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

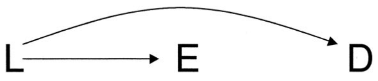


FIGURE 1. Common cause L of exposure E and outcome D.

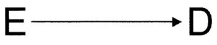


FIGURE 2. Causal effect of exposure E on outcome D.

the association measure is computed within levels of the common effect C, ie, the stratum-specific $\text{ARR}_{ED|C}$ 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 ($\text{ARR}_{ED} = 1.0$) 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 $\text{ARR}_{ED|C=1}$ 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

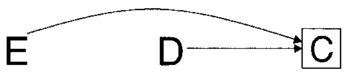


FIGURE 3. Conditioning on a common effect C of exposure E and outcome D.

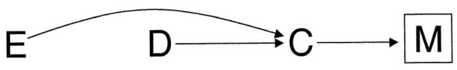


FIGURE 4. Conditioning on a common effect M of exposure E and outcome D.

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.

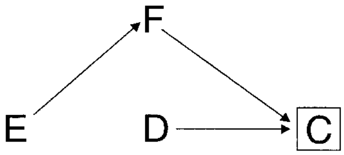


FIGURE 5. Selection bias in a case-control study. See text for details.

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-

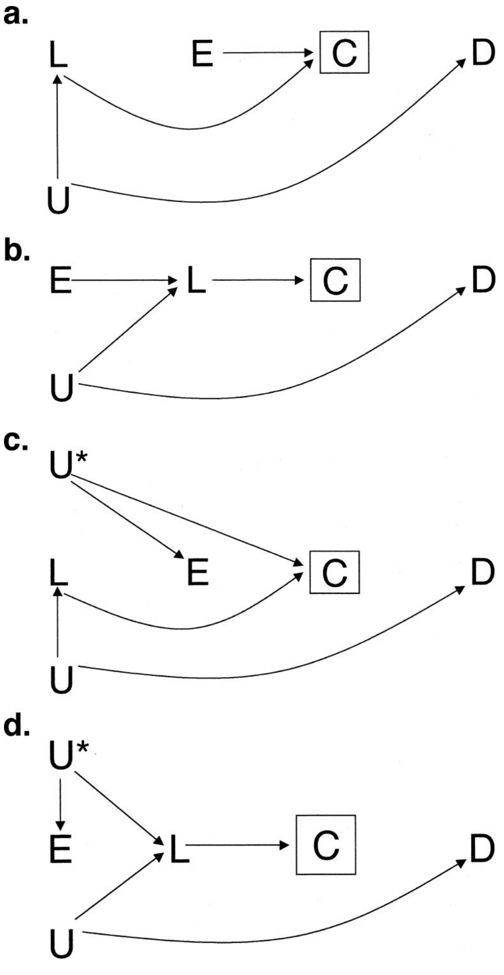


FIGURE 6. Selection bias in a cohort study. See text for details.

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

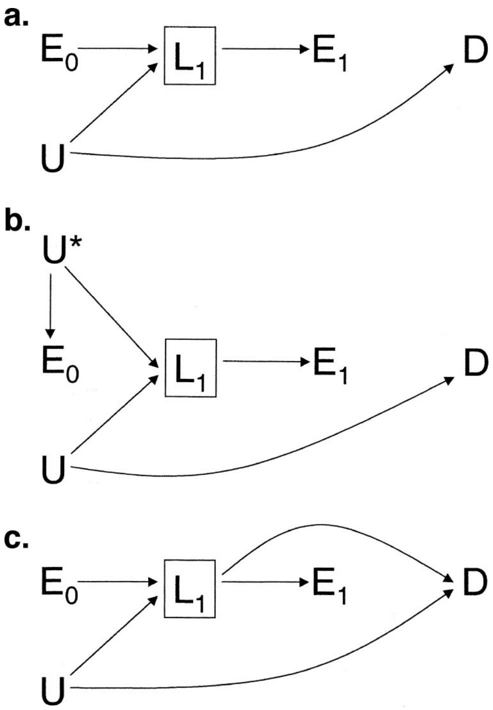


FIGURE 7. Adjustment for a variable affected by previous exposure.

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]

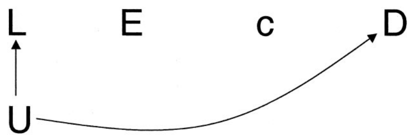


FIGURE 8. Causal diagram in the pseudopopulation created by inverse–probability weighting.

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

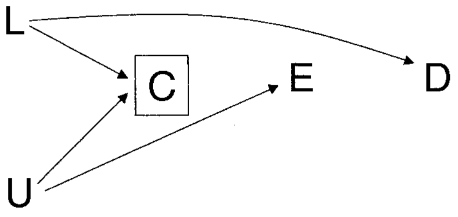


FIGURE 9. The firefighters' study.

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

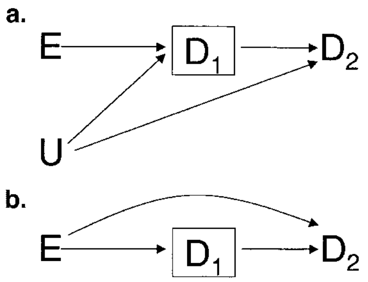
For a given subject, E has a causal effect on D if the subject's value of D had she been exposed differs from the value of D had she remained unexposed. Formally, letting $D_{i,\ e\ =\ 1}$ and $D_{i,e\ =\ 0}$ be subject's i (counterfactual or potential) outcomes when exposed and unexposed, respectively, 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, $\text{CRR}_{\text{ED}} = \Pr(D_{e\ =\ 1} = 1)/\Pr(D_{e\ =\ 0} = 1) = 1.0$. For an extension of counterfactual theory and methods to complex longitudinal data, see reference 19.

In a DAG, $\text{CRR}_{\text{ED}} = 1.0$ 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, $\text{CRR}_{\text{EC}} \neq 1.0$ 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, $\text{ARR}_{\text{ED}} = \Pr(D = 1|E = 1) / \Pr(D = 1|E = 0) = 1.0$. 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, $\text{CRR}_{\text{ED}} \neq \text{ARR}_{\text{ED}}$.

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 ARR_{ED_1} and ARR_{ED_2} are unbiased measures of the effect of E on death at times 1 and 2, respectively. (Because of the absence of confounding, ARR_{ED_1} and ARR_{ED_2} equal the causal risk ratios CRR_{ED_1} and CRR_{ED_2}, respectively.) Note that, even though E has no direct effect on D_2, ARR_{ED_2} (or, equivalently, CRR_{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 ARR_{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 $\text{ARR}_{\text{ED}_2}|\text{D}_1 = 0$. 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 $\text{E} \rightarrow \text{D}_1 \leftarrow \text{U} \rightarrow \text{D}_2$ 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 $\text{ARR}_{\text{ED}_2|\text{D}_1\ =\ 0,\text{U}}$ at D_2 given U is equal to 1.0 within each stratum of U because the path $\text{E} \rightarrow \text{D}_1 \leftarrow \text{U} \rightarrow \text{D}_2$ 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 $\text{ARR}_{\text{ED}_2|\text{D}_1} = 0$ 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 $\mathrm{ARR}_{ED_2|D_1\ =\ 0,U}$ 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

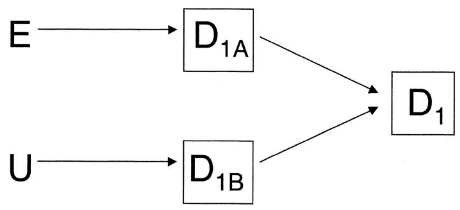
Although an arrow on a causal DAG represents a direct effect, a standard causal DAG does not distinguish a harmful effect from a protective effect. Similarly, a standard DAG does not indicate the presence of effect modification. For example, although Appendix Figure 1a implies that both E and U affect 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 $\mathrm{OR}_{UE|D_1} = 1$ 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, $\Pr[D_1 = 0|\ U = u,\ E = e]$, 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 $\Pr[D_1 = 0|\ U = u,\ E = e] = g(u)\ h(e)$ is equivalent to the model that assumes the survival ratio $\Pr[D_1 = 0|\ U = u,\ E = e]/\Pr[D_1 = 0|\ U = 0,\ E = 0]$ does not depend on u and is equal to $h(e)$. (Note that if $\Pr[D_1 = 0|\ U = u,\ E = e] = g(u)\ h(e)$, then $\Pr[D_1 = 1|\ U = u,\ E = e] = 1 - [g(u)\ h(e)]$ 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.

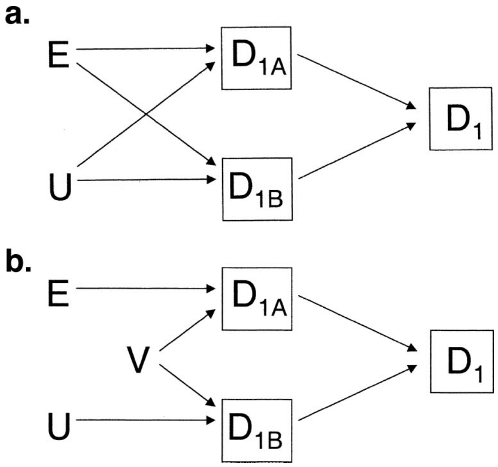
In summary, conditioning on a common effect always induces an association between its causes, but this association could be restricted to certain levels of the common effect.

A.4. Generalizations of Structure (3)

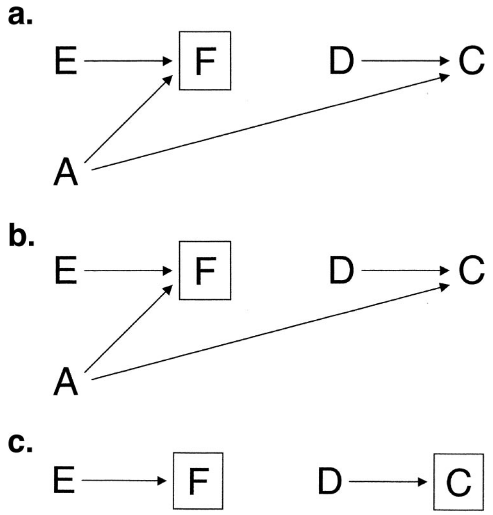
Consider Appendix Figure 4a representing a study restricted to firefighters ($F = 1$). E and D are unassociated among firefighters because the path $EFACD$ is blocked by C. If we then stratify on the covariate C like in Appendix Figure 4b, E and D are conditionally associated among firefighters in a 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 3. Multiplicative survival model does not hold.



Appendix Figure 4. Conditioning on 2 variables.