators exclude from the control group individuals hospitalized for diseases related to the exposure under investigation. This procedure eliminates a bias that would otherwise operate. The underlying principle has been discussed at length (1-3) and is treated in most epidemiology texts (4, 5). Some investigators take the exclusionary principle one step further and exclude not only those hospitalized for diseases related to the exposure under study but also those with a history of such diseases. This exclusion biases the estimate of the relative risk.

If the disease histories used to exclude potential controls are positively correlated with the exposure, the estimated relative risk will be inflated. This is probably the most common situation. Conversely, if the disease histories are negatively correlated with the exposure, the estimates will be deflated. The same prin-

and the bias introduced by exclusion based on disease history, quantifying the relationship between the mistaken exclusionary criterion and the resulting bias. We also provide examples.

THE EXCLUSIONARY PRINCIPLE

The following derivations pertain to a hospital-based case-control study, but they can easily be extended to a death-certificate-based case-control study. We assume that hospital referral patterns for the diseases under study are similar and that patients are hospitalized either for the disease of interest, D_1 , or the referent disease, D_2 . Let D_0 denote the disease-free status in the population from which the patients arise. Suppose $X = 1$ and $X = 0$ denote whether a subject is or is not exposed, respectively. The odds ratio (OR) relating exposure X to the disease group, D_i , and the disease-free referent group, D_0 , is given by

$$OR_i = \{P(X = 1|D_i)/P(X = 0|D_i)\} / \{P(X = 1|D_0)/P(X = 0|D_0)\}.$$

ciples apply to other types of case-control studies, e.g., those based on death certificates.

In the next two sections, we sketch the statistical basis of the exclusionary rule

In a hospital-based study, the D_0 group is unobserved so that its exposure odds (the denominator) must be estimated from another source. Computing the ratio of the D_1 and D_2 specific odds ratios gives

$$\frac{\{P(X = 1|D_1)/P(X = 0|D_1)\} / \{P(X = 1|D_0)/P(X = 0|D_0)\}}{\{P(X = 1|D_2)/P(X = 0|D_2)\} / \{P(X = 1|D_0)/P(X = 0|D_0)\}} \\ = \{P(X = 1|D_1)/P(X = 0|D_1)\} / \{P(X = 1|D_2)/P(X = 0|D_2)\}.$$

¹ Environmental Epidemiology Branch, Division of Cancer Cause and Prevention, National Cancer Institute, Department of Health and Human Services, Bethesda, MD.

Reprint requests to Dr. Jay H. Lubin, Environmental Epidemiology Branch, National Cancer In-

stitute, Landow Building, Room 3C09, Bethesda, MD 20205.

The authors wish to acknowledge Dr. Mitchell Gail for stimulating discussions on this and related topics.

The right side of this equation is precisely the formula for the odds ratio wherein D_2 is the referent group, although the estimated parameter is now OR_1/OR_2 . Since OR_1/OR_2 is the estimated parameter in these types of studies, in the absence of specific knowledge about OR_2 , the "exclusionary" principle states that the control diseases should be chosen from those thought to be unrelated to the exposure of interest, i.e., $OR_2 = 1$. Note that if analysis reveals that the odds ratio varies by choice of specific control disease type, valuable information concerning the association between the exposure and the control types may be obtained. It also

carried out only for the controls, the estimate of the odds ratio is biased.

Since the exclusions occur only among the control group, D_2 , we need only focus on its exposure odds. Suppose a constraint, C , is applied to the selection of eligible D_2 patients. Intuitively, if C is positively associated with exposure, excluding controls with C would reduce the probability of exposure in the control sample, thus deflating the control exposure odds and biasing the observed odds ratio upward. To be explicit, let C be 1 if the constraint is present and 0 if not. In terms of C , the probability of observing X can be written as

$$P(X|D_2) = P(X|D_2, C = 1) \times P(C = 1) + P(X|D_2, C = 0) \times P(C = 0). \quad (1)$$

should be pointed out that if the exposure of interest changes, the control diseases must be reevaluated for a relationship

Rewriting equation 1, we obtain the probability of observing X for a D_2 control without the constraint

$$P(X|D_2, C = 0) = \{P(X|D_2) - P(X|D_2, C = 1) \times P(C = 1)\} / P(C = 0). \quad (2)$$

with the new exposure.

CONTROL EXCLUSIONS BASED ON CONSTRAINTS

Unfortunately, some researchers misinterpret the exclusionary rule and, while selecting controls hospitalized for dis-

Exclusion based on the constraint C amounts to use of equation 2 for the control exposure odds rather than equation 1. These odds can be expressed as $\{P(X = 1|D_2)/P(X = 0|D_2)\}/B$, implying that the odds ratio being estimated is (OR_1/OR_2) times a bias B , where

$$B = \frac{1 - P(C = 1|D_2, X = 0) \times P(C = 1)/P(C = 1|D_2)}{1 - P(C = 1|D_2, X = 1) \times P(C = 1)/P(C = 1|D_2)}. \quad (3)$$

eases unrelated to exposure, reject patients who have a *history* of exposure-related illnesses. If this type of exclusion were applied consistently in the selection of both cases and controls, the odds ratios would be unbiased (the observed OR_1/OR_2 being conditional on the occurrence of a negative history), although perhaps of limited interest. However, if exclusion is

The bias is seen to be a function of the prevalence of the constraint within the population and the strength of the association between the constraining condition and the exposure. When one takes C to be the prevalence of some disease, an inflationary bias is introduced when disease C occurs more often with exposure than without.

ILLUSTRATION

Hypothetic data on lung cancer and cigarette use serves to illustrate the degree to which the bias can act. Suppose there are 90 cases of lung cancer (D_1) and an equal number of controls with a condition unlikely to be related to tobacco use, for example, accidents (D_2). Assuming no hospital referral differences, the association between lung cancer and cigarette smoking may be represented in a 2×2 table

	$X = 1$	$X = 0$	
D_1	85	5	90
D_2	60	30	90

where $X = 1$ denotes smoker and $X = 0$ denotes nonsmoker. The observed odds ratio is 8.5. Suppose accident patients were excluded if they have a history of cardiovascular disease ($C = 1$). In this instance, we can make the additional assumption that a history of cardiovascular disease and accidents are unrelated, so that $P(C = 1) = P(C = 1|D_2)$ and B reduces to the ratio of the probability of no cardiovascular disease history among nonsmokers and smokers. Among accident victims, suppose the probability of cardiovascular disease is 0.6 for smokers and 0.2 for nonsmokers. We would expect that among 90 controls, $0.6 \times 60 = 36$ of the smokers and $0.2 \times 30 = 6$ of the nonsmokers have a positive disease history, leaving 24 in each cell without a history. Thus, the exposure probability among accident patients without a history of cardiovascular disease is one-half, and with a sample of 90 subjects, the expected number in each cell would be 45. This would give an odds ratio of 17.0; the computed bias B is $(1.0 - 0.2)/(1.0 - 0.6) = 2.0$. Suppose the exclusion is based on a history of a disease of low prevalence, e.g., bladder cancer. If disease prevalence is 0.06 for smokers and 0.02 for nonsmokers

(the prevalence rate ratio being the same as for cardiovascular disease), then $B = 1.04$ and the expected odds ratio is 8.9.

DISCUSSION

A proper control series has an exposure experience that reflects that of the population from which the cases arise. A hospital control series may fail to reflect the population at risk because it includes people *admitted* to the hospital for conditions caused (or prevented) by the exposures of interest. Therefore, hospital controls *admitted* for exposure-related diseases should be excluded. By contrast, the exclusion from the control group of hospital patients because of a *history* of exposure-related diseases may render the controls incomparable to cases, because the selection of controls is subject to constraints which are not imposed equally on case selection. Such restrictions on the control series which are based on exposure-related criteria can induce an estimation bias, B , which may be quite marked depending on the prevalence of the constraint and its relationship to the exposure.

The formula for B provides a method of evaluating the possible bias with this type of exclusion; however, the case-control study does not provide data for its estimation from equation 3.

REFERENCES

1. Jick H, Vessey MP. Case-control studies in the evaluation of drug-induced illness. *Am J Epidemiol* 1978;107:1-7.
2. Ibrahim MA, ed. The case-control study: consensus and controversy. *J Chronic Dis* 1979;32:1-144.
3. Prentice RL, Breslow NE. Retrospective studies and failure time models. *Biometrika* 1978;65:153-8.
4. Lilienfeld AM. *Foundations of epidemiology*. New York: Oxford University Press, 1976.
5. Breslow NE, Day NE. *Statistical methods in cancer research*. Vol 1. The analysis of case-control studies. Scientific publication no. 32. Lyon: International Agency for Research on Cancer, 1980.