

## COMPETING RISKS BIAS ARISING FROM AN OMITTED RISK FACTOR

ARTHUR SCHATZKIN<sup>1</sup> AND ERIC SLUD<sup>2</sup>

Schatzkin, A. (NCI, Bethesda, MD 20892) and E. Slud. Competing risks bias arising from an omitted risk factor. *Am J Epidemiol* 1989;129:850-6.

The authors describe a form of selection bias that may arise when a second disease selectively removes from the population persons susceptible to the primary disease of interest. Two examples of this bias are given: 1) a lack of association between an exposure and the primary disease may appear as an inverse association, and 2) a direct association between exposure and primary disease may be greatly attenuated. These examples of bias require the presence of an unknown risk factor in addition to the exposure of interest.

epidemiologic methods; risk; risk factors

Cause-specific hazard rates are ordinarily estimated by the life table method. This method directly estimates the conditional probability of a person's dying from the specific cause in a certain time interval, given that the person survives to the beginning of the interval. The cause-specific hazard generated by this procedure has been called the "crude" hazard function (1). The life table method rests on the assumption that any other causes of loss to follow-up

("competing risks") operate independently of the cause of death under study (1-6). This independence assumption implies that subpopulations particularly susceptible (or resistant) to the primary disease are not selectively depleted (or overrepresented) among survivors of the other causes of loss to follow-up.

It has long been recognized, however, that the independence assumption is unrealistic in many instances (4, 7, 8), as when risk factors are common to several causes of death. The relation of smoking and health immediately comes to mind, smoking being implicated in the pathogenesis of coronary heart disease, cancer at multiple sites, cerebrovascular disease, and chronic lung disease (9). These causes of death are said to represent *dependent* competing risks.

We would like to construct hypothetical hazard rates that would obtain if all causes of loss to follow-up other than the specific primary cause could be suppressed. These hypothetical rates, variously called "net" (1) and "pure" (4) rates, presumably reflect the relevant pathobiologic processes underlying the primary disease of interest. When competing causes of loss to follow-up act

Received for publication July 21, 1987, and in final form June 13, 1988.

<sup>1</sup> Cancer Prevention Studies Branch, Cancer Prevention Research Program, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.

<sup>2</sup> Information Management Services, Silver Spring, MD, and Department of Mathematics, University of Maryland, College Park, MD.

Reprint requests to Dr. Arthur Schatzkin, National Cancer Institute, Blair Building, Room 6A-01, 9000 Rockville Pike, Bethesda, MD 20892-4200.

The authors are grateful for the suggestions and comments provided by Drs. David Byar, Charles C. Brown, Robert N. Hoover, and Larry G. Kessler of the National Cancer Institute and by Dr. L. Adrienne Cupples of the Boston University School of Public Health. The authors also thank Jennifer Gaegler of the National Cancer Institute for assistance in manuscript preparation and David Annett of Information Management Services, Silver Spring, Maryland, for graphics assistance.

independently of the primary disease under study, the crude and net hazard functions are identical. However, mathematicians have recognized in recent years that when the assumption of independence of competing risks does not hold, it is generally impossible to estimate the net cause-specific hazards (4, 5).

In this paper, we address the problem of interpreting survival data in two types of hypothetical settings in which the independence assumption fails and in which uncritical use of estimated crude cause-specific hazard functions would lead to serious biases.

The premise of the following examples is that, while some common exposures ( $E$  in the examples) are known and measured, others (such as  $X$  in the examples) are not. In brief, it is the ignored common risk factor  $X$  which can induce dependence between competing risks within  $E$  strata.

**EXAMPLE 1: LACK OF ASSOCIATION IS OBSERVED TO BE AN INVERSE ASSOCIATION**

Consider diseases 1 and 2 to be coronary heart disease and cancer, respectively, and the exposure of interest to be total serum cholesterol. Although the epidemiologic evidence has been inconsistent, a number of prospective cohort studies have found an inverse association between serum cholesterol and cancer (10). We emphasize that the following discussion is presented only for purposes of illustration and is not meant to be a substantive commentary on the cholesterol-cancer question.

We now suggest a set of conditions under which the cancer-cholesterol association might differ between the censored and uncensored populations. Suppose we have two exposures, cholesterol (corresponding to  $E$  in the general explanation above) and an additional exposure  $X$ , and two potential disease endpoints, coronary heart disease mortality and cancer mortality. We assume that as of time  $t = 1$  year (that is, after a single unit of time on study), the two ex-

posures, cholesterol and  $X$ , are related to the instantaneous hazards of coronary heart disease mortality and cancer mortality as shown in table 1. For purposes of illustration, we will consider both cholesterol and  $X$  dichotomous variables. The numbers in table 1 are annualized mortality (hazard) rates for cancer and coronary heart disease at one year within the given levels of cholesterol and  $X$ . Risks of death from cancer and coronary heart disease are assumed to act independently of one another within each of the four cholesterol- $X$  strata.

We also assume, only to simplify the presentation, that the (dichotomous) values of  $X$  are equally distributed within the high- and low-cholesterol categories, so that each cell within a  $2 \times 2$  table (table 1) has a probability of 0.25.

Rather than make the assumption that the time-1 hazard rates shown in table 1 all remain constant over time, we assumed that the hazard rates change over time according to a Weibull distribution (11). The Weibull shape parameters associated with the time-1 rates from table 1 are given in table 2. Although the specific figures in

TABLE 1  
Instantaneous mortality rates\* ( $\lambda\gamma$ ) at  $t = 1$  for diseases 1 and 2

	Disease 1		Disease 2	
	Low $X$	High $X$	Low $X$	High $X$
Low $E$				
High $E$	50	200	25	100

\* Annual rates per 100,000 person-years.

TABLE 2  
Weibull shape parameters ( $\gamma$ -values) used in examples 1 and 2

	Disease 1		Disease 2*	
			Low $X$	High $X$
Low $E$	1.6	1.6		
High $E$	1.6	2.4	1.6/1.7	2.0/1.9

\* The first  $\gamma$ -value in a given  $E$ - $X$  stratum is for example 1, while the second  $\gamma$ -value in a given stratum is for example 2.

tables 1 and 2 are arbitrary, the survival parameters in our examples have been chosen to yield age-specific mortality comparable to the actual age-specific mortality for US white males (12, 13). The mortality rates have been modeled to vary over time according to a power law, and the predicted mortality should be calculated according to Weibull survival functions.

One consequence of the specification of example 1 is that the patients at highest risk of failure from disease 1 (the "susceptibles") are selectively depleted from the high  $E$ -high  $X$  stratum. This can be seen quantitatively through the exhibit in table 3 of the respective proportions of the surviving population belonging at time  $t$  to each of the four cholesterol- $X$  strata. Of course, these proportions as of time 0 are all 0.25. The "depletion of susceptibles" in this example is reflected in the decrease over time of the observed proportion of high cholesterol (high  $X$  + low  $X$ ) individuals among those still at risk after time  $t$ .

Tables 1 and 2, taken together, imply the following:

TABLE 3  
*Changing population proportions over time within each cholesterol- $X$  stratum*

Year	High cholesterol	Low cholesterol
	Low $X$	
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

1) High  $X$  is a risk factor for cancer. The cancer hazard rate at time  $t = 1$  is greater for high  $X$  relative to low  $X$ . Moreover, the cancer hazard increases with time more rapidly for persons with high  $X$ , irrespective of their cholesterol values.

2) Cholesterol level is not a cancer risk factor.

3) Cholesterol is a necessary risk factor for coronary heart disease mortality and operates synergistically with  $X$  so that the coronary mortality rates at  $t = 1$  among persons with high cholesterol are about four times higher in the presence of high  $X$  than in the presence of low  $X$ . In addition, the coronary disease hazard increases more rapidly in the high cholesterol-high  $X$  stratum than in the other three strata.

On the basis of the hazards at  $t = 1$  presented in table 1 and the Weibull shape parameters shown in table 2, calculations were made of both the crude and net hazard-rate curves and corresponding survival functions from disease 2 deaths. The methods for calculating these curves from assumed survival distributions can be found in references 2 and 4. The results of these calculations, carried out separately for high- and low-cholesterol subjects, are depicted in figure 1.

Tables 1 and 2 imply that the true hazards of disease 2 mortality are the same for the high- and low-cholesterol groups. Moreover, the "crude" and "net" hazard functions for cancer death are identical within the low cholesterol group because coronary and cancer mortality are independent of one another (that is, the coronary heart disease mortality rate is the same within the low cholesterol stratum irrespective of the level of  $X$ ).

In figure 1, we see a divergence over time between the crude survival functions for the high- and low-cholesterol groups, with a lower proportion of subjects dying from cancer in the high cholesterol group compared with the low cholesterol group. In spite of the fact that the cancer hazard was truly unrelated to cholesterol level, low cholesterol, relative to high cholesterol, would

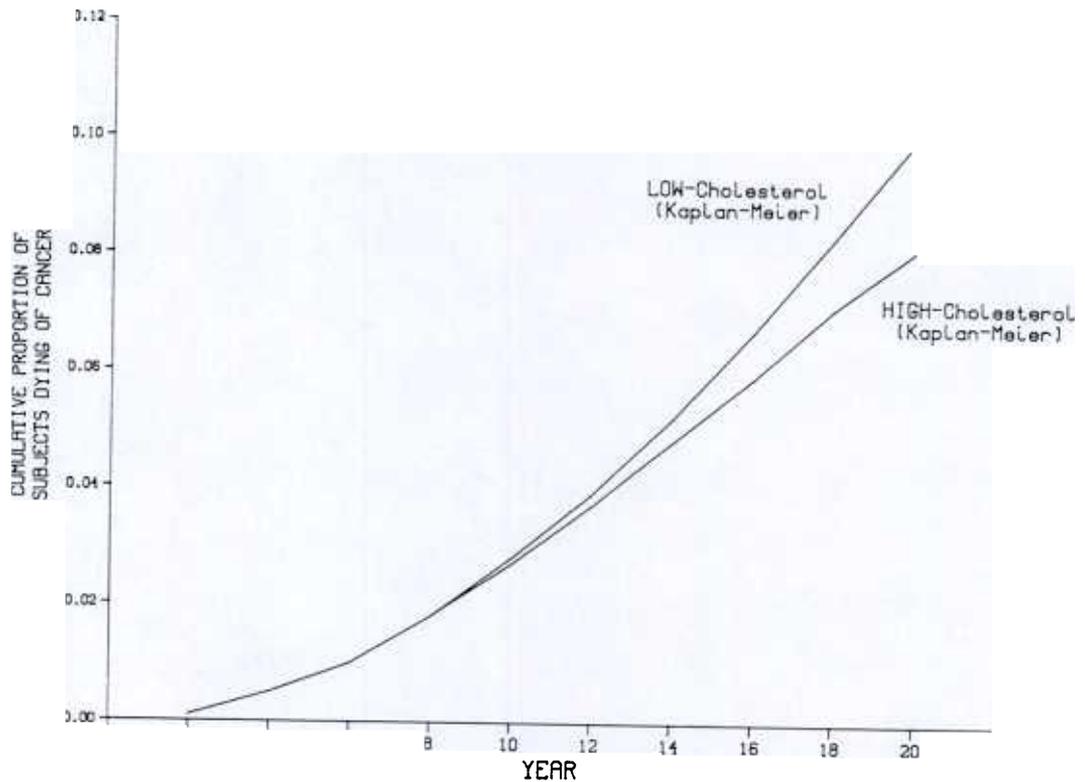


FIGURE 1. Survival curves for example 1. The net high cholesterol curve coincides with the low cholesterol curves.

appear from a life table analysis of observable data without regard to  $X$  to be associated with an excess risk of dying of cancer.

#### EXAMPLE 2: A DIRECT ASSOCIATION IS MARKEDLY ATTENUATED

In this example, we explore whether the direct association of an exposure  $E$  with a disease (disease 2) can be apparently diminished by the action of dependent competing risks bias.

We once again used the time-1 hazard (mortality) rates presented in table 1. The Weibull shape parameters for disease 1 remained the same, but slightly different shape parameters were used for disease 2 (table 2). The same assumptions— independence of disease rates within cholesterol strata and equal distribution of  $X$ -values across cholesterol categories—hold as in the first example.

As before, we calculated for disease 2 the

true (net) and crude survival functions for each of the two cholesterol groups. Figure 2 shows the results of these calculations. Three curves are presented: low  $E$  with the net and crude hazard curves being identical, true (net) high  $E$ , and crude high  $E$ . We see that the true disease 2 mortality is considerably greater for high  $E$  compared with low  $E$ . However, the excess mortality for high  $E$  is markedly reduced when one compares the crude hazard curves for the high- and low-cholesterol groups.

#### SENSITIVITY OF THE BIAS TO CHANGES IN UNDERLYING ASSUMPTIONS

The calculations for examples 1 and 2 are clearly driven by our time-1 rates (table 1) and Weibull shape parameters (table 2). Generally, higher rates in all exposure strata for diseases 1 and 2, which might characterize an older population, would cause the bias to appear earlier in follow-

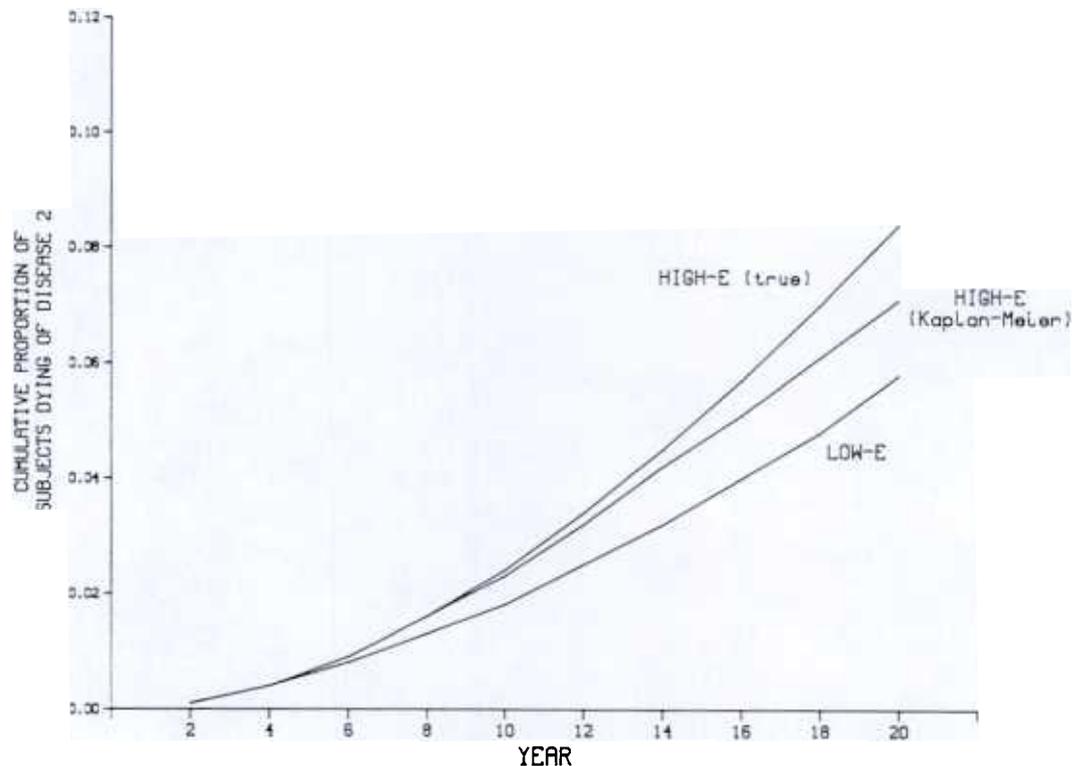


FIGURE 2. Survival curves for example 2.

up. Conversely, if the high-high to high-low difference were smaller, or if the overall rates were smaller, then the bias from dependent competing risks would present less of a problem. As we showed, the extent of the bias appears especially sensitive to the relative magnitudes of the disease 1 mortality rates in the high-high stratum versus the high-low stratum, that is, to the extent of positive interaction of  $X$  and  $E$  on disease 1. In examples of this type, we have found that removing the variation in hazard rates over time (that is, replacing all shape parameters by 1, thereby making the Weibull distributions exponential) reduces the bias due to dependence. The magnitude of the bias might also change if more than two simple exposure categories were used, or if the two risk factors were correlated. (In our two examples, the subjects were evenly distributed across the risk factor strata, implying that the two risk factors were independent. However, additional calculations indicated that the magnitude of

bias was relatively insensitive to moderate changes in the population frequencies of the four strata.)

#### DISCUSSION

We have described a form of selection bias based on the dynamic interrelation of at least two diseases and two risk factors. Two elements join to create this bias: 1) a kind of "censoring interaction," whereby censoring from one disease is greater in the presence of two risk factors than in the presence of one, and 2) absence of information on the second risk factor ( $X$ ).

If information on  $X$  were available, the disease 2 hazards for the high  $E$  and low  $E$  groups could be derived from a "mixture" of the high  $X$  and low  $X$  hazards (weighted equally, half and half) within each of these high  $E$  and low  $E$  groups. In the examples, knowledge of  $X$  would have allowed calculation of stratumwise life tables within the four  $E$ - $X$  strata. However, we presented the examples from the perspective that  $X$  was

unobserved. The different hazards of mortality within the four  $E$ - $X$  strata make disease 1 and disease 2 failures dependent within each of the high  $E$  and low  $E$  subpopulations. The crude survival functions for the high  $E$  and low  $E$  strata (without regard to  $X$ ) therefore differ from the survival functions constructed by mixing over  $X$  strata. This difference constitutes a competing risks bias attributable to an omitted exposure-covariate.

$X$  might be conceived of as a "susceptibility" factor, in which case the competing risks bias illustrated here can be conceived of as a "depletion of susceptibles" or "resistant survivor" phenomenon. This latter effect is well known in occupational studies (14) and may explain the declining mortality ratios for smokers compared with non-smokers with increasing age in some cohort studies (15). We emphasize, however, that  $X$  need not be an inherent genetic characteristic, but may well be a common exposure (or set of exposures). In other words, susceptibility to "depletion" by the competing cause of death (coronary disease, in our example) results from the presence of the second exposure(s) (high  $X$ ), which may be an exogenous exposure as well as some intrinsic characteristic.

Although we have demonstrated the possible magnitude of competing risks bias in the context of a cohort study, the bias may also be operative in the case-control setting. In our cholesterol-cancer example, the selective removal by coronary heart disease death of high cholesterol subjects with high  $X$  would occur in the general population just as it would in a cohort. The pool of cases eligible for the case-control study would therefore have already been "depleted," and the same competing risks bias illustrated in the cohort study would affect results from the case-control study.

There is no easy way to deal with potential competing risks bias. We have shown quantitatively how dependent competing risks might bias epidemiologic observations, but there is no a priori way of knowing whether the exposure-disease associa-

tion differs between the censored and uncensored populations, or whether the magnitude of such a difference is great enough to confer any meaningful bias. Our purpose here was to demonstrate that in certain situations the magnitude of the competing risks bias is worthy of serious consideration. (For a related theoretical example in which dependent competing risks might reverse the apparent effect of a dichotomous covariate on survival, see reference 16.)

If one knew what  $X$  was and could measure it, then one could perform stratified analyses of the relation between cholesterol and cancer within strata of  $X$ , or, alternatively, include the appropriate interaction terms for cholesterol and  $X$  in a statistical model. If  $X$  were known in our examples, then a correct stratified life table analysis could be performed, in which case the lack of association between cholesterol and cancer in example 1 would be immediately evident.

This argues for careful attention to interaction in epidemiologic analyses, since it is possible that  $X$  might be "captured" by some combination of measured risk factors that could then be analyzed for interaction with the exposure ( $E$ ) of interest. From a more biologic perspective, it is plausible that the development of intermediate endpoint markers (of cancer, for example) might make it possible to detect one incipient disease before censoring from another disease occurs, or to detect two diseases simultaneously, and thereby avoid the competitive censoring which leads to biased analyses.

#### REFERENCES

1. Gail M. Competing risks. In: Kotz S, Johnson NL, eds. Encyclopedia of statistical sciences. Vol 2. New York: John Wiley & Sons, 1982;75-81.
2. Birnbaum ZW. On the mathematics of competing risks. Hyattsville, MD: National Center for Health Statistics, 1979. (DHEW publication no. (PHS)79-1351). (Series 2, number 77).
3. David HA, Moeschberger ML. The theory of competing risks. New York: Macmillan Publishing Co, Inc, 1978.
4. Cornfield J. The estimation of the probability of

- developing a disease in the presence of competing risks. *Am J Public Health* 1957;47:601-7.
5. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci USA* 1975;72:20-2.
  6. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53:457-81.
  7. Chiang CL. An introduction to stochastic processes and their applications. Huntington, NY: Robert E Krieger Publishing Company, 1980.
  8. Slud EV, Rubinstein LV. Dependent competing risks and summary survival curves. *Biometrika* 1983;70:643-9.
  9. US Public Health Service. The health consequences of smoking: a report of the Surgeon General. 1972. Washington, DC: US Department of Health, Education, and Welfare, 1972. (DHEW publication no. (HSM)72-6516).
  10. McMichael AJ, Jensen OM, Parkin DM, et al. Dietary and endogenous cholesterol and human cancer. *Epidemiol Rev* 1984;6:192-216.
  11. Lee ET. Statistical methods for survival data analysis. Belmont, CA: Lifetime Learning Publications, 1980.
  12. National Center for Health Statistics. Health, United States, 1981. Hyattsville, MD: US Department of Health and Human Services, US Public Health Service, 1981. (DHHS publication no. 82-1232).
  13. Horm JW, Asire AJ, Young JL, et al, eds. SEER program: cancer incidence and mortality in the United States, 1973-1981. Bethesda, MD: US Department of Health and Human Services, US Public Health Service, National Institutes of Health, 1984. (NIH publication no. 85-1837).
  14. Marrett LD, Walter SD, Meigs JW. Coffee drinking and bladder cancer in Connecticut. *Am J Epidemiol* 1983;117:113-27.
  15. Castelli WP, Garrison RJ, Wilson PWF, et al. The incidence of coronary heart disease and lipoprotein cholesterol levels: The Framingham Study. *JAMA* 1986;256:2835-8.
  16. Slud E, Byar D. How dependent causes of death can make risk factors appear protective. *Biometrics* 1988;44:265-9.