THE B FILES

Case studies of bias in real life epidemiologic studies

Bias File 2. Should we stop drinking coffee? The story of coffee and pancreatic cancer

Compiled by

Madhukar Pai, MD, PhD

Jay S Kaufman, PhD

Department of Epidemiology, Biostatistics & Occupational Health

McGill University, Montreal, Canada

madhukar.pai@mcgill.ca & jay.kaufman@mcgill.ca

McGill

Bias File 2. Should we stop drinking coffee? The story of coffee and pancreatic cancer

The story

Brian MacMahon (1923 - 2007) was a British-American epidemiologist who chaired the Department of Epidemiology at Harvard from 1958 until 1988. In 1981, he published a paper in the *New England Journal of Medicine*, a case-control study on coffee drinking and pancreatic cancer [MacMahon B, *et al.*. 1981]. The study concluded that "coffee use might account for a substantial proportion of the cases of this disease in the United States." According to some reports, after this study came out, MacMahon stopped drinking coffee and replaced coffee with tea in his office. This publication provoked a storm of protest from coffee drinkers and industry groups, with coverage in the *New York Times*, *Time* magazine and *Newsweek*. Subsequent studies, including one by MacMahon's group, failed to confirm the association. So, what went wrong and why?

The study

From the original abstract:

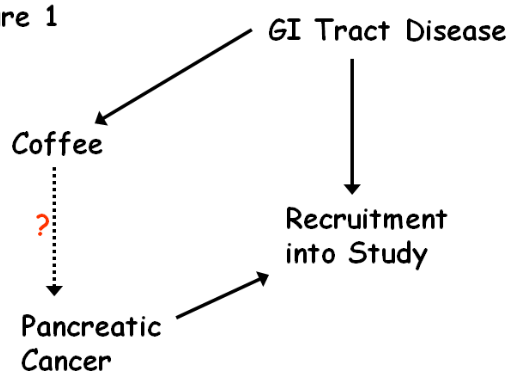
We questioned 369 patients with histologically proved cancer of the pancreas and 644 control patients about their use of tobacco, alcohol, tea, and coffee. There was a weak positive association between pancreatic cancer and cigarette smoking, but we found no association with use of cigars, pipe tobacco, alcoholic beverages, or tea. A strong association between coffee consumption and pancreatic cancer was evident in both sexes. The association was not affected by controlling for cigarette use. For the sexes combined, there was a significant dose-response relation (P approximately 0.001); after adjustment for cigarette smoking, the relative risk associated with drinking up to two cups of coffee per day was 1.8 (95% confidence limits, 1.0 to 3.0), and that with three or more cups per day was 2.7 (1.6 to 4.7). This association should be evaluated with other data; if it reflects a causal relation between coffee drinking and pancreatic cancer, coffee use might account for a substantial proportion of the cases of this disease in the United States.

The bias

The MacMahon study had several problems and several experts have debated these in various journals, but a widely recognized bias was related to control selection. A nice, easy to read explanation can be found in the Gordis text [Gordis L, 2009], but a 1981 paper by Feinstein drew attention to this problem). Controls in the MacMahon study were selected from a group of patients hospitalized by the same physicians who had diagnosed and hospitalized the cases' disease. The idea was to make the selection process of cases and controls similar. It was also logistically easier to get controls using this method. However, as the exposure factor was coffee drinking, it turned out that patients seen by the physicians who diagnosed pancreatic cancer often had gastrointestinal disorders and were thus advised not to drink coffee (or had chosen to reduce coffee drinking by themselves). So, this led to the selection of controls with higher prevalence of gastrointestinal disorders, and these controls had an unusually low odds of exposure (coffee intake). These in turn may have led to a spurious positive association between coffee intake and pancreatic cancer that could not be subsequently confirmed.

This problem can be explored further using causal diagrams. Since the study used a case-control design, cases were sampled from the source population with higher frequency than the controls, which is represented by the directed arrow between “pancreatic cancer” and “recruitment into study” in Figure 1. However, controls were selected by being hospitalized by the same doctors who treated the cases. If they were not hospitalized for pancreatic cancer, they must have been hospitalized for some other disease, which gave them a higher representation of GI tract disease than observed in the source population. Patients with GI tract disease may have been discouraged from drinking coffee, which gave controls a lower prevalence of exposure than seen in the source population. This is shown in Figure 1 as a directed arc from “GI tract disease” to coffee and to “recruitment into study”

Figure 1



Collider stratification bias occurs when one conditions (in the design or the analysis) on a common child of two parents. In this case, restricting the observations to people recruited into the study (Figure 2) changes the correlation structure so that it is no longer the same as in the source population. Specifically, pancreatic cancer and GI tract diseases may be uncorrelated in the general population. However, among patients hospitalized by the doctors who had admitted patients with pancreatic cancer, the ones who didn’t have pancreatic disease were more likely to have something else: a GI tract disease. Therefore, restriction to the population of the doctors who hospitalized the cases induces a negative correlation between these two diseases in the data set.

Figure 2

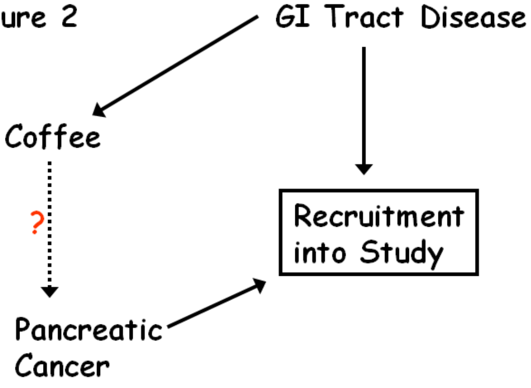
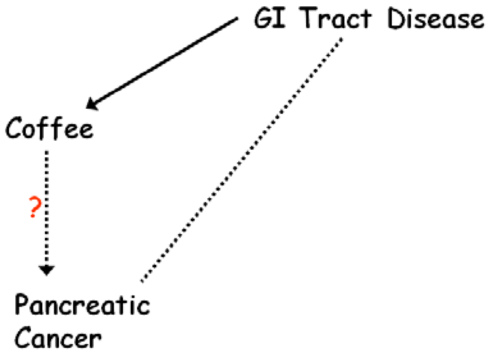


Figure 3 shows a graph of the data for the study population, as opposed to the source population. Restriction to the subjects recruited from the hospital has created a correlation between GI tract disease and pancreatic cancer. Since GI tract disease lowers exposure, an unblocked backdoor path is now opened, which leads to confounding of the estimated exposure effect (shown with a dashed line and a question mark). Specifically, since the induced correlation is negative, and the effect of GI tract disease on coffee is negative, the exposure estimate for coffee on pancreatic cancer will be biased upward (Vander Stoep et al 1999).

Figure 3 (those included in the case-control study)



The lesson

Control selection is a critical element of case-control studies, and even the best among us can make erroneous choices. Considerable thought needs to go into this critical step in study design. As Rothman et al. emphasize in their textbook (Modern Epidemiology, 2008), the two important rules for control selection are:

1. Controls should be selected from the same population - the source population (i.e. study base) - that gives rise to the study cases. If this rule cannot be followed, there needs to be solid evidence that the population supplying controls has an exposure distribution identical to that of the population that is the source of cases, which is a very stringent demand that is rarely demonstrable.

2. Within strata of factors that will be used for stratification in the analysis, controls should be selected independently of their exposure status, in that the sampling rate for controls should not vary with exposure.

A more general concern than the issue of control selection in case-control studies is the problem of selection bias (Hernán et al 2004). Whenever the epidemiologist conditions statistically (e.g. by stratification, exclusion or adjustment) on a factor affected by exposure and affected by outcome, a spurious correlation will occur in the study data-set that does not reflect an association in the real world from which the data were drawn. If there is already a non-null association between exposure and outcome, it can be shifted upwards or downwards by this form of bias.

Sources and suggested readings*

1. MacMahon B, Yen S, Trichopoulos D *et al.* Coffee and cancer of the pancreas. *N Engl J Med* 1981;304: 630–633.

2. Schmeck HM. Critics say coffee study was flawed. New York Times, June 30, 1981.

3. Gordis L. Epidemiology. Saunders, 2008.

4. Feinstein A et al. Coffee and Pancreatic Cancer. The Problems of Etiologic Science and Epidemiologic Case-Control Research. JAMA 1981;246:957-961.

5. Rothman K, Greenland S, Lash T. Modern epidemiology. Lippincott Williams & Wilkins, 3rd edition, 2008.

6. Coffee and Pancreatic Cancer. An Interview With Brian MacMahon. EpiMonitor, April/May, 1981.

7. Vander Stoep A, Beresford SA, Weiss NS. A didactic device for teaching epidemiology students how to anticipate the effect of a third factor on an exposure-outcome relation. Am J Epidemiol. 1999 Jul 15;150(2):221.

8. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25.

Image credit: Epidemiology: July 2004 - Volume 15 - Issue 4 - pp 504-508

*From this readings list, the most relevant papers are enclosed.

COFFEE AND CANCER OF THE PANCREAS

Brian MacMahon, M.D., Stella Yen, M.D., Dimitrios Trichopoulos, M.D., Kenneth Warren, M.D., and George Nardi, M.D.

Abstract We questioned 369 patients with histologically proved cancer of the pancreas and 644 control patients about their use of tobacco, alcohol, tea, and coffee. There was a weak positive association between pancreatic cancer and cigarette smoking, but we found no association with use of cigars, pipe tobacco, alcoholic beverages, or tea. A strong association between coffee consumption and pancreatic cancer was evident in both sexes. The association was not affected by controlling for cigarette use. For the sexes combined, there was a significant dose-response relation ($P \sim 0.001$); after adjustment for cigarette smoking, the relative risk associated with drinking up to two cups of coffee per day was 1.8 (95 per cent confidence limits, 1.0 to 3.0), and that with three or more cups per day was 2.7 (1.6 to 4.7). This association should be evaluated with other data; if it reflects a causal relation between coffee drinking and pancreatic cancer, coffee use might account for a substantial proportion of the cases of this disease in the United States. (N Engl J Med. 1981; 304:630-3.)

OVER the past few decades, cancer of the pancreas has emerged as one of the most important neoplasias in human beings. It now accounts for approximately 20,000 deaths annually in the United States. Causative factors have been sought in several previous studies, but only cigarette smoking has emerged as a consistent, though relatively weak, exogenous risk factor. We report the results of a study that was planned to reevaluate the relation of this disease to smoking and to examine the role of alcohol consumption as a possible confounding variable. Data were also obtained on intake of tea and coffee — factors that have not been adequately investigated in this disease.

Methods

We conducted a case-control interview study. The cases were patients with histologic diagnoses of cancer of the exocrine pancreas who were in any of 11 large hospitals in the Boston metropolitan area and Rhode Island between October 1974 and August 1979. Patients with tumors of the islet cells, periampullary duodenal mucosa, or ampulla of Vater were not included. We identified 578 patients and interviewed 405 of them. Twenty patients died and 35 were discharged before an interview could be arranged; 78 were too sick to be interviewed, 14 had language difficulties, and 26 refused the interview. Also excluded from the analysis were eight nonwhite patients, four residents of countries other than the United States, eight patients older than 79 years, and 16 patients whose interview information was judged by the interviewer to be of questionable reliability. The analysis is based on data from the remaining 369 patients.

To assemble a control series, the interviewers also attempted to question all other patients who were under the care of the same physician in the same hospital at the time of an interview with a patient with pancreatic cancer. Either before the interview (if the information was known) or afterward, patients with diseases of the pancreas or hepatobiliary tract or diseases known to be associated with smoking or alcohol consumption were excluded. The principal diagnostic categories excluded (in addition to diseases of the biliary tract or pancreas) were cardiovascular disease, diabetes mellitus, respiratory or bladder cancer, and peptic ulcer. From a total of 1118 eligible patients, we interviewed 700; nine died and 131 were discharged before the interview, 179 were too ill, 26 had language problems, and 73 refused. After exclusion of 17 nonwhites, five foreign residents, four persons older than 79 years, and 30 persons whose interviews were judged to be unreliable, the control series used for the analysis consisted of 644 patients. Minor differences between tables in the stated numbers of cases and controls resulted from absence of specific items being analyzed in a few questionnaires.

The control series was composed of two principal diagnostic groups: 273 patients with cancer other than cancers of the pancreas and biliary tract, respiratory tract, or bladder and 371 patients with other disorders. Of the control patients with cancer, the tumor was in the breast in 65 patients, colon in 60, rectum in 25, stomach in 24, small intestine in nine, ovary in eight, prostate in eight, and cervix in seven; there were also 16 with melanoma and 15 with lymphoma. No other cancer was found in more than four subjects. Diagnoses in the controls without cancer were of a wide variety, although because of the nature of the practices of many of the physicians who were responsible for patients with cancer of the pancreas, patients with gastroenterologic conditions were probably overrepresented in relation to a general hospital population. The principal diagnoses were hernia in 70 patients; colitis, enteritis, or diverticulitis in 41; bowel obstruction, adhesions, or fistula in 26; gastritis in 17; other gastroenterologic conditions in 47; benign tumors in 29; varicose veins or phlebitis in 21; genitourinary disorders in 20; neurologic disorders in 20; gynecologic disorders in 16; and other conditions in 64.

In the analyses, the patients with pancreatic cancer were compared with the control patients with cancer and independently with the control group without cancer. The findings were quite similar, and only the results with the combined control group are presented here.

Several questions in the interview probed the duration and intensity of smoking of cigarettes, cigars, and pipes. Questions on alcoholic beverages asked about the frequency of use before the onset of illness, the age span over which such use occurred, and the type of beverage used most frequently. The questions on tea and coffee were limited to the number of cups consumed in a typical day before the current illness was evident.

Tests of significance and estimates of adjusted relative risks and their confidence limits were derived with the method of Mantel and Haenszel[1] and its extension.[2] The data were stratified by age in 10-year groups and by sex where appropriate. All confidence limits are 95 per cent intervals. Most analyses were performed with the calculator programs developed by Rothman and Boice.[3]

Results

Tobacco

There was no difference between cases and controls in the use of cigars or pipe tobacco. Among men, the relative risk associated with use of cigars (with nonsmokers as the referent group) was 1.0 (confidence interval, 0.7 to 1.4), and that with use of a pipe was 1.0 (confidence interval, 0.7 to 1.4).

From the Department of Epidemiology, Harvard School of Public Health, 677 Huntington Ave., Boston, MA 02115, where reprint requests should be addressed to Dr. MacMahon.

Supported by a grant (5 P01 CA 06373) from the National Cancer Institute.

ARREN, M.D.,

tment for cig-
sociated with
r day was 1.8
3.0), and that
.7 (1.6 to 4.7).
ith other data;
offee drinking
ht account for
of this disease
1981; 304:630-

the control series
Minor differences
l controls resulted
n a few question-

ncipal diagnostic
ers of the pancre-
and 371 patients
cancer, the tumor
m in 25, stomach
tate in eight, and
and 15 with lym-
four subjects. Di-
f a wide variety,
s of many of the
cancer of the pan-
s were probably
population. The
olitis, enteritis, or
, or fistula in 26;
ns in 47; benign
nitourinary disor-
c disorders in 16;

cancer were com-
dependently with
ere quite similar,
oup are presented

duration and in-
Questions on al-
e before the onset
red, and the type
on tea and coffee
a typical day be-

relative risks and
od of Mantel and
atified by age in
confidence limits
formed with the
Boice.³

s and controls
ong men, the
rs (with non-
confidence in-
pipe was 1.0

Table 1. Distribution of Cases and Controls According to Cigarette-Smoking Habits and Estimates of Risk Ratios.

Sex	Category	Never Smoked	Ex-Smokers	Current Smokers		Total *
				<1 pack/day	>1 pack/day	
Men	Cases (no.)	40	99	22	57	218
	Controls (no.)	74	122	35	75	306
	Adjusted relative risk †	1.0	1.4	1.1	1.4	1.4
	95% confidence interval	—	0.9–2.3	0.5–2.2	0.9–2.4	0.9–2.2
Women	Cases (no.)	62	41	20	26	149
	Controls (no.)	160	86	36	55	337
	Adjusted relative risk †	1.0	1.3	1.5	1.6	1.5
	95% confidence interval	—	0.8–2.2	0.8–2.8	0.9–2.9	1.0–2.2

*Adjusted relative risks and 95 per cent confidence intervals in this column are for consumers of any amount (including ex-consumers) as compared with nonconsumers.

†Mantel–Haenszel estimates of risk ratios, adjusted over categories of age in decades. In all comparisons, the referent category was subjects who had never smoked. Chi-square (Mantel extension) with equally spaced scores, adjusted over age in decades: 1.2 for men, 4.1 for women.

The data on use of cigarettes are shown in Table 1. There was a weak positive association. Although only the data for women showed a significant dose-response relation, the estimate of the relative risk associated with smoking at any time for both sexes combined was 1.4; the difference from the referent risk was significant (confidence interval, 1.1 to 1.9).

Alcohol

Table 2 shows a comparison of use of alcoholic beverages by cases and by controls. No notable or significant association appeared. The combined estimate of relative risk associated with drinking at any time was 0.9, with a confidence interval of 0.6 to 1.3, and that associated with regular drinking was 0.8 (confidence interval, 0.5 to 1.3).

No difference between cases and controls was found in the statements about the type of alcoholic beverage used most frequently (data not shown).

Table 2. Distribution of Cases and Controls According to Alcohol-Drinking Habits and Estimates of Risk Ratios.

Sex	Category	Alcohol Drinking			Total
		none	occasional	regular	
Men	Cases (no.)	16	113	89	218
	Controls (no.)	27	157	123	307
	Adjusted relative risk *	1.0	1.3	1.3	1.3
	95% confidence interval	—	0.7–2.6	0.6–2.6	0.7–2.5
Women	Cases (no.)	33	99	17	149
	Controls (no.)	59	221	57	337
	Adjusted relative risk *	1.0	0.8	0.5	0.8
	95% confidence interval	—	0.5–1.3	0.3–1.1	0.5–1.3

*Chi-square (Mantel extension) with equally spaced scores, adjusted over age in decades: 0.2 for men, 2.7 for women. All data are analyzed as in Table 1.

Tea

The tea consumption of cases and controls is shown in Table 3. A slight inverse association appeared in both sexes, but it was not significant in either.

Coffee

An unexpected association of pancreatic cancer with coffee consumption was evident (Table 4). Among men, each category of coffee consumption had a statistically significant excess risk as compared with that of nondrinkers of coffee, but the dose-response relation was flat. Among women, both categories of consumers of three or more cups per day had significantly elevated risks, and the dose-response relation (as measured by the Mantel test) was highly significant ($P<0.001$). For the sexes combined, with a simultaneous adjustment for sex and age, the trend was also highly significant (chi-square, 11.0), and the adjusted relative risks for consumers of no cups per day, one to two, three to four, and at least five were 1.0, 2.1, 2.8, and 3.2, respectively.

Table 3. Distribution of Cases and Controls According to Tea-Drinking Habits and Estimates of Risk Ratios.

Sex	Category	Tea Drinking (Cups per Day)			Total
		0	1-2	>3	
Men	Cases (no.)	61	134	21	216
	Controls (no.)	72	205	29	306
	Adjusted relative risk *	1.0	0.7	0.8	0.7
	95% confidence interval	—	0.5–1.1	0.4–1.5	0.5–1.1
Women	Cases (no.)	40	85	25	150
	Controls (no.)	75	191	70	336
	Adjusted relative risk *	1.0	0.7	0.6	0.7
	95% confidence interval	—	0.5–1.2	0.3–1.2	0.5–1.2

*Chi-square (Mantel extension) with equally spaced scores, adjusted over age in decades: 1.4 for men, 1.9 for women. All data are analyzed as in Table 1.

Interaction

Since no association was observed with use of alcoholic drinks, tea, pipe tobacco, or cigars, the principal interaction of interest was that between cigarette use and coffee use. This relation was explored in the analysis presented in Table 5. The data showed a consistent association of pancreatic cancer with coffee drinking within each category of smoking, and the data for all smokers and nonsmokers showed a consistent trend with coffee drinking after adjustment for smoking. With the Mantel extension, the chi-square value for the trend with coffee consumption (after adjustment for smoking as well as age and sex) was 10.6 ($P \sim 0.001$). The association with smoking within categories of coffee consumption was less clear, and the relative risks for ex-smokers and current smokers, adjusted for coffee consumption, did not differ significantly from unity.

Discussion

Our findings with regard to association of cancer of the pancreas with cigarette use and alcohol consumption are consistent with those of previous investigators. The association with cigarette use has been most extensively explored. Weakly positive associations were found in two other case-control studies[4,5] and in the large cohort studies in British physicians,[6] American veterans,[7] and the American Cancer Society population.[8] The relative risks for cigarette smokers as compared with nonsmokers were 2.3 in the larger case-control study and 1.6, 1.8, and an average of 2.2 in the three cohort studies. These values are comparable to the figure of 1.4 in our study. In one small case-control study, a weak and nonsignificant association was found only in women; among men, there was no difference in cigarette-smoking habits between cases and controls.[5] However, the inclusion of patients with smoking-related diseases among the hospitalized controls in that study would have served to conceal a weak relation. Adjustment for coffee consumption did not entirely remove the association with cigarette smoking in our own data, although the association was not significant after such adjustment. The possible confounding influence of coffee consumption was not evaluated in the other studies.

The relation between alcohol use and pancreatic cancer has been less extensively studied, but a lack of association has been found in one case-control study[4] and in a proportional mortality analysis of a large series of deaths of alcoholics.[9] An association with wine drinking was reported in one study, but the numbers were relatively small, the difference was not conventionally significant, and potential confounding factors were not evaluated.[5] Overall, it seems unlikely that alcohol consumption has any role in the origin of cancer of the pancreas — an observation that is of some interest in the light of the obvious role of this substance in chronic pancreatitis.

In a recently reported case-control study involving 94 patients with pancreatic adenocarcinoma and a similar number of hospital controls, Lin and Kessler noted that the cases tended to drink more decaffeinated coffee than did the controls.[5] In view of the relatively recent use of decaffeinated coffee on a large scale, it seems unlikely that this particular type of beverage has a causal relation to cases of pancreatic cancer appearing at present. It seems more likely that the high consumption of decaffeinated coffee noted by Lin and Kessler is a reflection of generally high coffee consumption by these patients in the past. These authors gave no data on the use of regular coffee by their subjects.

Table 4. Distribution of Cases and Controls by Coffee-Drinking Habits and Estimates of Risk Ratios.

Sex	Category	Coffee Drinking (Cups per Day)				Total
		0	1–2	3–4	>5	
Men	Cases (no.)	9	94	53	60	216
	Controls (no.)	32	119	74	82	307
	Adjusted relative risk *	1.0	2.6	2.3	2.6	2.6
	95% confidence interval	—	1.2–5.5	1.0–5.3	1.2–5.8	1.2–5.4
Women	Cases (no.)	11	59	53	28	151
	Controls (no.)	56	152	80	48	336
	Adjusted relative risk *	1.0	1.6	3.3	3.1	2.3
	95% confidence interval	—	0.8–3.4	1.6–7.0	1.4–7.0	1.2–4.6

*Chi-square (Mantel extension) with equally spaced scores, adjusted over age in decades: 1.5 for men, 13.7 for women. All data are analyzed as in Table 1.

Table 5. Estimates of Relative Risk of Cancer of the Pancreas Associated with Use of Coffee and Cigarettes.*

Cigarette Smoking	Coffee Drinking (Cups per Day)			Total †
	0	1–2	≥3	
Never	1.0	2.1	3.1	1.0
Ex-smokers	1.3	4.0	3.0	1.3 (0.9–1.8)
Current smokers	1.2	2.2	4.6	1.2 (0.9–1.8)
Total †	1.0	1.8 (1.0–3.0)	2.7 (1.6–4.7)	

*The referent category is the group that uses neither cigarettes nor coffee. Estimates are adjusted for sex and for age in decades.

†Values are adjusted for the other variable, in addition to age and sex, and are expressed in relation to the lowest category of each variable. Values in parentheses are 95 per cent confidence intervals of the adjusted estimates.

Although the positive association with coffee consumption that we observed must be evaluated with other data before serious consideration is given to the possibility of a causal relation, it is worth noting that some of the descriptive features of the epidemiology of cancer of the pancreas seem to be consistent with such a relation. The apparent increase in frequency of cancer of the pancreas in recent decades[10] and the low rates observed in Mormons[11,12] and Seventh-Day Adventists[13] would be compatible with a causative role for either coffee consumption or cigarette smoking. However, the relatively small excess of men with the disease in proportion to women would seem to be more suggestive of a role for coffee rather than for cigarettes. Some 10 years ago, correlating trade statistics in 20 countries with rates of death from cancer, Stocks reported a positive correlation between coffee consumption and rates of pancreatic cancer; the association was present in both sexes, although it was significant only in men.[14] We note also the recent report of the simultaneous occurrence of cancer of the pancreas in a husband and wife who both added "coffee syrup" to ground coffee before percolating it.[15]

Our use of a control group composed of hospitalized patients must be discussed. It is possible that these patients reduced their coffee consumption because of illness and that their replies were affected, even though the question was related to the time before the onset of their illness. Indeed, Rosenberg et al. reported a lower proportion of coffee consumers among hospitalized women with chronic disease than among women admitted for emergencies.[16] However, the differences noted by Rosenberg et al. between patients with acute and chronic illness were much smaller than those between the cases and controls in our study. Although the majority of control patients in our series had chronic disease, pancreatic cancer itself is a chronic disease, and in theory it would seem as likely as any other disorder to induce a change in coffee consumption. It is a matter for speculation whether such a bias is likely to be greater in our case series or in patients with the diagnoses represented in our control series. It is inconceivable that this bias would account for the total difference between cases and controls, but it is possible that risk may be either overestimated or underestimated on this account. We note, however, that the relative risks shown in Table 4 were similar whether the patients with other cancers or the patients with nonmalignant disorders were used as the control group.

If the association between coffee consumption and pancreatic cancer is confirmed and found to be causal, the relation will have some importance in quantitative terms. Cancer of the pancreas is now the fourth most common fatal malignant disease in the United States. If the distribution of coffee consumption in our control group reflects that in the general population, with relative risks of 1.8 associated with the use of one to two cups daily and 2.7 associated with three or more cups daily, we estimate the proportion of pancreatic cancer that is potentially attributable to coffee consumption to be slightly more than 50 per cent.[17] This estimate emphasizes the need to determine whether the association exists in other data and to evaluate its causal or noncausal nature.

We are indebted to the administrative, nursing, and record-room staffs of Beth Israel Hospital, Carney Hospital, Lahey Clinic, Massachusetts General Hospital, New England Baptist Hospital, New England Deaconess Hospital, Peter Bent Brighan Hospital, Rhode Island Hospital, Tufts-New England Medical Center, University Hospital, and the Veterans Administration Hospital of Jamaica Plain, Mass.; to the physicians on their staffs who gave us permission to interview patients; and to Mrs. Kim Neave and Miss Mary Curran for conducting the interviews.

REFERENCES

1. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959; 22:719-48.
2. Mantel N. Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. J Am Stat Assoc. 1963; 58:690-700.
3. Rothman KJ, Boice JD. Epidemiologic analysis with a programmable calculator. Washington, D.C.: Government Printing Office, 1979. (DHEW publication no. (NIH)79-1649).
4. Wynder EL, Mabuchi K, Maruchi N, Fortner JG. Epidemiology of cancer of the pancreas. J Natl Cancer Inst. 1973; 50:645-67.
5. Lin RS, Kessler II. A multifactorial model for pancreatic cancer in man: epidemiologic evidence. JAMA. 1981; 245:147-52.
6. Doll R, Peto R. Mortality in relation to smoking: 20 years' observations of male British doctors. Br Med J. 1976; 2:1525-36.
7. Kahn HA. The Dorn study of smoking and mortality among U.S. veterans: report on eight and one-half years of observation. Natl Cancer Inst Monogr. 1966; 19:1-125.
8. Hammond EC. Smoking in relation to the death rates of one million men and women. Natl Cancer Inst Monogr. 1966; 19:127-204.
9. Monson RR, Lyon JL. Proportional mortality among alcoholics. Cancer. 1975; 36:1077-9.
10. Devesa SS, Silverman DT. Cancer incidence and mortality trends in the United States: 1935-74. J Natl Cancer Inst. 1978; 60:545-71.
11. Lyon JL, Gardner JW, West DW. Cancer incidence in Mormons and non-Mormons in Utah during 1967-75. J Natl Cancer Inst. 1980; 65:1055-62.
12. Enstrom JE. Cancer mortality among Mormons in California during 1968-75. J Natl Cancer Inst. 1980; 65:1073-82.
13. Phillips RL, Garfinkel L, Kuzma JW, Beeson WL, Lotz T, Brin B. Mortality among California Seventh-Day Adventists for selected cancer sites. J Natl Cancer Inst. 1980; 65:1097-108.
14. Stocks P. Cancer mortality in relation to national consumption of cigarettes, solid fuel, tea and coffee. Br J Cancer. 1970; 24:215-25.
15. Ferguson LJ, Watts JM. Simultaneous cancer of the pancreas occurring in husband and wife. Gut. 1980; 21:537-40.
16. Rosenberg L, Slone D, Shapiro S, Kaufman DW, Stolley PD, Miettinen OS. Coffee drinking and myocardial infarction in young women. Am J Epidemiol. 1980; 111:675-81.
17. Cole P, MacMahon B. Attributable risk percent in case-control studies. Br J Prev Soc Med. 1971; 25:242-4.

Welcome to TimesPeople
Get Started
TimesPeople Lets You Share and Discover the Best of NYTimes.com 0:15 AM
Recommend

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS
Log In | Register Now

The New York Times
Sunday, August 16, 2009
Science

Search All NYTimes.com
Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

CRITICS SAY COFFEE STUDY WAS FLAWED

By HAROLD M. SCHMECK JR.
Published: June 30, 1981

SIGN IN TO RECOMMEND
E-MAIL
SEND TO PHONE
PRINT
REPRINTS
SHARE

THERE were flaws in a study showing links between coffee drinking and a common form of cancer, several medical scientists and physicians said in letters published in the latest issue of The New England Journal of Medicine.

In March, the journal carried a report showing statistical links between coffee drinking and cancer of the pancreas, the fourth most common cause of cancer deaths among Americans.

"This otherwise excellent paper may be flawed in one critical way," said a letter from Dr. Steven Shedlofsky of the Veterans Administration Hospital in White River Junction, Vt. He questioned the comparison of pancreatic cancer patients with persons hospitalized for noncancerous diseases of the digestive system.

Such patients, he noted, might be expected to give up coffee drinking because of their illness. This, he argued, would tilt the proportion of coffee drinkers away from the "control" group who were being compared with the cancer patients. Amplifying the letter in an interview, Dr. Shedlofsky said many patients with digestive diseases give up coffee because they believe it aggravates their discomfort, and others do so because their doctors have advised them to.

Dr. Thomas C. Chalmers, president of the Mount Sinai Medical Center and dean of its medical school, commented that the investigators who questioned patients on their prehospitalization coffee habits knew in advance which ones had cancer. This could have introduced unintentional bias in the results, Dr. Chalmers asserted.

Among the comments from other physicians were these: the question of whether noncancerous illness might have kept the control patients from drinking coffee was raised; a correspondent pointed out the problem inherent in trying to judge coffee consumption simply by asking about typical daily consumption before hospitalization; and another noted the possible role of other health habits that are closely related to coffee drinking. These habits included cigarette smoking and the use of sugar, milk, cream or nondairy "creamers" with the coffee.

The authors of the original report, led by Dr. Brian MacMahon of the Harvard School of Public Health, defended their study against all of the comments. They agreed that concern was "reasonable" over the large number of patients in their control group who had gastrointestinal disorders. But they said the association between coffee drinking and cancer of the pancreas was present in all the control groups.

The introduction of unintentional bias was unlikely, they said, because the study team had no hypotheses about coffee when it began the study. Coffee drinking only emerged as statistically important when most of the data had already been gathered, they said.
Differences Between Sexes

MOST POPULAR
E-MAILED | BLOGGED

1. Patient Money: The Expense of Eating With Celiac Disease
2. Noticed: It's Hip to Be Round
3. Bob Herbert: Hard to Believe!
4. Gail Collins: To Be Old and in Woodstock
5. Getting Your Wireless Network Up to Speed
6. Multicultural Stages in a Small Oregon Town
7. Well: Fatty Foods Affect Memory and Exercise
8. Believers Invest in the Gospel of Getting Rich
9. Shortcuts: New Worries About Children With Cellphones
10. Paul Krugman: Republican Death Trip

Go to Complete List »

» Sleep Study

» Coffee Allergy

The study showed no difference in risk between men who said they drank only about two cups of coffee a day and those who drank much more. Among women, however, the risk seemed to be related to the amount consumed. Some of the physicians who commented on the study considered the lack of a dose effect in men puzzling and a cause of doubt concerning the overall implications of the study.

In their original report, Dr. MacMahon and his colleagues treated their evidence cautiously, asserting that further studies were needed to determine whether coffee drinking was actually a factor in causing the cancers. If it is a matter of cause and effect, they said, and if the findings apply to the nation as a whole, coffee drinking might be a factor in slightly more than half of the roughly 20,000 cases a year of that form of cancer in the United States.

Coffee industry spokesmen, who were critical of the report when it was published in March, estimate that more than half of Americans over the age of 10 drink coffee.

Sign in to Recommend

More Articles in Science >

INSIDE NYTIMES.COM

HEALTH »

Roving Runner: Baseball Nostalgia in the Bronx

WORLD »

Kiev Residents Wonder if Mayor Is Fit for Office

OPINION »

But They Were Next in Line for Takeoff

Airplane passengers should demand approval of the merciful Airline Passengers Bill of Rights.

FASHION & STYLE »

SMASH IMPERIALISM ABOLISH ROTC NO EXPANSION
NO CONCESSIONS TO THE CORPORATION FIGHT TO WIN!

The Spirit of '69, Circa 1972

ARTS »

In Dresden, High Culture and Ugly Reality Clash

OPINION »

Weekend Opinionator: Cheney v. Bush

Home | World | U.S. | N.Y. / Region | Business | Technology | Science | Health | Sports | Opinion | Arts | Style | Travel | Jobs | Real Estate | Autos | Back to Top

EPIDEMIOLOGY Fourth Edition

LEON GORDIS, MD, MPH, DrPH
Professor of Epidemiology
Johns Hopkins Bloomberg School of Public Health
Professor of Pediatrics
Johns Hopkins School of Medicine
Baltimore, Maryland

SAUNDERS
ELSEVIER

although it is attractive to choose as hospitalized controls a disease group that is obviously unrelated to the putative causative factor under investigation, such controls are unlikely to be representative of the general reference population. As a result, it will not be clear whether it is the cases or the controls who differ from the general population.

The issue of which diagnostic groups would be eligible for use as controls and which would be ineligible (and therefore excluded) is very important. Let us say we are conducting a case-control study of lung cancer and smoking: we select as cases patients who have been hospitalized with lung cancer, and as controls we select patients who have been hospitalized with emphysema. What problem would this present? Because we know that there is a strong relationship between smoking and emphysema, our controls, the emphysema patients, would include a high number of smokers. Consequently, any relationship of smoking to lung cancer would not be detectable in this study, because we would have selected as controls a group of persons in which there is a greater-than-expected prevalence of smoking. We might therefore want to exclude from our control group those persons who have other smoking-related diagnoses, such as coronary heart disease, bladder cancer, pancreatic cancer, and emphysema. Such exclusions might yield a control group with a lower-than-expected prevalence of smoking and the exclusion process becomes complex. One alternative is to not exclude any groups from selection as controls in the design of the study, but to analyze the study data separately for different diagnostic subgroups that constitute the control group.

PROBLEMS IN CONTROL SELECTION

The following example demonstrates the problem of exclusions in the process of control selection:

In 1981, MacMahon and coworkers[6] reported a case-control study of cancer of the pancreas. The cases were patients with a histologically confirmed diagnosis of pancreatic cancer in 11 Boston and Rhode Island hospitals from 1974 to 1979. Controls were selected from all patients who were hospitalized at the same time as the cases; and they were selected from other inpatients hospitalized by the attending physicians who had hospitalized the cases. One finding in this study was an apparent dose–response relationship between coffee consumption and cancer of the pancreas, particularly in women (Table 10-6).

When such a relationship is observed, it is difficult to know whether the disease is *caused* by the coffee consumption or by some factor closely related to the coffee consumption. Because smoking is a known risk factor for cancer of the pancreas, and because coffee consumption is closely related to cigarette smoking (it is rare to find a smoker who does not drink coffee), did MacMahon and others observe an association of coffee consumption with pancreatic cancer because the coffee caused the pancreatic cancer, or because coffee consumption is related to cigarette smoking, and cigarette smoking is known to be a risk factor for cancer of the pancreas? Recognizing this problem, the authors analyzed the data after stratifying for smoking history. The relationship with coffee consumption held both for current smokers and for those who had never smoked (Table 10-7).

This report aroused great interest in both the scientific and lay communities, particularly among coffee manufacturers. Given the widespread exposure of human beings to coffee, if the reported relationship were true, it would have major public health implications.

Let us examine the design of this study. The cases were white patients with cancer of the pancreas at 11 Boston and Rhode Island hospitals. The controls are of particular interest: They were patients with other diseases who were hospitalized by the same physicians who had hospitalized the cases. That is, when a case had been identified, the attending physician was asked if another of his or her patients who was hospitalized at the same time for another condition could be interviewed as a control. This unusual method of control selection had a practical advantage: One of the major obstacles in obtaining participation of hospital controls in case-control studies is that permission to contact the patient is requested of the attending physician. The physicians are often not motivated to have their patients serve as controls, because the patients do not have the disease that is the focus of the study. By asking physicians who had already given permission for patients with pancreatic cancer to participate, the likelihood was increased that permission would be granted for patients with other diseases to participate as controls.

Did that practical decision introduce any problems? The underlying question that the investigators wanted to answer was whether patients with cancer of the pancreas drank more coffee than did people without cancer of the pancreas in the same population (Fig. 10-3). What MacMahon and coworkers

TABLE 10-6. **Distribution of Cases and Controls by Coffee-Drinking Habits and Estimates of Risk Ratios**

		Coffee Consumption (Cups/Day)				
Sex	**Category**	***0***	***1–2***	***3–4***	***≥5***	**Total**
M	Number of cases	9	94	53	60	216
	Number of controls	32	119	74	82	307
	Adjusted relative risk*	1.0	2.6	2.3	2.6	2.6
	95% Confidence interval	–	1.2–5.5	1.0–5.3	1.2–5.8	1.2–5.4
F	Number of cases	11	59	53	28	151
	Number of controls	56	152	80	48	336
	Adjusted relative risk*	1.0	1.6	3.3	3.1	2.3
	95% Confidence interval	–	0.8–3.4	1.6–7.0	1.4–7.0	1.2–4.6

*Chi-square (Mantel extension) with equally spaced scores, adjusted over age in decades: 1.5 for men, 13.7 for women. Mantel-Haenszel estimates of risk ratios, adjusted over categories of age in decades. In all comparisons, the referent category was subjects who never drank coffee.

From MacMahon B, Yen S, Trichopoulos D, et al: Coffee and cancer of the pancreas. N Engl J Med 304(11):630–633, 1981.

found was that the level of coffee consumption in cases was greater than the level of coffee consumption in controls.

The investigators would like to be able to establish that the level of coffee consumption observed in the controls is what would be expected in the general population without pancreatic cancer and that cases therefore demonstrate *excessive* coffee consumption (Fig. 10-4). But the problem is this: Which physicians are most likely to admit patients with cancer of the pancreas to the hospital? Gastroenterologists are often the admitting physicians. Many of their other hospitalized patients (who served as controls) also have gastrointestinal problems, such as esophagitis and peptic ulcer. So in this study, the persons who served as controls may very well have reduced their intake of coffee, either because of a physician's instructions or because of their own realization that reducing their coffee intake could relieve their symptoms. We cannot assume that the controls' levels of coffee consumption are representative of the level of coffee consumption expected in the general popula-

TABLE 10-7. **Estimates of Relative Risk* of Cancer of the Pancreas Associated with Use of Coffee and Cigarettes**

	Coffee Drinking (Cups/Day)			
Cigarette Smoking Status	***0***	***1–2***	***≥3***	**Total†**
Never smoked	1.0	2.1	3.1	1.0
Ex-smokers	1.3	4.0	3.0	1.3
Current smokers	1.2	2.2	4.6	1.2 (0.9–1.8)
Total*	1.0	1.8 (1.0–3.0)	2.7 (1.6–4.7)	

*The referent category is the group that uses neither cigarettes nor coffee. Estimates are adjusted for sex and age in decades.

†Values are adjusted for the other variables, in addition to age and sex, and are expressed in relation to the lowest category of each variable.

Values in parentheses are 95% confidence intervals of the adjusted estimates.

From MacMahon B, Yen S, Trichopoulos D, et al: Coffee and cancer of the pancreas. N Engl J Med 304(11):630–633, 1981.

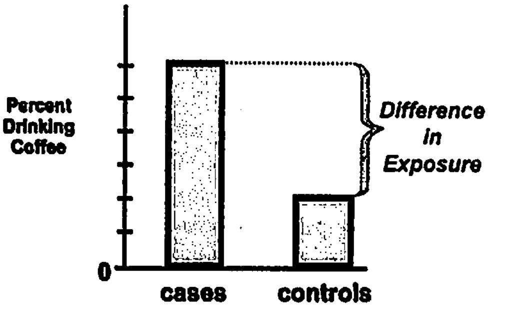


Figure 10-3. Interpreting the results of a case-control study of coffee drinking and pancreatic cancer.

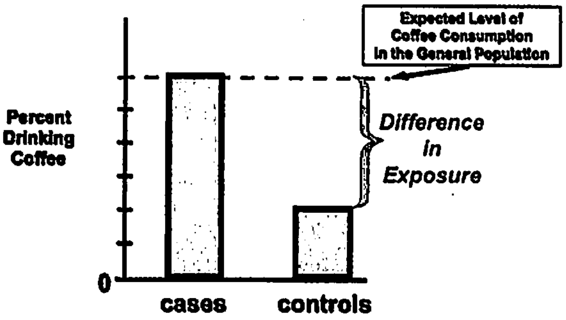


Figure 10-5. Interpreting the results of case-control studies: Is the higher level the expected level of exposure?

tion; their rate of coffee consumption may be abnormally low. Thus, the observed difference in coffee consumption between pancreatic cancer cases and controls may not necessarily have been the result of cases drinking more coffee than expected, but rather of the controls drinking less coffee than expected (Fig. 10-5).

MacMahon and his colleagues subsequently repeated their analysis but separated controls with gastrointestinal illness from controls with other conditions. They found that the risk associated with coffee consumption was indeed higher when the comparison was with controls with gastrointestinal illness but that the relationship between coffee consumption and pancreatic cancer persisted, albeit at a lower level, even when the comparison was with controls with other illnesses. Several years later, Hsieh and coworkers reported a new study that attempted to replicate these results; it did not support the original findings.[7]

In summary, when a difference in exposure is observed between cases and controls, we must ask whether the level of exposure observed in the controls is really the level expected in the population in which the study was carried out or whether—perhaps given the manner of selection—the controls may have a particularly high or low level of exposure that might not be representative of the level in the population in which the study was carried out.

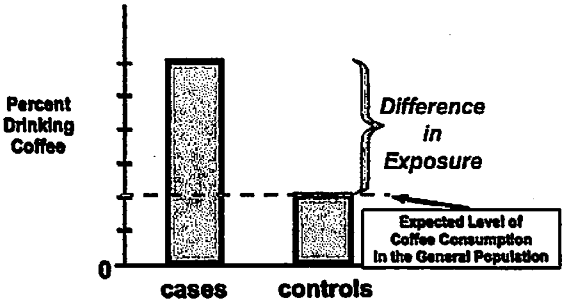


Figure 10-4. Interpreting the results of case-control studies: Is the lower level the expected level of exposure?

MATCHING

A major concern in conducting a case-control study is that cases and controls may differ in characteristics or exposures other than the one that has been targeted for study. If more cases than controls are found to have been exposed, we may be left with the question of whether the observed association could be due to differences between the cases and controls in factors other than the exposure being studied. For example, if more cases than controls are found to have been exposed, and if most of the cases are poor and most of the controls are affluent, we would not know whether the factor determining development of disease is exposure to the factor being studied or another characteristic associated with being poor. To avoid such a situation, we would like to ensure that the distribution of the cases and controls by socioeconomic status is similar, so that a difference in exposure will likely constitute the critical difference, and the presence or absence of disease is not likely to be attributable to a difference in socioeconomic status.

One approach to dealing with this problem in the design and conduct of the study is to match the cases and controls for factors about which we may be con-

Commentary

JAMA
THE JOURNAL of the
American Medical Association
Aug 28, 1981 Vol 246, No. 9

Coffee and Pancreatic Cancer

The Problems of Etiologic Science and Epidemiologic Case-Control Research

THE RECENT report that coffee may cause pancreatic cancer[1] was presented in a pattern that has become distressingly familiar. The alleged carcinogen is a commonly used product. The report was given widespread publicity before the supporting evidence was available for appraisal by the scientific community, and the public received renewed fear and uncertainty about the cancerous hazards lurking in everyday life.

The research on coffee and pancreatic cancer was done with the case-control technique that has regularly been used in epidemiologic circumstances where the more scientifically desirable forms[2] of clinical investigation—a randomized controlled trial or a suitably performed observational cohort study—are either impossible or unfeasible. In case-control studies, the investigators begin at the end, rather than at the beginning, of the cause-effect pathway. The cases are selected from persons in whom the target disease has already developed. The controls are selected from persons in whom that disease has not been noted. The cases and controls are then investigated in a backward temporal direction, with inquiries intended to determine antecedent exposure to agents that may have caused the disease. If the ratio of antecedent exposure to a particular agent is higher in the cases than in the controls, and if the associated mathematical calculations are "statistically significant," the agent is suspected of having caused the disease.

In the recently reported study[1] of coffee and pancreatic cancer, the investigators began by assembling records for 578 cases of patients with "histologic diagnoses of cancer of the exocrine pancreas." The investigators next created two "control" groups, having other diagnoses. The cases and controls were then interviewed regarding antecedent exposure to tobacco, alcohol, tea, and coffee. When the data were analyzed for groups demarcated according to gender and quantity of coffee consumption, the calculated relative-risk ratios for pancreatic cancer were the values shown in Table 1.

From these and other features of the statistical analysis, the investigators concluded that "a strong association between coffee consumption and pancreatic cancer was evident in both sexes." The conclusions were presented with the customary caveats about the need for more research and with the customary restraints shown in such expressions as "coffee use *might* [our italics] account for a substantial proportion" of pancreatic cancers. Nevertheless, the impression was strongly conveyed that coffee had been indicted as a carcinogen.

Although the major public attention has been given to the "Results" and "Discussion" sections of the published report, readers concerned with scientific standards of evidence will want to focus on the "Methods." The rest of this commentary contains a review of pertinent principles of case-control methodology, together with a critique of the way these principles were applied in the coffee-pancreatic cancer study to formulate a hypothesis, assemble the case and control groups, collect the individual data, and interpret the results.

Scientific Hypotheses and 'Fishing Expeditions'

Most case-control studies are done to check the hypothesis that the target disease has been caused by a specified suspected agent, but after the cases and controls are assembled the investigators can also collect data about many other possible etiologic agents. The process of getting and analyzing data for these other agents is sometimes called a "fishing expedition," but the process seems entirely reasonable. If we do not know what causes a disease, we might as well check many different possibilities. On the other hand, when an unsuspected agent yields a positive result, so that the causal hypothesis is generated by the data rather than by the investigator, the results of the fishing expedition require cautious interpretation. Many scientists would not even call the positive association a "hypothesis" until the work has been reproduced in another investigation.

The investigators who found a positive association between coffee consumption and pancreatic cancer have been commendably forthright in acknowledging that they were looking for something else. When the original analyses showed nothing substantial to incriminate the two principal suspects—tobacco and alcohol—the exploration of alternative agents began. The investigators do not state how many additional agents were examined besides

Table 1.—Relative-Risk Ratios According to Gender and Quantity of Coffee Consumption

	Coffee Consumption, Cups per Day			
	0	1-2	3-4	≥5
Men	1.0	2.6	2.3	2.6
Women	1.0	1.6	3.3	3.1

From the Robert Wood Johnson Clinical Scholars Program, Yale University School of Medicine, New Haven, Conn (Drs Feinstein and Horwitz), and the Cooperative Studies Program Support Center, Veterans Administration Hospital, West Haven, Conn (Dr Feinstein), and the McGill Cancer Center, McGill University (Dr Spitzer), and the Kellogg Center for Advanced Studies in Primary Care, Montreal General Hospital (Drs Spitzer and Battista), Montreal.

Reprint requests to Robert Wood Johnson Clinical Scholar Program, Yale University School of Medicine, 333 Cedar St, Box 3333, New Haven, CT 06510 (Dr Feinstein).

tea and coffee, but tea was exonerated in the subsequent analyses, while coffee yielded a positive result.

The investigators suggest that this result is consistent with coffee-as-carcinogen evidence that had appeared in a previous case-control study[3] of pancreatic cancer. In fact, however, coffee was not indicted in that previous study. The previous investigators found an elevated risk ratio for only decaffeinated coffee, and they drew no conclusion about it, having found elevated risks for several other phenomena that led to the decision that pancreatic cancer had a nonspecific multifactorial etiology. Thus, the new hypothesis that coffee may cause pancreatic cancer not only arises from a "fishing expedition," but also contradicts the results found in previous research.

Selection and Retention of Cases and Controls

Because the investigators begin at the end of the causal pathway and must explore it with a reversal of customary scientific logic, the selection of cases and controls is a crucial feature of case-control studies. Both groups are chosen according to judgments made by the investigators. The decisions about the cases are relatively easy. They are commonly picked from a registry or some other listing that will provide the names of persons with the target disease. For the controls, who do not have the target disease, no standard method of selection is available, and they have come from an extraordinarily diverse array of sources. The sources include death certificates, tumor registries, hospitalized patients, patients with specific categories of disease, patients hospitalized on specific clinical services, other patients of the same physicians, random samples of geographically defined communities, people living in "retirement" communities, neighbors of the cases, or personal friends of the cases.

One useful way of making these decisions less arbitrary is to choose cases and controls according to the same principles of eligibility and observation that would be used in a randomized controlled trial of the effects of the alleged etiologic agent. In such a trial, a set of admission criteria would be used for demarcating persons to be included (or excluded) in the group who are randomly assigned to be exposed or non-exposed to the agent. Special methods would then be used to follow the members of the exposed and non-exposed groups thereafter, and to examine them for occurrence of the target disease. Those in whom this disease develops would become the cases, and all other people would be the controls.

When cases and controls are chosen for a case-control study, the selection can be made from persons who would have been accepted for admission to such a randomized trial and who have been examined with reasonably similar methods of observation. As a scientific set of guidelines for choosing eligible patients, the randomized-trial principles could also help avoid or reduce many of the different forms of bias that beset case-control studies. Among these difficulties are several biases to be discussed later, as well as other problems such as clinical susceptibility bias, surveillance bias, detection bias, and "early death" bias, which are beyond the scope of this discussion and have been described elsewhere.[4-9]

The randomized-trial principles can also help illuminate the problems created and encountered by the investigators in the study of coffee and pancreatic cancer. In a randomized trial, people without pancreatic cancer would be assigned either to drink or not to drink coffee. Anyone with clinical contraindications against coffee drinking or indications for it (whatever they might be) would be regarded as ineligible and not admitted. Everyone who did enter the trial, however, would thereafter be included in the results as the equivalent of either a case, if later found to have pancreatic cancer, or a control. The cases would be "incidence cases," with newly detected pancreatic cancer, whose diagnoses would be verified by a separate panel of histological reviewers. All of the other admitted persons would eventually be classified as unaffected "controls," no matter what ailments they acquired, as long as they did not have pancreatic cancer. If large proportions of the potential cases and controls were lost to follow-up, the investigators would perform detailed analyses to show that the remaining patients resembled those who were lost, thus providing reasonable assurance that the results were free from migration bias.[2]

In the coffee-pancreatic cancer study, the source of the cases was a list of 578 patients with "histologic diagnoses of cancer of the exocrine pancreas." The histologic material was apparently not obtained and reviewed; and the authors do not indicate whether the patients were newly diagnosed "incidence cases," or "prevalence cases" who had been diagnosed at previous admissions. Regardless of the incidence-prevalence distinction, however, the published data are based on only 369 (64%) of the 578 patients who were identified as potential cases. Most of the "lost" patients were not interviewed, with 98 potential cases being too sick or already dead when the interviewer arrived. The investigators report no data to indicate whether the "lost" cases were otherwise similar to those who were retained.

In choosing the control group, the investigators made several arbitrary decisions about whom to admit or exclude. The source of the controls was "all other patients who were under the care of the same physician in the same hospital at the time of an interview with a patient with pancreatic cancer." From this group, the investigators then excluded anyone with any of the following diagnoses: diseases of the pancreas; diseases of the hepatobiliary tract; cardiovascular disease; diabetes mellitus; respiratory cancer; bladder cancer; or peptic ulcer. Since none of these patients would have been excluded as nonpancreatic-cancer controls if they acquired these diseases after entry into a randomized trial of coffee drinking, their rejection in this case-control study is puzzling. The investigators give no reasons for excluding patients with "diseases of the pancreas or hepatobiliary tract." The reason offered for the other rejections is that the patients had "diseases known to be associated with smoking or alcohol consumption." The pertinence of this stipulation for a study of coffee is not readily apparent.

Since the investigators do not state how many potential controls were eliminated, the proportionate impact of the exclusions cannot be estimated. The remaining list of eligible control patients, however, contained 1,118 people, of whom only a little more than half—644 patients—became the actual control group used for analyses. Most of the "lost" controls were not interviewed because of death, early discharge, severity of illness, refusal to participate, and language problems. Of the 700 interviewed controls, 56 were subsequently excluded because they were nonwhite, foreign, older than 79 years, or "unreliable." No data are offered to demonstrate that the 644 actual controls were similar to the 474 "eligible" controls who were not included.

The many missing controls and missing interviews could have led to exclusion biases[10,11] whose effects cannot be evaluated in this study. The investigators have also given no attention to the impact of selective hospitalization bias, perceived by Berkson[4] and empirically demonstrated by Roberts et al,[6] that can sometimes falsely elevate relative-risk ratios in a hospital population to as high as 17 times their true value in the general population. For example, in a hospitalized population, Roberts et al[6] found a value of 5.0 for the relative-risk ratio of arthritic and rheumatic complaints in relation to laxative usage; but in the general population that contained the hospitalized patients, the true value was 1.5. Whatever may have been the effects of selective hospitalization in the current study (including the possibility of having masked real effects of tobacco and alcohol), the way that the cases and controls were chosen made the study particularly vulnerable to the type of bias described in the next section.

Protopathic Bias in Cases and Controls

"Protopathic" refers to early disease. A protopathic problem occurs if a person's exposure to a suspected etiologic agent is altered because of the early manifestations of a disease, and if the altered (rather than the original) level of exposure is later associated with that disease. By producing changes in a person's life-style or medication, the early manifestations of a disease can create a bias unique to case-control studies.[12] In a randomized trial or observational cohort study, the investigator begins with each person's baseline state and follows it to the subsequent outcome. If exposure to a suspected etiologic agent is started, stopped, or altered during this pathway, the investigator can readily determine whether the change in exposure took place before or after occurrence of the outcome. In a case-control study, however, the investigator beginning with an outcome cannot be sure whether it preceded or followed changes in exposure to the suspected agent. If the exposure was altered because the outcome had already occurred and if the timing of this change is not recognized by the investigator, the later level of exposure (or non-exposure) may be erroneously linked to the outcome event.

For example, in circumstances of ordinary medical care, women found to have benign breast disease might be told by their physicians to avoid or stop any form of estrogen therapy. If such women are later included as cases in a case-control study of etiologic factors in benign breast disease, the antecedent exposure to estrogens will have been artifactually reduced in the case group. Oral contraceptives or other forms of estrogen therapy may then be found to exert a fallacious "protection" against the development of benign breast disease.

The problem of protopathic bias will occur in a case-control study if the amount of previous exposure to the suspected etiologic agent was preferentially altered—either upward or downward—because of clinical manifestations that represented early effects of the same disease that later led to the patient's selection as either a case or control. The bias is particularly likely to arise if the preferential decisions about exposure were made in opposite directions in the cases and controls. The coffee-pancreatic cancer study was particularly susceptible to this type of bi-directional bias. The customary intake of coffee may have been increased by members of the pancreatic-cancer case group who were anxious about vague abdominal symptoms that had not yet become diagnosed or even regarded as "illness." Conversely, control patients with such gastrointestinal ailments as regional enteritis or dyspepsia may have been medically advised to stop or reduce their drinking of coffee. With a strict set of admission criteria, none of these patients would be chosen as cases or controls, because the use of the alleged etiologic agent would have been previously altered by the same ailment that led to the patient's selection for the study.

This problem of protopathic bias is a compelling concern in the investigation under review here. Because so many potential control patients were excluded, the remaining control group contained many people with gastrointestinal diseases for which coffee drinking may have been previously reduced or eliminated. Of the 644 controls, 249 (39%) had one of the following diagnoses: cancer of the stomach, bowel, or rectum; colitis, enteritis, or diverticulitis; bowel obstruction, adhesions, or fistula; gastritis; or "other gastroenterologic conditions." If coffee drinking is really unrelated to pancreatic cancer, but if many of these 249 patients had premonitory symptoms that led to a cessation or reduction in coffee drinking "before the current illness was evident," the subsequent distortions could easily produce a false-positive association.

The existence of this type of bias could have been revealed or prevented if the investigators had obtained suitable data. All that was needed during the interview with each case or control patient was to ask about duration of coffee drinking, changes in customary pattern of consumption, and reasons for any changes. Unfortunately, since coffee was not a major etiologic suspect in the research, this additional information was not solicited. After the available data were analyzed, when the investigators became aware of a possible problem, they tried to minimize its potential importance by asserting that "although the majority of control patients in our series had chronic disease, pancreatic cancer is itself a chronic disease, and in theory it would seem as likely as any other disorder to induce a change in coffee [consumption]." This assertion does not address the point at issue. The bias under discussion arises from changes in exposure status because of the early clinical manifestations of a disease, not from the chronic (or acute) characteristics of the conditions under comparison.

The investigators also claimed that "it is inconceivable that this bias would account for the total difference between cases and controls." The conception is actually quite easy. To make the demonstration clear, let us eliminate gender distinctions and coffee quantification in the investigators' Table 4, which can then be converted into a simple fourfold table (Table 2). In this table, the odds ratio, which estimates the relative-risk ratio, is (347/20)/(555/88)=2.75, which is the same magnitude as the relative risks cited by the investigators.

Let us now assume that 5% of the coffee-drinker cases were formerly non-coffee-drinkers. If so, 17 people in the case group would be transferred downward from the coffee drinkers to the nondrinkers. Although 249 members

Table 2.—Status of Study Subjects According to Coffee Consumption

	Cases	Controls
Coffee-drinkers	347	555
Non-coffee-drinkers	20	88
Total	**367**	**643**

Table 3.—Hypothetical* Status of Study Subjects Shown in Table 2

	Cases	Controls
Coffee-drinkers	330	573
Non-coffee-drinkers	37	70
Total	**367**	**643**

*Based on estimate that 5% of coffee-drinkers in case group were previously non-coffee-drinkers and that 20% of non-coffee-drinkers in control group ceased coffee consumption because of symptoms.

of the control group had gastrointestinal conditions that might have led to a cessation of coffee consumption, let us conservatively estimate that only 20% of the 88 controls listed in the non-coffee-drinkers category were previous coffee-drinkers who had stopped because of symptoms. If so, 18 of the non-coffee-drinking controls should move upward into the coffee-drinking group. With these reclassifications, the adjusted fourfold table would be as presented in Table 3. For this new table, the odds ratio is (330/37)/(573/70)=1.09, and the entire positive association vanishes.

Acquisition of Basic Data

All of the difficulties just described arise as consequences of basic decisions made in choosing cases and controls. After these decisions are completed, the case-control investigator acquires information about each person's antecedent exposure. This information becomes the basic research data, analogous to the description of each patient's outcome in a randomized controlled trial. The information about exposure should therefore be collected with thorough scientific care, using impeccable criteria to achieve accuracy, and, when necessary, using objective (or "blinded") methods to prevent biased observations.

These scientific requirements are seldom fulfilled in epidemiologic research. The primary data about exposure are verified so infrequently in case-control studies that prominent epidemiologists[13] have begun to make public pleas for improved scientific standards and methods. In the few instances where efforts have been made to confirm recorded data,[14,15] to repeat interviews at a later date[16] or to check the agreement of data obtained from different sources,[11] the investigators have encountered discrepancies of major magnitude. In one of these studies,[17] when the agent of exposure (occupation as a fisherman) was confirmed, the original numbers of exposed people were reduced by 17%. Had these numbers not been corrected, the study would have produced misleading conclusions.

Although errors of similar magnitude could easily have occurred in the coffee-pancreatic cancer investigation, the investigators did not publish even a brief text of the actual questions used for the interviews, and no efforts are mentioned to check the quality of the data that were obtained in the single interview with each patient. Family members or friends were not asked to confirm the patients' answers; the information was not checked against previous records; and no patients were reinterviewed after the original interrogation to see whether subsequent responses agreed with what was said previously. Although a verification of each interview is difficult to achieve in a large study, the scientific quality of the data could have been checked in a selected sample.

Because of the high likelihood of the protopathic bias noted earlier, the quality of the coffee-drinking data is a major problem in the study under review. The investigators state that "the questions on tea and coffee were limited to the number of cups consumed in a typical day before the current illness was evident." This approach would not produce reliable data, since it does not indicate what and when was a "typical day," who decided what was the "time before the current illness was evident," or who determined which of the patient's symptoms were the first manifestation of "illness" either for pancreatic cancer or for the diverse diseases contained in the control group.

Although the investigators acknowledge the possibility that "patients reduced their coffee consumption because of illness," nothing was done to check this possibility or to check the alternative possibility that other patients may have increased their customary amounts of coffee drinking. In addition to no questions about changes in coffee consumption, the patients were also asked nothing about duration. Thus, a patient who had started drinking four cups a day in the past year would have been classified as having exactly the same exposure as a patient who had been drinking four cups a day for 30 years.

The Problem of Multiple Contrasts

When multiple features of two groups are tested for "statistically significant" differences, one or more of those features may seem "significant" purely by chance. This multiple-contrast problem is particularly likely to arise during a "fishing expedition." In the customary test of statistical significance, the investigator contrasts the results for a single feature in two groups. The result of this single-feature two-group contrast is declared significant if the P value falls below a selected boundary, which is called the α level. Because α is commonly set at .05, medical literature has become replete with statements that say "the results are statistically significant at $P<.05$." For a single two-group contrast at an α level of .05, the investigator has one chance in 20 (which can also be expressed as contrary odds of 19 to 1) of finding a false-positive result if the contrasted groups are really similar.

For the large series of features that receive two-group contrasts during a "fishing expedition," however, statistical significance cannot be decided according to the same α level used for a single contrast. For example, in the coffee-pancreatic cancer study, the cases and controls were divided for two-group contrasts of such individual exposures (or non-exposures) as cigars, pipes, cigarettes, alcohol, tea, and coffee. (If other agents were also checked, the results are not mentioned.) With at least six such two-group contrasts, the random chance of finding a single false-positive association where none really exists is no longer .05. If the characteristics are mutually independent, the chance is at least $.26[=1-(.95)^6]$. Consequently, when six different agents are checked in the same study, the odds against finding a spurious positive result are reduced from 19 to 1 and become less than 3 to 1 [=.74/.26].

To guard against such spurious conclusions during multiple contrasts, the customary statistical strategy is to make stringent demands on the size of the P value required for "significance." Instead of being set at the customary value of .05, the α level is substantially lowered. Statisticians do not agree on the most desirable formula for determining this lowered boundary, but a frequent procedure is to divide the customary α level by k, where k is the number of comparisons.[18] Thus, in the current study, containing at least six comparisons, the decisive level of α would be set at no higher than .05/6=.008.

In the published report, the investigators make no comment about this multiple-contrast problem and they do not seem to have considered it in their analyses. In one of the results, a P value is cited as "<.001," but most of the cogent data for relative risks are expressed in "95% confidence intervals," which were calculated with α=.05. Many of those intervals would become expanded to include the value of 1, thereby losing "statistical significance," if α were re-set at the appropriate level of .008 or lower.

Comment

The foregoing discussion has been confined to the main reasons for doubting the reported association between coffee and pancreatic cancer. Readers who are interested in evaluating other features of the study can check its constituent methods by referring to the criteria listed in several published proposals[8-10] of scientific standards for case-control research.

A separate problem, to be mentioned only in passing, is the appropriateness of forming conclusions and extensively diffusing results from a study in which the hypothesis develops as an analytic surprise in the data. Scientists and practitioners in the field of human health face difficult dilemmas about the risks and benefits of their activities. The old principle of avoiding harm whenever possible holds true whether a person or a population is at risk. Whether to shout "Fire!" in a crowded theater is a difficult decision, even if a fire is clearly evident. The risk of harm seems especially likely if such shouts are raised when the evidence of a blaze is inconclusive or meager. Aside from puzzled medical practitioners and a confused lay public, another possible victim is the developing science of chronic disease epidemiology. Its credibility can withstand only a limited number of false alarms.

Because the epidemiologic case-control study is a necessary, currently irreplaceable research mechanism in etiologic science, its procedures and operating paradigms need major improvements in scientific quality. In the evaluation of cause-effect relationships for therapeutic agents, the experimental scientific principles of a randomized trial have sometimes required huge sample sizes and massive efforts that have made the trials become an "indispensable ordeal."[19] In the evaluation of cause-effect relationships for etiologic agents, the case-control technique has eliminated the "ordeal" of a randomized controlled trial by allowing smaller sample sizes, the analysis of natural events and data, and a reversed observational direction. Since the use of scientific principles remains "indispensable," however, the development and application of suitable scientific standards in case-control research is a prime challenge in chronic disease epidemiology today.

The current methodologic difficulties arise because case-control investigators, having recognized that etiologic agents cannot be assigned with experimental designs, and having necessarily abandoned the randomization principle in order to work with naturally occurring events and data, have also abandoned many other scientific principles that are part of the experimental method and that could be employed in observational research. The verification and suitably unbiased acquisition of basic raw data regarding diagnoses and exposures do not require randomized trials; and the patients admitted to an observational study can be selected in accordance with the same eligibility criteria and the same subsequent diagnostic procedures that would have been used in a randomized trial.[20] These scientific experimental principles, however, are still frequently disregarded in case-control research, despite the celebrated warning of the distinguished British statistician, Sir Austin Bradford Hill.[21] In discussing the use of observational substitutes for experimental trials, he said that the investigator "must have the experimental approach firmly in mind" and must work "in such a way as to fulfill, as far as possible, experimental requirements."

ALVAN R. FEINSTEIN, MD
RALPH I. HORWITZ, MD
WALTER O. SPITZER, MD
RENALDO N. BATTISTA, MD

1. MacMahon B, Yen S, Trichopoulos D, et al: Coffee and cancer of the pancreas. *N Engl J Med* 1981;304:630-633.

2. Feinstein AR: Clinical biostatistics: XLVIII. Efficacy of different research structures in preventing bias in the analysis of causation. *Clin Pharmacol Ther* 1979;26:129-141.

3. Lin RS, Kessler II: A multifactorial model for pancreatic cancer in man. *JAMA* 1981;245:147-152.

4. Berkson J: Limitations of the application of four-fold tables to hospital data. *Biometrics Bull* 1946;2:47-53.

5. Neyman J: Statistics: Servant of all sciences. *Science* 1955;122:401.

6. Roberts RS, Spitzer WO, Delmore T, et al: An empirical demonstration of Berkson's bias. *J Chronic Dis* 1978;31:119-128.

7. Horwitz RI, Feinstein AR: Methodologic standards and contradictory results in case-control research. *Am J Med* 1979;66:556-564.

8. Feinstein AR: Methodologic problems and standards in case-control research. *J Chronic Dis* 1979;32:35-41.

9. Sackett DL: Bias in analytic research. *J Chronic Dis* 1979;32:51-63.

10. Horwitz RI, Feinstein AR, Stewart KR: Exclusion bias and the false relationship of reserpine/breast cancer, abstracted. *Clin Res* 1981;29:563.

11. Horwitz RI, Feinstein AR, Stremlau JR: Alternative data sources and discrepant results in case-control studies of estrogens and endometrial cancer. *Am J Epidemiol* 1980;111:389-394.

12. Horwitz RI, Feinstein AR: The problem of 'protopathic bias' in case-control studies. *Am J Med* 1980;68:255-258.

13. Gordis L: Assuring the quality of questionnaire data in epidemiologic research. *Am J Epidemiol* 1979;109:21-24.

14. Chambers LW, Spitzer WO, Hill GB, et al: Underreporting of cancer in medical surveys: A source of systematic error in cancer research. *Am J Epidemiol* 1976;104:141-145.

15. Chambers LW, Spitzer WO: A method of estimating risk for occupational factors using multiple data sources: The Newfoundland lip cancer study. *Am J Public Health* 1977;67:176-179.

16. Klemetti A, Saxen L: Prospective versus retrospective approach in the search for environmental causes of malformations. *Am J Public Health* 1967;57:2071-2075.

17. Spitzer WO, Hill GB, Chambers LW, et al: The occupation of fishing as a risk factor in cancer of the lip. *N Engl J Med* 1975;293:419-424.

18. Brown BW Jr, Hollander M: *Statistics: A Biomedical Introduction.* New York, John Wiley & Sons Inc, 1977, pp 231-234.

19. Fredrickson DS: The field trial: Some thoughts on the indispensable ordeal. *Bull NY Acad Med* 1968;44:985-993.

20. Horwitz RI, Feinstein AR: A new research method, suggesting that anticoagulants reduce mortality in patients with myocardial infarction. *Clin Pharmacol Ther* 1980;27:258.

21. Hill AB: Observation and experiment. *N Engl J Med* 1953;248:995-1001.

A Structural Approach to Selection Bias

Miguel A. Hernán, Sonia Hernández-Díaz,† and James M. Robins**

Abstract: The term "selection bias" encompasses various biases in epidemiology. We describe examples of selection bias in case-control studies (eg, inappropriate selection of controls) and cohort studies (eg, informative censoring). We argue that the causal structure underlying the bias in each example is essentially the same: conditioning on a common effect of 2 variables, one of which is either exposure or a cause of exposure and the other is either the outcome or a cause of the outcome. This structure is shared by other biases (eg, adjustment for variables affected by prior exposure). A structural classification of bias distinguishes between biases resulting from conditioning on common effects ("selection bias") and those resulting from the existence of common causes of exposure and outcome ("confounding"). This classification also leads to a unified approach to adjust for selection bias.

(*Epidemiology* 2004;15: 615–625)

Epidemiologists apply the term "selection bias" to many biases, including bias resulting from inappropriate selection of controls in case-control studies, bias resulting from differential loss-to-follow up, incidence–prevalence bias, volunteer bias, healthy-worker bias, and nonresponse bias.

As discussed in numerous textbooks,[1–5] the common consequence of selection bias is that the association between exposure and outcome among those selected for analysis differs from the association among those eligible. In this article, we consider whether all these seemingly heterogeneous types of selection bias share a common underlying causal structure that justifies classifying them together. We use causal diagrams to propose a common structure and show how this structure leads to a unified statistical approach to adjust for selection bias. We also show that causal diagrams can be used to differentiate selection bias from what epidemiologists generally consider confounding.

CAUSAL DIAGRAMS AND ASSOCIATION

Directed acyclic graphs (DAGs) are useful for depicting causal structure in epidemiologic settings.[6–12] In fact, the structure of bias resulting from selection was first described in the DAG literature by Pearl[13] and by Spirtes et al.[14] A DAG is composed of variables (nodes), both measured and unmeasured, and arrows (directed edges). A causal DAG is one in which 1) the arrows can be interpreted as direct causal effects (as defined in Appendix A.1), and 2) all common causes of any pair of variables are included on the graph. Causal DAGs are acyclic because a variable cannot cause itself, either directly or through other variables. The causal DAG in Figure 1 represents the dichotomous variables L (being a smoker), E (carrying matches in the pocket), and D (diagnosis of lung cancer). The lack of an arrow between E and D indicates that carrying matches does not have a causal effect (causative or preventive) on lung cancer, ie, the risk of D would be the same if one intervened to change the value of E.

Besides representing causal relations, causal DAGs also encode the causal determinants of statistical associations. In fact, the theory of causal DAGs specifies that an association between an exposure and an outcome can be produced by the following 3 causal structures[13,14]:

1. Cause and effect: If the exposure E causes the outcome D, or vice versa, then they will in general be associated. Figure 2 represents a randomized trial in which E (antiretroviral treatment) prevents D (AIDS) among HIV-infected subjects. The (associational) risk ratio ARR_{ED} differs from 1.0, and this association is entirely attributable to the causal effect of E on D.
2. Common causes: If the exposure and the outcome share a common cause, then they will in general be associated even if neither is a cause of the other. In Figure 1, the common cause L (smoking) results in E (carrying matches) and D (lung cancer) being associated, ie, again, $ARR_{ED} \neq 1.0$.
3. Common effects: An exposure E and an outcome D that have a common effect C will be conditionally associated if

Submitted 21 March 2003; final version accepted 24 May 2004.

From the *Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts; and the †Slone Epidemiology Center, Boston University School of Public Health, Brookline, Massachusetts.

Miguel Hernán was supported by NIH grant K08-AI-49392 and James Robins by NIH grant R01-AI-32475.

Correspondence: Miguel Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. E-mail: miguel_hernan@post.harvard.edu

ISSN: 1044-3983/04/1505-0615
DOI: 10.1097/01.ede.0000135174.63482.43

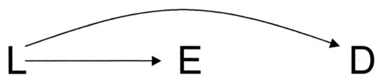


FIGURE 1. Common cause L of exposure E and outcome D.

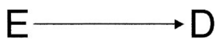


FIGURE 2. Causal effect of exposure E on outcome D.

the association measure is computed within levels of the common effect C, ie, the stratum-specific ARR$_{ED|C}$ will differ from 1.0, regardless of whether the crude (equivalently, marginal, or unconditional) ARR$_{ED}$ is 1.0. More generally, a conditional association between E and D will occur within strata of a common effect C of 2 other variables, one of which is either exposure or a cause of exposure and the other is either the outcome or a cause of the outcome. Note that E and D need not be unconditionally associated simply because they have a common effect. In the Appendix we describe additional, more complex, structural causes of statistical associations.

That causal structures (1) and (2) imply a crude association accords with the intuition of most epidemiologists. We now provide intuition for why structure (3) induces a conditional association. (For a formal justification, see references 13 and 14.) In Figure 3, the genetic haplotype E and smoking D both cause coronary heart disease C. Nonetheless, E and D are marginally unassociated (ARR$_{ED}$ = 1.0) because neither causes the other and they share no common cause. We now argue heuristically that, in general, they will be conditionally associated within levels of their common effect C.

Suppose that the investigators, who are interested in estimating the effect of haplotype E on smoking status D, restricted the study population to subjects with heart disease (C = 1). The square around C in Figure 3 indicates that they are conditioning on a particular value of C. Knowing that a subject with heart disease lacks haplotype E provides some information about her smoking status because, in the absence of E, it is more likely that another cause of C such as D is present. That is, among people with heart disease, the proportion of smokers is increased among those without the haplotype E. Therefore, E and D are inversely associated conditionally on C = 1, and the conditional risk ratio ARR$_{ED|C=1}$ is less than 1.0. In the extreme, if E and D were the only causes of C, then among people with heart disease, the absence of one of them would perfectly predict the presence of the other.

As another example, the DAG in Figure 4 adds to the DAG in Figure 3 a diuretic medication M whose use is a consequence of a diagnosis of heart disease. E and D are also associated within levels of M because M is a common effect of E and D.

There is another possible source of association between 2 variables that we have not discussed yet. As a result of sampling variability, 2 variables could be associated by chance even in the absence of structures (1), (2), or (3). Chance is not a structural source of association because chance associations become smaller with increased sample size. In contrast, structural associations remain unchanged. To focus our discussion on structural rather than chance associations, we assume we have recorded data in every subject in a very large (perhaps hypothetical) population of interest. We also assume that all variables are perfectly measured.

A CLASSIFICATION OF BIASES ACCORDING TO THEIR STRUCTURE

We will say that bias is present when the association between exposure and outcome is not in its entirety the result of the causal effect of exposure on outcome, or more precisely when the causal risk ratio (CRR$_{ED}$), defined in Appendix A.1, differs from the associational risk ratio (ARR$_{ED}$). In an ideal randomized trial (ie, no confounding, full adherence to treatment, perfect blinding, no losses to follow up) such as the one represented in Figure 2, there is no bias and the association measure equals the causal effect measure.

Because nonchance associations are generated by structures (1), (2), and (3), it follows that biases could be classified on the basis of these structures:

1. Cause and effect could create bias as a result of reverse causation. For example, in many case-control studies, the outcome precedes the exposure measurement. Thus, the association of the outcome with measured exposure could in part reflect bias attributable to the outcome's effect on measured exposure.[7,8] Examples of reverse causation bias include not only recall bias in case-control studies, but also more general forms of information bias like, for example, when a blood parameter affected by the presence of cancer is measured after the cancer is present.
2. Common causes: In general, when the exposure and outcome share a common cause, the association measure

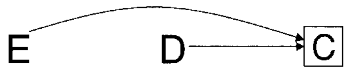


FIGURE 3. Conditioning on a common effect C of exposure E and outcome D.

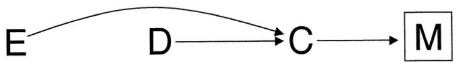


FIGURE 4. Conditioning on a common effect M of exposure E and outcome D.

differs from the effect measure. Epidemiologists tend to use the term *confounding* to refer to this bias.

3. Conditioning on common effects: We propose that this structure is the source of those biases that epidemiologists refer to as selection bias. We argue by way of example.

EXAMPLES OF SELECTION BIAS

Inappropriate Selection of Controls in a Case-Control Study

Figure 5 represents a case-control study of the effect of postmenopausal estrogens (E) on the risk of myocardial infarction (D). The variable C indicates whether a woman in the population cohort is selected for the case-control study (yes = 1, no = 0). The arrow from disease status D to selection C indicates that cases in the cohort are more likely to be selected than noncases, which is the defining feature of a case-control study. In this particular case-control study, investigators selected controls preferentially among women with a hip fracture (F), which is represented by an arrow from F to C. There is an arrow from E to F to represent the protective effect of estrogens on hip fracture. Note Figure 5 is essentially the same as Figure 3, except we have now elaborated the causal pathway from E to C.

In a case-control study, the associational exposure–disease odds ratio ($AOR_{ED|C\,=\,1}$) is by definition conditional on having been selected into the study (C = 1). If subjects with hip fracture F are oversampled as controls, then the probability of control selection depends on a consequence F of the exposure (as represented by the path from E to C through F) and "inappropriate control selection" bias will occur (eg, $AOR_{ED|C\,=\,1}$ will differ from 1.0, even when like in Figure 5 the exposure has no effect on the disease). This bias arises because we are conditioning on a common effect C of exposure and disease. A heuristic explanation of this bias follows. Among subjects selected for the study, controls are more likely than cases to have had a hip fracture. Therefore, because estrogens lower the incidence of hip fractures, a control is less likely to be on estrogens than a case, and hence $AOR_{ED|C\,=\,1}$ is greater than 1.0, even though the exposure does not cause the outcome. Identical reasoning would explain that the expected $AOR_{ED|C\,=\,1}$ would be greater than the causal OR_{ED} even had the causal OR_{ED} differed from 1.0.

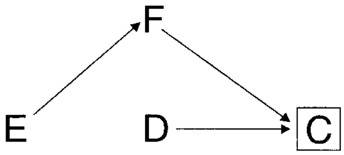


FIGURE 5. Selection bias in a case-control study. See text for details.

Berkson's Bias

Berkson[15] pointed out that 2 diseases (E and D) that are unassociated in the population could be associated among hospitalized patients when both diseases affect the probability of hospital admission. By taking C in Figure 3 to be the indicator variable for hospitalization, we recognize that Berkson's bias comes from conditioning on the common effect C of diseases E and D. As a consequence, in a case-control study in which the cases were hospitalized patients with disease D and controls were hospitalized patients with disease E, an exposure R that causes disease E would appear to be a risk factor for disease D (ie, Fig. 3 is modified by adding factor R and an arrow from R to E). That is, $AOR_{RD|C\,=\,1}$ would differ from 1.0 even if R does not cause D.

Differential Loss to Follow Up in Longitudinal Studies

Figure 6a represents a follow-up study of the effect of antiretroviral therapy (E) on AIDS (D) risk among HIV-

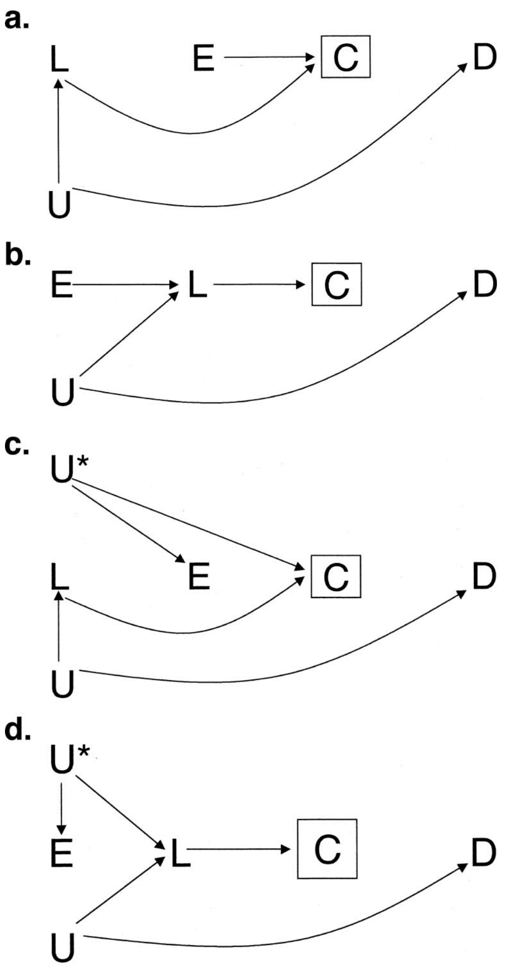


FIGURE 6. Selection bias in a cohort study. See text for details.

infected patients. The greater the true level of immunosuppression (U), the greater the risk of AIDS. U is unmeasured. If a patient drops out from the study, his AIDS status cannot be assessed and we say that he is censored (C = 1). Patients with greater values of U are more likely to be lost to follow up because the severity of their disease prevents them from attending future study visits. The effect of U on censoring is mediated by presence of symptoms (fever, weight loss, diarrhea, and so on), CD4 count, and viral load in plasma, all summarized in the (vector) variable L, which could or could not be measured. The role of L, when measured, in data analysis is discussed in the next section; in this section, we take L to be unmeasured. Patients receiving treatment are at a greater risk of experiencing side effects, which could lead them to dropout, as represented by the arrow from E to C. For simplicity, assume that treatment E does not cause D and so there is no arrow from E to D ($CRR_{ED} = 1.0$). The square around C indicates that the analysis is restricted to those patients who did not drop out (C = 0). The associational risk (or rate) ratio $ARR_{ED|C = 0}$ differs from 1.0. This "differential loss to follow-up" bias is an example of bias resulting from structure (3) because it arises from conditioning on the censoring variable C, which is a common effect of exposure E and a cause U of the outcome.

An intuitive explanation of the bias follows. If a treated subject with treatment-induced side effects (and thereby at a greater risk of dropping out) did in fact not drop out (C = 0), then it is generally less likely that a second cause of dropping out (eg, a large value of U) was present. Therefore, an inverse association between E and U would be expected. However, U is positively associated with the outcome D. Therefore, restricting the analysis to subjects who did not drop out of this study induces an inverse association (mediated by U) between exposure and outcome, ie, $ARR_{ED|C = 0}$ is not equal to 1.0.

Figure 6a is a simple transformation of Figure 3 that also represents bias resulting from structure (3): the association between D and C resulting from a direct effect of D on C in Figure 3 is now the result of U, a common cause of D and C. We now present 3 additional structures, (Figs. 6b–d), which could lead to selection bias by differential loss to follow up.

Figure 6b is a variation of Figure 6a. If prior treatment has a direct effect on symptoms, then restricting the study to the uncensored individuals again implies conditioning on the common effect C of the exposure and U thereby introducing a spurious association between treatment and outcome. Figures 6a and 6b could depict either an observational study or an experiment in which treatment E is randomly assigned, because there are no common causes of E and any other variable. Thus, our results demonstrate that randomized trials are not free of selection bias as a result of differential loss to follow up because such selection occurs after the randomization.

Figures 6c and d are variations of Figures 6a and b, respectively, in which there is a common cause U* of E and another measured variable. U* indicates unmeasured lifestyle/personality/educational variables that determine both treatment (through the arrow from U* to E) and either attitudes toward attending study visits (through the arrow from U* to C in Fig. 6c) or threshold for reporting symptoms (through the arrow from U* to L in Fig. 6d). Again, these 2 are examples of bias resulting from structure (3) because the bias arises from conditioning on the common effect C of both a cause U* of E and a cause U of D. This particular bias has been referred to as M bias.[12] The bias caused by differential loss to follow up in Figures 6a–d is also referred to as bias due to informative censoring.

Nonresponse Bias/Missing Data Bias

The variable C in Figures 6a–d can represent missing data on the outcome for any reason, not just as a result of loss to follow up. For example, subjects could have missing data because they are reluctant to provide information or because they miss study visits. Regardless of the reasons why data on D are missing, standard analyses restricted to subjects with complete data (C = 0) will be biased.

Volunteer Bias/Self-selection Bias

Figures 6a–d can also represent a study in which C is agreement to participate (yes = 1, no = 0), E is cigarette smoking, D is coronary heart disease, U is family history of heart disease, and U* is healthy lifestyle. (L is any mediator between U and C such as heart disease awareness.) Under any of these structures, there would be no bias if the study population was a representative (ie, random) sample of the target population. However, bias will be present if the study is restricted to those who volunteered or elected to participate (C = 1). Volunteer bias cannot occur in a randomized study in which subjects are randomized (ie, exposed) only after agreeing to participate, because none of Figures 6a–d can represent such a trial. Figures 6a and b are eliminated because exposure cannot cause C. Figures 6c and d are eliminated because, as a result of the random exposure assignment, there cannot exist a common cause of exposure and any another variable.

Healthy Worker Bias

Figures 6a–d can also describe a bias that could arise when estimating the effect of a chemical E (an occupational exposure) on mortality D in a cohort of factory workers. The underlying unmeasured true health status U is a determinant of both death (D) and of being at work (C). The study is restricted to individuals who are at work (C = 1) at the time of outcome ascertainment. (L could be the result of blood tests and a physical examination.) Being exposed to the chemical is a predictor of being at work in the near future, either directly (eg, exposure can cause disabling asthma), like in Figures 6a and b, or through a common cause U* (eg, certain exposed jobs are eliminated for economic reasons and the workers laid off) like in Figures 6c and d.

This "healthy worker" bias is an example of bias resulting from structure (3) because it arises from conditioning on the censoring variable C, which is a common effect of (a cause of) exposure and (a cause of) the outcome. However, the term "healthy worker" bias is also used to describe the bias that occurs when comparing the risk in certain group of workers with that in a group of subjects from the general population. This second bias can be depicted by the DAG in Figure 1 in which L represents health status, E represents membership in the group of workers, and D represents the outcome of interest. There are arrows from L to E and D because being healthy affects job type and risk of subsequent outcome, respectively. In this case, the bias is caused by structure (1) and would therefore generally be considered to be the result of confounding.

These examples lead us to propose that the term selection bias in causal inference settings be used to refer to any bias that arises from conditioning on a common effect as in Figure 3 or its variations (Figs. 4–6).

In addition to the examples given here, DAGs have been used to characterize various other selection biases. For example, Robins[7] explained how certain attempts to eliminate ascertainment bias in studies of estrogens and endometrial cancer could themselves induce bias[16]; Hernán et al.[8] discussed incidence–prevalence bias in case-control studies of birth defects; and Cole and Hernán[9] discussed the bias that could be introduced by standard methods to estimate direct effects.[17,18] In Appendix A.2, we provide a final example: the bias that results from the use of the hazard ratio as an effect measure. We deferred this example to the appendix because of its greater technical complexity. (Note that standard DAGs do not represent "effect modification" or "interactions" between variables, but this does not affect their ability to represent the causal structures that produce bias, as more fully explained in Appendix A.3).

To demonstrate the generality of our approach to selection bias, we now show that a bias that arises in longitudinal studies with time-varying exposures[19] can also be understood as a form of selection bias.

Adjustment for Variables Affected by Previous Exposure (or its causes)

Consider a follow-up study of the effect of antiretroviral therapy (E) on viral load at the end of follow up ($D = 1$ if detectable, $D = 0$ otherwise) in HIV-infected subjects. The greater a subject's unmeasured true immunosuppression level (U), the greater her viral load D and the lower the CD4 count L (low = 1, high = 0). Treatment increases CD4 count, and the presence of low CD4 count (a proxy for the true level of immunosuppression) increases the probability of receiving treatment. We assume that, in truth but unknown to the data analyst, treatment has no causal effect on the outcome D. The DAGs in Figures 7a and b represent the first 2 time points of the study. At time 1, treatment E_1 is decided after observing the subject's risk factor profile L_1. (E_0 could be decided after observing L_0, but the inclusion of L_0 in the DAG would not essentially alter our main point.) Let E be the sum of E_0 and E_1. The cumulative exposure variable E can therefore take 3 values: 0 (if the subject is not treated at any time), 1 (if the subject is treated at time one only or at time 2 only), and 2 (if the subject is treated at both times). Suppose the analyst's interest lies in comparing the risk had all subjects been always treated ($E = 2$) with that had all subjects never been treated ($E = 0$), and that the causal risk ratio is 1.0 ($CRR_{ED} = 1$, when comparing $E = 2$ vs. $E = 0$).

To estimate the effect of E without bias, the analyst needs to be able to estimate the effect of each of its components E_0 and E_1 simultaneously and without bias.[17] As we will see, this is not possible using standard methods, even when data on L_1 are available, because lack of adjustment for L_1 precludes unbiased estimation of the causal effect of E_1 whereas adjustment for L_1 by stratification (or, equivalently, by conditioning, matching, or regression adjustment) precludes unbiased estimation of the causal effect of E_0.

Unlike previous structures, Figures 7a and 7b contain a common cause of the (component E_1 of) exposure E and the outcome D, so one needs to adjust for L_1 to eliminate

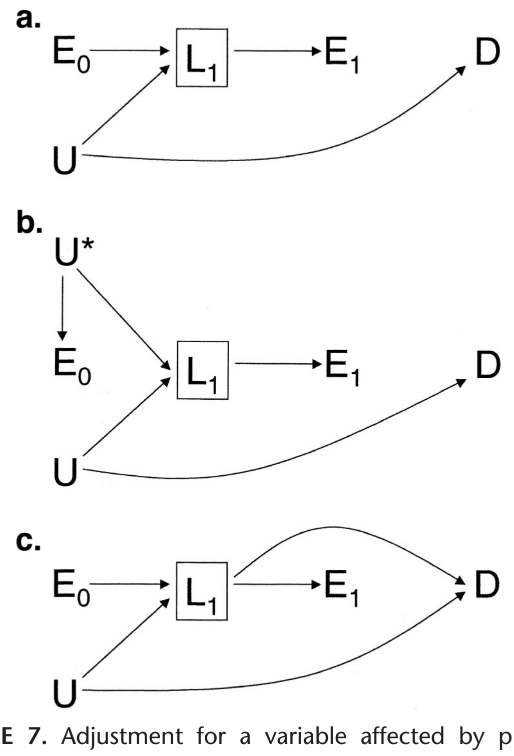


FIGURE 7. Adjustment for a variable affected by previous exposure.

confounding. The standard approach to confounder control is stratification: the associational risk ratio is computed in each level of the variable L_1. The square around the node L_1 denotes that the associational risk ratios ($ARR_{ED|L\ =\ 0}$ and $ARR_{ED|L\ =\ 1}$) are conditional on L_1. Examples of stratification-based methods are a Mantel-Haenzsel stratified analysis or regression models (linear, logistic, Poisson, Cox, and so on) that include the covariate L_1. (Not including interaction terms between L_1 and the exposure in a regression model is equivalent to assuming homogeneity of $ARR_{ED|L\ =\ 0}$ and $ARR_{ED|L\ =\ 1}$.) To calculate $ARR_{ED|L\ =\ 1}$, the data analyst has to select (ie, condition on) the subset of the population with value $L_1 = 1$. However, in this example, the process of choosing this subset results in selection on a variable L_1 affected by (a component E_0 of) exposure E and thus can result in bias as we now describe.

Although stratification is commonly used to adjust for confounding, it can have unintended effects when the association measure is computed within levels of L_1 and in addition L_1 is caused by or shares causes with a component E_0 of E. Among those with low CD4 count ($L_1 = 1$), being on treatment ($E_0 = 1$) makes it more likely that the person is severely immunodepressed; among those with a high level of CD4 ($L_1 = 0$), being off treatment ($E_0 = 0$) makes it more likely that the person is not severely immunodepressed. Thus, the side effect of stratification is to induce an association between prior exposure E_0 and U, and therefore between E_0 and the outcome D. Stratification eliminates confounding for E_1 at the cost of introducing selection bias for E_0. The net bias for any particular summary of the time-varying exposure that is used in the analysis (cumulative exposure, average exposure, and so on) depends on the relative magnitude of the confounding that is eliminated and the selection bias that is created. In summary, the associational (conditional) risk ratio $ARR_{ED|L_1}$, could be different from 1.0 even if the exposure history has no effect on the outcome of any subjects.

Conditioning on confounders L_1 which are affected by previous exposure can create selection bias even if the confounder is not on a causal pathway between exposure and outcome. In fact, no such causal pathway exists in Figures 7a and 7b. On the other hand, in Figure 7C the confounder L_1 for subsequent exposure E_1 lies on a causal pathway from earlier exposure E_0 to an outcome D. Nonetheless, conditioning on L_1 still results in selection bias. Were the potential for selection bias not present in Figure 7C (e.g., were U not a common cause of L_1 and D), the association of cumulative exposure E with the outcome D within strata of L_1 could be an unbiased estimate of the direct effect[18] of E not through L_1 but still would not be an unbiased estimate of the overall effect of E on D, because the effect of E_0 mediated through L_1 is not included.

ADJUSTING FOR SELECTION BIAS

Selection bias can sometimes be avoided by an adequate design such as by sampling controls in a manner to ensure that they will represent the exposure distribution in the population. Other times, selection bias can be avoided by appropriately adjusting for confounding by using alternatives to stratification-based methods (see subsequently) in the presence of time-dependent confounders affected by previous exposure.

However, appropriate design and confounding adjustment cannot immunize studies against selection bias. For example, loss to follow up, self-selection, and, in general, missing data leading to bias can occur no matter how careful the investigator. In those cases, the selection bias needs to be explicitly corrected in the analysis, when possible.

Selection bias correction, as we briefly describe, could sometimes be accomplished by a generalization of inverse probability weighting[20–23] estimators for longitudinal studies. Consider again Figures 6a–d and assume that L is measured. Inverse probability weighting is based on assigning a weight to each selected subject so that she accounts in the analysis not only for herself, but also for those with similar characteristics (ie, those with the same vales of L and E) who were not selected. The weight is the inverse of the probability of her selection. For example, if there are 4 untreated women, age 40–45 years, with CD4 count >500, in our cohort study, and 3 of them are lost to follow up, then these 3 subjects do not contribute to the analysis (ie, they receive a zero weight), whereas the remaining woman receives a weight of 4. In other words, the (estimated) conditional probability of remaining uncensored is 1/4 = 0.25, and therefore the (estimated) weight for the uncensored subject is 1/0.25 = 4. Inverse probability weighting creates a pseudopopulation in which the 4 subjects of the original population are replaced by 4 copies of the uncensored subject.

The effect measure based on the pseudopulation, in contrast to that based on the original population, is unaffected by selection bias provided that the outcome in the uncensored subjects truly represents the unobserved outcomes of the censored subjects (with the same values of E and L). This provision will be satisfied if the probability of selection (the denominator of the weight) is calculated conditional on E and on all additional factors that independently predict both selection and the outcome. Unfortunately, one can never be sure that these additional factors were identified and recorded in L, and thus the causal interpretation of the resulting adjustment for selection bias depends on this untestable assumption.

One might attempt to remove selection bias by stratification (ie, by estimating the effect measure conditional on the L variables) rather than by weighting. Stratification could yield unbiased conditional effect measures within levels of L under the assumptions that all relevant L variables were measured *and* that the exposure does not cause or share a common cause with any variable in L. Thus, stratification would work (ie, it would provide an unbiased conditional effect measure) under the causal structures depicted in Figures 6a and c, but not under those in Figures 6b and d. Inverse probability weighting appropriately adjusts for selection bias under all these situations because this approach is not based on estimating effect measures conditional on the covariates L, but rather on estimating unconditional effect measures after reweighting the subjects according to their exposure and their values of L.

Inverse probability weighting can also be used to adjust for the confounding of later exposure E_1 by L_1, even when exposure E_0 either causes L_1 or shares a common cause with L_1 (Figs. 7a–7c), a situation in which stratification fails. When using inverse probability weighting to adjust for confounding, we model the probability of exposure or treatment given past exposure and past L so that the denominator of a subject's weight is, informally, the subject's conditional probability of receiving her treatment history. We therefore refer to this method as inverse-probability-of-treatment weighting.[22]

One limitation of inverse probability weighting is that all conditional probabilities (of receiving certain treatment or censoring history) must be different from zero. This would not be true, for example, in occupational studies in which the probability of being exposed to a chemical is zero for those not working. In these cases, g-estimation[19] rather than inverse probability weighting can often be used to adjust for selection bias and confounding.

The use of inverse probability weighting can provide unbiased estimates of causal effects even in the presence of selection bias because the method works by creating a pseudopopulation in which censoring (or missing data) has been abolished and in which the effect of the exposure is the same as in the original population. Thus, the pseudopopulation effect measure is equal to the effect measure had nobody been censored. For example, Figure 8 represents the pseudopulation corresponding to the population of Figure 6a when the weights were estimated conditional on L and E. The censoring node is now lower-case because it does not correspond to a random variable but to a constant (everybody is uncensored in the pseudopopulation). This interpretation is desirable when censoring is the result of loss to follow up or nonresponse, but questionably helpful when censoring is the result of competing risks. For example, in a study aimed at estimating the effect of certain exposure on the risk of Alzheimer's disease, we might not wish to base our effect estimates on a pseudopopulation in which all other causes of death (cancer, heart disease, stroke, and so on) have been removed, because it is unclear even conceptually what sort of medical intervention would produce such a population. Another more pragmatic reason is that no feasible intervention could possibly remove just one cause of death without affecting the others as well.[24]

DISCUSSION

The terms "confounding" and "selection bias" are used in multiple ways. For instance, the same phenomenon is sometimes named "confounding by indication" by epidemiologists and "selection bias" by statisticians/econometricians. Others use the term "selection bias" when "confounders" are unmeasured. Sometimes the distinction between confounding and selection bias is blurred in the term "selection confounding."

We elected to refer to the presence of common causes as "confounding" and to refer to conditioning on common effects as "selection bias." This structural definition provides a clearcut classification of confounding and selection bias, even though it might not coincide perfectly with the traditional, often discipline-specific, terminologies. Our goal, however, was not to be normative about terminology, but rather to emphasize that, regardless of the particular terms chosen, there are 2 distinct causal structures that lead to these biases. The magnitude of both biases depends on the strength of the causal arrows involved.[12,25] (When 2 or more common effects have been conditioned on, an even more general formulation of selection bias is useful. For a brief discussion, see Appendix A.4.)

The end result of both structures is the same: noncomparability (also referred to as lack of exchangeability) between the exposed and the unexposed. For example, consider a cohort study restricted to firefighters that aims to estimate the effect of being physically active (E) on the risk of heart disease (D) (as represented in Fig. 9). For simplicity, we have assumed that, although unknown to the data analyst, E does not cause D. Parental socioeconomic status (L) affects the

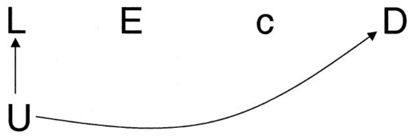


FIGURE 8. Causal diagram in the pseudopopulation created by inverse–probability weighting.

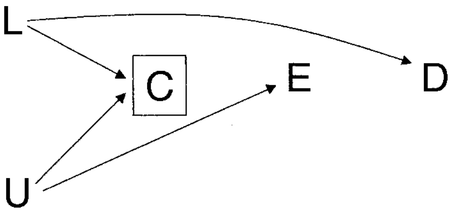


FIGURE 9. The firefighters' study.

risk of becoming a firefighter (C) and, through childhood diet, of heart disease (D). Attraction toward activities that involve physical activity (an unmeasured variable U) affects the risk of becoming a firefighter and of being physically active (E). U does not affect D, and L does not affect E. According to our terminology, there is no confounding because there are no common causes of E and D. Thus, if our study population had been a random sample of the target population, the crude associational risk ratio ARR_{ED} would have been equal to the causal risk ratio CRR_{ED} of 1.0.

However, in a study restricted to firefighters, the crude ARR_{ED} and CRR_{ED} would differ because conditioning on a common effect C of causes of exposure and outcome induces selection bias resulting in noncomparability of the exposed and unexposed firefighters. To the study investigators, the distinction between confounding and selection bias is moot because, regardless of nomenclature, they must stratify on L to make the exposed and the unexposed firefighters comparable. This example demonstrates that a structural classification of bias does not always have consequences for either the analysis or interpretation of a study. Indeed, for this reason, many epidemiologists use the term "confounder" for any variable L on which one has to stratify to create comparability, regardless of whether the (crude) noncomparability was the result of conditioning on a common effect or the result of a common cause of exposure and disease.

There are, however, advantages of adopting a structural or causal approach to the classification of biases. First, the structure of the problem frequently guides the choice of analytical methods to reduce or avoid the bias. For example, in longitudinal studies with time-dependent confounding, identifying the structure allows us to detect situations in which stratification-based methods would adjust for confounding at the expense of introducing selection bias. In those cases, inverse probability weighting or g-estimation are better alternatives. Second, even when understanding the structure of bias does not have implications for data analysis (like in the firefighters' study), it could still help study design. For example, investigators running a study restricted to firefighters should make sure that they collect information on joint risk factors for the outcome and for becoming a firefighter. Third, selection bias resulting from conditioning on preexposure variables (eg, being a firefighter) could explain why certain variables behave as "confounders" in some studies but not others. In our example, parental socioeconomic status would not necessarily need to be adjusted for in studies not restricted to firefighters. Finally, causal diagrams enhance communication among investigators because they can be used to provide a rigorous, formal definition of terms such as "selection bias."

ACKNOWLEDGMENTS

We thank Stephen Cole and Sander Greenland for their helpful comments.

REFERENCES

1. Rothman KJ, Greenland S. *Modern Epidemiology*, 2nd ed. Philadelphia: Lippincott-Raven; 1998.
2. Szklo M0, Nieto FJ. *Epidemiology. Beyond the Basics*. Gaithersburg, MD: Aspen; 2000.
3. MacMahon B, Trichopoulos D. *Epidemiology. Principles & Methods*, 2nd ed. Boston: Little, Brown and Co; 1996.
4. Hennekens CH, Buring JE. *Epidemiology in Medicine*. Boston: Little, Brown and Co; 1987.
5. Gordis L. *Epidemiology*. Philadelphia: WB Saunders Co; 1996.
6. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–48.
7. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology*. 2001;11:313–320.
8. Hernán MA, Hernández-Diaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155:176–184.
9. Cole SR, Hernán MA. Fallibility in the estimation of direct effects. *Int J Epidemiol*. 2002;31:163–165.
10. Maclure M, Schneeweiss S. Causation of bias: the episcope. *Epidemiology*. 2001;12:114–122.
11. Greenland S, Brumback BA. An overview of relations among causal modeling methods. *Int J Epidemiol*. 2002;31:1030–1037.
12. Greenland S. Quantifying biases in causal models: classical confounding versus collider-stratification bias. *Epidemiology*. 2003;14:300–306.
13. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82: 669–710.
14. Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search. Lecture Notes in Statistics 81*. New York: Springer-Verlag; 1993.
15. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics*. 1946;2:47–53.
16. Greenland S, Neutra RR. An analysis of detection bias and proposed corrections in the study of estrogens and endometrial cancer. *J Chronic Dis*. 1981;34:433–438.
17. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to the healthy worker survivor effect [published errata appear in *Mathematical Modelling*. 1987;14: 917–921]. *Mathematical Modelling*. 1986;7:1393–1512.
18. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3:143–155.
19. Robins JM. Causal inference from complex longitudinal data. In: Berkane M, ed. *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics 120*. New York: Springer-Verlag; 1997:69–117.
20. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47:663–685.
21. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*. 2000;56:779–788.
22. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11:561–570.
23. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
24. Greenland S. Causality theory for policy uses of epidemiologic measures. In: Murray CJL, Salomon JA, Mathers CD, et al., eds. *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/WHO; 2002.
25. Walker AM. *Observation and Inference: An introduction to the Methods of Epidemiology*. Newton Lower Falls: Epidemiology Resources Inc; 1991.
26. Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*. 1996;7:498–501.

APPENDIX

A.1. Causal and Associational Risk Ratio

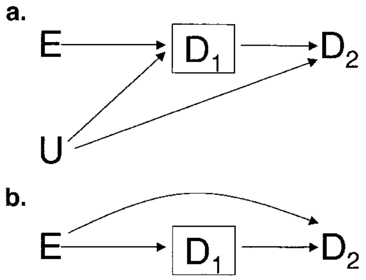
For a given subject, *E* has a causal effect on *D* if the subject's value of *D* had she been exposed differs from the value of *D* had she remained unexposed. Formally, letting $D_{i,\ e\ =\ 1}$ and $D_{i,e\ =\ 0}$ be subject's *i* (counterfactual or potential) outcomes when exposed and unexposed, respectively, we say there is a causal effect for subject *i* if $D_{i,\ e\ =\ 1} \neq D_{i,\ e\ =\ 0}$. Only one of the counterfactual outcomes can be observed for each subject (the one corresponding to his observed exposure), ie, $D_{i,\ e} = D_i$ if $E_i = e$, where D_i and E_i represent subject *i*'s observed outcome and exposure. For a population, we say that there is no average causal effect (preventive or causative) of *E* on *D* if the average of *D* would remain unchanged whether the whole population had been treated or untreated, ie, when $\Pr(D_{e\ =\ 1} = 1) = \Pr(D_{e\ =\ 0} = 1)$ for a dichotomous *D*. Equivalently, we say that *E* does not have a causal effect on *D* if the causal risk ratio is one, ie, $\mathrm{CRR}_{\mathrm{ED}} = \Pr(D_{e\ =\ 1} = 1)/\Pr(D_{e\ =\ 0} = 1) = 1.0$. For an extension of counterfactual theory and methods to complex longitudinal data, see reference 19.

In a DAG, $\mathrm{CRR}_{\mathrm{ED}} = 1.0$ is represented by the lack of a directed path of arrows originating from *E* and ending on *D* as, for example, in Figure 5. We shall refer to a directed path of arrows as a causal path. On the other hand, in Figure 5, $\mathrm{CRR}_{\mathrm{EC}} \neq 1.0$ because there is a causal path from *E* to *C* through *F*. The lack of a direct arrow from *E* to *C* implies that *E* does not have a direct effect on *C* (relative to the other variables on the DAG), ie, the effect is wholly mediated through other variables on the DAG (ie, *F*).

For a population, we say that there is no association between *E* and *D* if the average of *D* is the same in the subset of the population that was exposed as in the subset that was unexposed, ie, when $\Pr(D = 1|E = 1) = \Pr(D = 1|E = 0)$ for a dichotomous *D*. Equivalently, we say that *E* and *D* are unassociated if the associational risk ratio is 1.0, ie, $\mathrm{ARR}_{\mathrm{ED}} = \Pr(D = 1|E = 1) / \Pr(D = 1|E = 0) = 1.0$. The associational risk ratio can always be estimated from observational data. We say that there is bias when the causal risk ratio in the population differs from the associational risk ratio, ie, $\mathrm{CRR}_{\mathrm{ED}} \neq \mathrm{ARR}_{\mathrm{ED}}$.

A.2. Hazard Ratios as Effect Measures

The causal DAG in Appendix Figure 1a describes a randomized study of the effect of surgery *E* on death at times 1 (D_1) and 2 (D_2). Suppose the effect of exposure on D_1 is protective. Then the lack of an arrow from *E* to D_2 indicates that, although the exposure *E* has a direct protective effect (decreases the risk of death) at time 1, it has no direct effect on death at time 2. That is, the exposure does not influence the survival status at time D_2 of any subject who would survive past time 1 when unexposed (and thus when exposed). Suppose further that *U* is an unmeasured haplotype that decreases the subject's risk of death at all times. The associational risk ratios $\mathrm{ARR}_{\mathrm{ED}_1}$ and $\mathrm{ARR}_{\mathrm{ED}_2}$ are unbiased measures of the effect of *E* on death at times 1 and 2, respectively. (Because of the absence of confounding, $\mathrm{ARR}_{\mathrm{ED}_1}$ and $\mathrm{ARR}_{\mathrm{ED}_2}$ equal the causal risk ratios $\mathrm{CRR}_{\mathrm{ED}_1}$ and $\mathrm{CRR}_{\mathrm{ED}_2}$, respectively.) Note that, even though *E* has no direct effect on D_2, $\mathrm{ARR}_{\mathrm{ED}_2}$ (or, equivalently, $\mathrm{CRR}_{\mathrm{ED}_2}$) will be less than 1.0 because it is a measure of the effect of E on total mortality through time 2.



Appendix Figure 1. Effect of exposure on survival.

Consider now the time-specific associational hazard (rate) ratio as an effect measure. In discrete time, the hazard of death at time 1 is the probability of dying at time 1 and thus is the same as $\mathrm{ARR}_{\mathrm{ED}_1}$. However, the hazard at time 2 is the probability of dying at time 2 among those who survived past time 1. Thus, the associational hazard ratio at time 2 is then $\mathrm{ARR}_{\mathrm{ED}_2}|D_1 = 0$. The square around D_1 in Appendix Figure 1a indicates this conditioning. Exposed survivors of time 1 are less likely than unexposed survivors of time 1 to have the protective haplotype *U* (because exposure can explain their survival) and therefore are more likely to die at time 2. That is, conditional on $D_1 = 0$, exposure is associated with a higher mortality at time 2. Thus, the hazard ratio at time 1 is less than 1.0, whereas the hazard ratio at time 2 is greater than 1.0, ie, the hazards have crossed. We conclude that the hazard ratio at time 2 is a biased estimate of the direct effect of exposure on mortality at time 2. The bias is selection bias arising from conditioning on a common effect D_1 of exposure and of *U*, which is a cause of D_2 that opens the noncausal (ie, associational) path $\mathrm{E} \rightarrow \mathrm{D}_1 \leftarrow \mathrm{U} \rightarrow \mathrm{D}_2$ between *E* and D_2.[13] In the survival analysis literature, an unmeasured cause of death that is marginally unassociated with exposure such as *U* is often referred to as a frailty.

In contrast to this, the conditional hazard ratio $\mathrm{ARR}_{\mathrm{ED}_2|\mathrm{D}_1\ =\ 0,\mathrm{U}}$ at D_2 given *U* is equal to 1.0 within each stratum of *U* because the path $\mathrm{E} \rightarrow \mathrm{D}_1 \leftarrow \mathrm{U} \rightarrow \mathrm{D}_2$ between *E* and D_2 is now blocked by conditioning on the noncollider *U*. Thus, the conditional hazard ratio correctly indicates the absence of a direct effect of *E* on D_2. The fact that the unconditional hazard ratio $\mathrm{ARR}_{\mathrm{ED}_2|\mathrm{D}_1} = 0$ differs from the common-stratum specific hazard ratios of 1.0 even though *U* is independent of *E,* shows the noncollapsibility of the hazard ratio.[26]

Unfortunately, the unbiased measure $\text{ARR}_{ED_2|D_1\,=\,0,U}$ of the direct effect of E on D_2 cannot be computed because U is unobserved. In the absence of data on U, it is impossible to know whether exposure has a direct effect on D_2. That is, the data cannot determine whether the true causal DAG generating the data was that in Appendix Figure 1a versus that in Appendix Figure 1b.

A.3. Effect Modification and Common Effects in DAGs

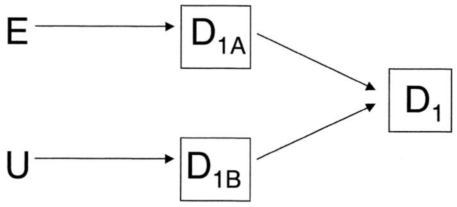
Although an arrow on a causal DAG represents a direct effect, a standard causal DAG does not distinguish a harmful effect from a protective effect. Similarly, a standard DAG does not indicate the presence of effect modification. For example, although Appendix Figure 1a implies that both E and U affect death D_1, the DAG does not distinguish among the following 3 qualitatively distinct ways that U could modify the effect of E on D_1:

1. The causal effect of exposure E on mortality D_1 is in the same direction (ie, harmful or beneficial) in both stratum $U = 1$ and stratum $U = 0$.
2. The direction of the causal effect of exposure E on mortality D_1 in stratum $U = 1$ is the opposite of that in stratum $U = 0$ (ie, there is a qualitative interaction between U and E).
3. Exposure E has a causal effect on D_1 in one stratum of U but no causal effect in the other stratum, eg, E only kills subjects with $U = 0$.

Because standard DAGs do not represent interaction, it follows that it is not possible to infer from a DAG the direction of the conditional association between 2 marginally independent causes (E and U) within strata of their common effect D_1. For example, suppose that, in the presence of an undiscovered background factor V that is unassociated with E or U, having either $E = 1$ or $U = 1$ is sufficient and necessary to cause death (an "or" mechanism), but that neither E nor U causes death in the absence of V. Then among those who died by time 1 ($D_1 = 1$), E and U will be negatively associated, because it is more likely that an unexposed subject ($E = 0$) had $U = 1$ because the absence of exposure increases the chance that U was the cause of death. (Indeed, the logarithm of the conditional odds ratio $\text{OR}_{UE|D_1} = 1$ will approach minus infinity as the population prevalence of V approaches 1.0.) Although this "or" mechanism was the only explanation given in the main text for the conditional association of independent causes within strata of a common effect; nonetheless, other possibilities exist. For example, suppose that in the presence of the undiscovered background factor V, having both $E = 1$ and $U = 1$ is sufficient and necessary to cause death (an "and" mechanism) and that neither E nor U causes death in the absence of V. Then, among those who die by time 1, those who had been exposed ($E = 1$) are more likely to have the haplotype ($U = 1$), ie, E and U are positively correlated. A standard DAG such as that in Appendix Figure 1a fails to distinguish between the case of E and U interacting through an "or" mechanism from the case of an "and" mechanism.

Although conditioning on common effect D_1 always induces a conditional association between independent causes E and U in at least one of the 2 strata of D_1 (say, $D_1 = 1$), there is a special situation under which E and U remain conditionally independent within the other stratum (say, $D_1 = 0$). This situation occurs when the data follow a multiplicative survival model. That is, when the probability, $\Pr[D_1 = 0 \mid U = u, E = e]$, of survival (ie, $D_1 = 0$) given E and U is equal to a product $g(u)\ h(e)$ of functions of u and e. The multiplicative model $\Pr[D_1 = 0 \mid U = u,\ E = e] = g(u)\ h(e)$ is equivalent to the model that assumes the survival ratio $\Pr[D_1 = 0 \mid U = u,\ E = e]/\Pr[D_1 = 0 \mid U = 0,\ E = 0]$ does not depend on u and is equal to $h(e)$. (Note that if $\Pr[D_1 = 0 \mid U = u,\ E = e] = g(u)\ h(e)$, then $\Pr[D_1 = 1 \mid U = u,\ E = e] = 1 - [g(u)\ h(e)]$ does not follow a multiplicative mortality model. Hence, when E and U are conditionally independent given $D_1 = 0$, they will be conditionally dependent given $D_1 = 1$.)

Biologically, this multiplicative survival model will hold when E and U affect survival through totally independent mechanisms in such a way that U cannot possibly modify the effect of E on D_1, and vice versa. For example, suppose that the surgery E affects survival through the removal of a tumor, whereas the haplotype U affects survival through increasing levels of low-density lipoprotein-cholesterol levels resulting in an increased risk of heart attack (whether or not a tumor is present), and that death by tumor and death by heart attack are independent in the sense that they do not share a common cause. In this scenario, we can consider 2 cause-specific mortality variables: death from tumor D_{1A} and death from heart attack D_{1B}. The observed mortality variable D_1 is equal to 1 (death)when either D_{1A} or D_{1B} is equal to 1, and D_1 is equal to 0 (survival) when both D_{1A} and D_{1B} equal 0. We assume the measured variables are those in Appendix Figure 1a so data on underlying cause of death is not recorded. Appendix Figure 2 is an expansion of Appendix Figure 1a that represents this scenario (variable D_2 is not represented because it is not essential to the current



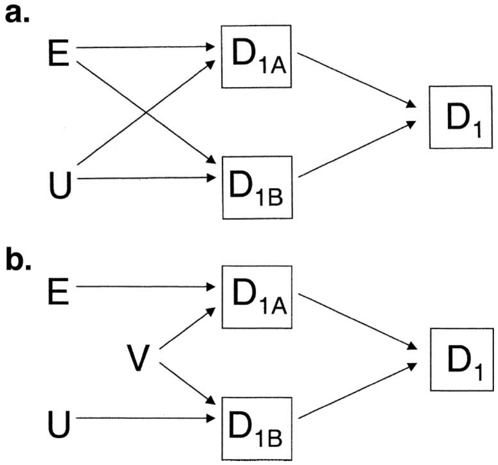
Appendix Figure 2. Multiplicative survival model.

discussion). Because $D_1 = 0$ implies both $D_{1A} = 0$ and $D_{1B} = 0$, conditioning on observed survival ($D_1 = 0$) is equivalent to simultaneously conditioning on $D_{1A} = 0$ and $D_{1B} = 0$ as well. As a consequence, we find by applying d-separation[13] to Appendix Figure 2 that E and U are conditionally independent given $D_1 = 0$, ie, the path, between E and U through the conditioned on collider D_1 is blocked by conditioning on the noncolliders D_{1A} and D_{1B}.[8] On the other hand, conditioning on $D_1 = 1$ does not imply conditioning on any specific values of D_{1A} and D_{1B} as the event $D_1 = 1$ is compatible with 3 possible unmeasured events $D_{1A} = 1$ and $D_{1B} = 1$, $D_{1A} = 1$ and $D_{1B} = 0$, and $D_{1A} = 0$ and $D_{1B} = 1$. Thus, the path between E and U through the conditioned on collider D_1 is not blocked, and thus E and U are associated given $D_1 = 1$.

What is interesting about Appendix Figure 2 is that by adding the unmeasured variables D_{1A} and D_{1B}, which functionally determine the observed variable D_1, we have created an annotated DAG that succeeds in representing both the conditional independence between E and U given $D_1 = 0$ and the their conditional dependence given $D_1 = 1$. As far as we are aware, this is the first time such a conditional independence structure has been represented on a DAG.

If E and U affect survival through a common mechanism, then there will exist an arrow either from E to D_{1B} or from U to D_{1A}, as shown in Appendix Figure 3a. In that case, the multiplicative survival model will not hold, and E and U will be dependent within both strata of D_1. Similarly, if the causes D_{1A} and D_{1B} are not independent because of a common cause V as shown in Appendix Figure 3b, the multiplicative survival model will not hold, and E and U will be dependent within both strata of D_1.

In summary, conditioning on a common effect always induces an association between its causes, but this association could be restricted to certain levels of the common effect.



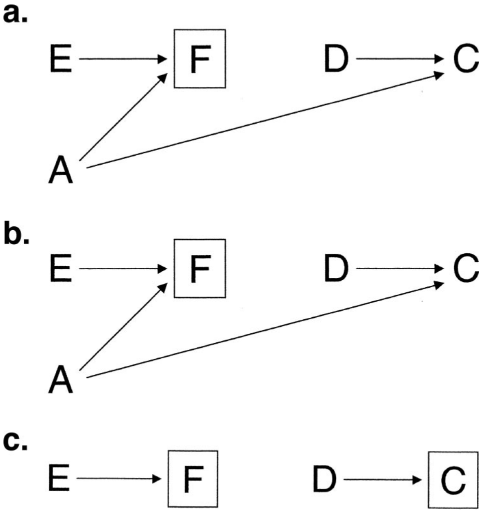
Appendix Figure 3. Multiplicative survival model does not hold.

A.4. Generalizations of Structure (3)

Consider Appendix Figure 4a representing a study restricted to firefighters ($F = 1$). E and D are unassociated among firefighters because the path $EFACD$ is blocked by C. If we then stratify on the covariate C like in Appendix Figure 4b, E and D are conditionally associated among firefighters in a given stratum of C; yet C is neither caused by E nor by a cause of E. This example demonstrates that our previous formulation of structure (3) is insufficiently general to cover examples in which we have already conditioned on another variable F before conditioning on C. Note that one could try to argue that our previous formulation works by insisting that the set (F,C) of all variables conditioned be regarded as a single supervariable and then apply our previous formulation with this supervariable in place of C. This fix-up fails because it would require E and D to be conditionally associated within joint levels of the super variable (C, F) in Appendix Figure 4c as well, which is not the case.

However, a general formulation that works in all settings is the following. A conditional association between E and D will occur within strata of a common effect C of 2 other variables, one of which is either the exposure or statistically associated with the exposure and the other is either the outcome or statistically associated with the outcome.

Clearly, our earlier formulation is implied by the new formulation and, furthermore, the new formulation gives the correct results for both Appendix Figures 4b and 4c. A drawback of this new formulation is that it is not stated purely in terms of causal structures, because it makes reference to (possibly noncausal) statistical associations. Now it actually is possible to provide a fully general formulation in terms of causal structures but it is not simple, and so we will not give it here, but see references 13 and 14.



Appendix Figure 4. Conditioning on 2 variables.