Confounding—Guilt by Association! (Student's Version)

Introduction

Generally, epidemiologic studies are directed at answering questions about health-related events in a community. One question is, What is the extent of a disease or health event in a particular community? To answer this question, the epidemiologist will consult various sources of health records, such as hospital admissions, disease registries, doctors' records, and so on. Such information is crucial for a community in planning and sustaining a health care system for its people.

Beyond planning for health services, the epidemiologist is interested in asking another set of questions: Why are the people in the community experiencing a particular health event? Can the cause of the disease be identified? What might be the factors that alter a person's risk for a particular health problem? Can these factors be controlled or eliminated, thereby reducing the risk of a particular disease or health problem?

To simplify the language that we use in this module, we will use the terms disease and exposure. When we refer to **disease,** we mean the health event or outcome that we are interested in studying. This is often a medical condition (such as cancer or heart disease), but it can also be a psychologic or social problem (e.g., depression, homelessness or poor academic performance). It can also be something positive, like recovery from AIDS. When we refer to **exposure,** we mean exposure to the factor that we are interested in investigating as a possible cause of the disease. A wide range of factors can be studied, things like exposure to radiation or industrial chemicals, behavioral practices such as poor diet or lack of exercise, and even personal characteristics such as gender or age. So when we use the words disease and exposure, remember that we are using them as a convenient shorthand for a wide variety of outcomes and causes.

Questions

1. Can you think of other examples that correspond to our definition of disease? Can you think of other examples that correspond to our definition of exposure?

A classic case of identification of an association between an exposure and a disease is the study of smoking and cancer. From the first suspicion in the 1940s that there was a relation between smoking and cancer, epidemiologists have led the charge in designing studies, analyzing data and constructing logical arguments to demonstrate what is now a commonly accepted assertion: Smoking causes cancer. It may seem, with hindsight, that it should have been very easy for

epidemiologists to demonstrate that smoking causes cancer. The usual method of demonstrating such a causal relation is to perform an experiment. However, in the case of smoking, the medical community was working with two severe handicaps. First, there is no ethical way to perform an experiment to test whether smoking causes lung cancer in humans.

2. If there were no ethical issues to consider, how would you design an experiment to test whether smoking causes lung cancer in humans? Why would such an experiment be unethical?

The second problem was the tobacco companies. Should cigarettes be found to cause cancer, the companies' economic well-being would be severely threatened. Unfortunately for the health of very many individuals over a very long time, the tobacco companies had what seemed to be a solid argument: Without a randomized controlled experiment, a causal connection between cigarettes and cancer could not be shown.

The argument of the tobacco companies is based on the sound statistical principle that association does not necessarily imply causation. In the case of smoking and cancer, the argument would run something like this. Yes, the tobacco companies would agree, it is true that the proportion of people who get cancer is higher for cigarette smokers than nonsmokers, and thus there is an association between smoking and cancer. However, they continue, it may be that people who smoke cigarettes tend to have other exposures and make other lifestyle choices, and those are the real culprits in causing cancer. Or, at least, they assert, there is no logical way to single out smoking as the culprit.

3. Can you think of some examples of other exposures or lifestyle choices that might be the real culprits in causing lung cancer?

The problem here is that the tobacco company folks had a point. If people who smoke are more likely than nonsmokers to live in areas of high pollution, for example, it could well be that the pollution is causing the cancer and the smoking habit is just an innocent bystander, falsely accused merely because it is associated with the true cancer-causing agent, pollution. In epidemiologic terms, the tobacco companies were claiming that air pollution (or any other factor that can cause cancer) is a confounding variable. A **confounding variable** is a variable (say, pollution) that can cause the disease under study (cancer) and is also associated with the exposure of interest (smoking). The existence of confounding variables in smoking studies made it difficult to establish a clear causal link between smoking and cancer unless appropriate methods were used to adjust for the effect of the confounders. These associations are shown in Figures 1 and 2.

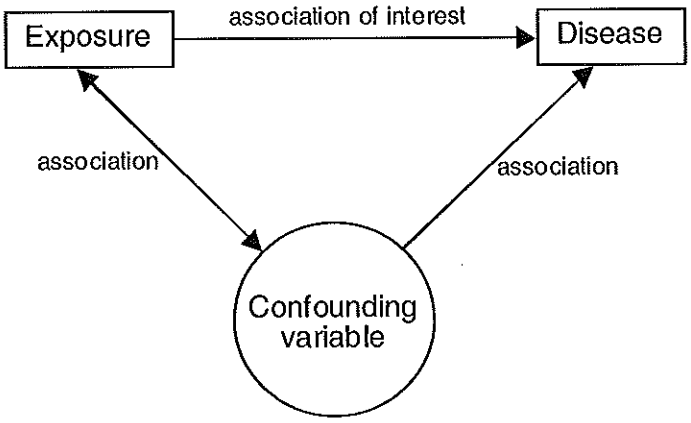

Figure 1. Associations of a hypothetical exposure, disease, and confounding variable

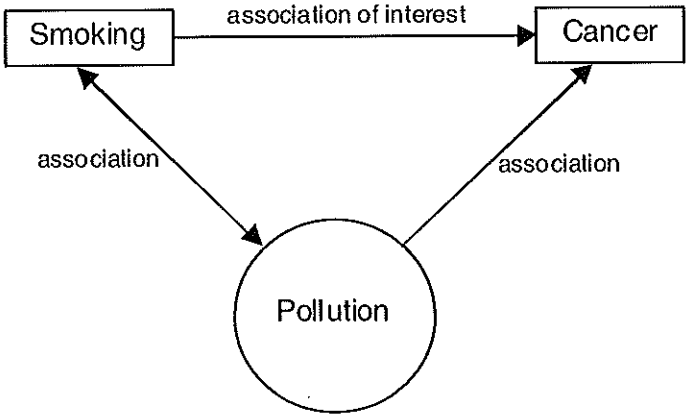

Figure 2. Associations of smoking, cancer, and pollution

How do epidemiologists identify potential confounding variables? When such variables are suspected, how can epidemiologists construct causal arguments in the face of these possible confounding variables? That is, how might smoking be proven to cause cancer if there are confounding variables that represent plausible rival explanations for cancer? These are the questions you will learn to answer in this module.

Making a case that a particular exposure is the cause of a particular health event is much like an exciting murder mystery or detective story. The detective must gather evidence about the crime, examine the witnesses, and then finally, on the last page, unmask the true criminal. The epidemiologist—let's call him or her Dr. Watson, of course—begins with the crime, a particular disease or health problem. The prime suspect is a particular exposure, such as smoking, thought to

be causally related to the disease. The prime suspect denies it, of course, and points to other equally plausible suspects. It is now Dr. Watson's responsibility to investigate those other suspects, those plausible rival guilty parties—those confounding variables! We will begin by rounding up the usual suspects, which means confronting our first problem: How do we identify them?

Rounding Up the Suspects: Quantifying the Association Between Exposure and Disease

The search for factors that might be causally related to a disease begins with the idea that people who have the exposure should have a different frequency of the disease from those who do not have the exposure. If an agent, such as a mosquito, causes West Nile Fever, for example, people who have been bitten by mosquitoes should have a higher frequency of the disease than those not bitten.

On the other hand, the exposure could be a vaccine, in which case those who have been given the vaccine should have less of a chance of getting the disease than those who did not receive the vaccine—the vaccine should act as a protective factor. Indications of differences in the chance of getting the disease would appear in actual data as different proportions of people having the disease, depending on exposure. Let's see how this works.

In epidemiology, a common type of study is the cohort study, in which a group of people is identified and followed over a period of time. For all individuals in the cohort, the investigator keeps track of whether or not they are exposed and whether or not they develop the disease. The information is usually presented in a 2 × 2 table such as the following:

Table 1. 2 × 2 Layout: Cohort Study

	Develop Disease	Do Not Develop Disease	Total
Exposed	*a*	*b*	*a* + *b*
Not Exposed	*c*	*d*	*c* + *d*

The first thing we can do with the numbers arranged this way is to calculate what proportion of exposed people and what proportion of unexposed people developed the disease.

4. Try to fill in the blanks below, using the letters in the cells in Table 1. To get you started, the first row has been completed for you.

Number of exposed people who develop disease	a
Number of exposed people who do not develop disease	
Total number of exposed people	
Number of unexposed people who develop disease	
Number of unexposed people who do not develop disease	
Total number of unexposed people	
Proportion of exposed people who develop disease	
Proportion of unexposed people who develop disease	

The proportions you calculated above (i.e., $a/(a+b)$, $c/(c+d)$) are called risks—they represent the risk that a person has of developing the disease. Another way to say this is that these proportions represent the *probability* that an individual would develop the disease over a specified period of time.

If the proportion of those exposed who develop the disease is greater than (>) the proportion of those not exposed who develop the disease, we would say that the exposure and the disease are positively associated. Expressed algebraically,

$$\frac{a}{a+b} > \frac{c}{c+d}$$

If the exposure is to a protective factor, the proportion of those exposed who develop the disease is less than (<) the proportion of those not exposed who develop the disease, and we would say that the exposure and the disease are negatively associated:

$$\frac{a}{a+b} < \frac{c}{c+d}$$

If the exposure is unrelated to the onset of the disease, we would expect the proportions to be equal,

$$\frac{a}{a+b} = \frac{c}{c+d}$$

in which case we would say that there is no association.

At this point we have a way of identifying whether or not there is an association and of determining whether the association is positive or negative, but we do not yet have a measure of the strength or magnitude of the association.

5. The proportions we have defined above represent proportions of individuals who have developed a disease. Can you think of a way to use these proportions to quantify the magnitude of association between exposure and disease?

The relative risk is one method of measuring the association between exposure and disease in cohort studies. The **relative risk,** as the name suggests, represents the probability of developing a disease among exposed individuals relative to the probability in unexposed individuals. Relative risks allow us to quantify how many times as likely individuals are to get the disease if exposed compared with if they were not exposed.

The relative risk (RR) is simply the ratio of the two risks we defined earlier, that is, the ratio of the risk of disease in the exposed compared with the risk of disease in the unexposed.

6. The RR is equal to the ratio of the risk of disease in the exposed to the risk of disease in the unexposed. Using the formulas for risk that you constructed in Question 4, see whether you can construct the formula for the RR.

What we are measuring with the relative risk is the degree of association between the exposure and development of the disease. If the relative risk is greater than 1, our interpretation is that the exposed individuals have a higher probability (or risk) of developing the disease. The greater the relative risk, the more strongly the exposure is associated with a higher frequency of disease. A relative risk less than 1 would be interpreted as indicating that the exposure leads to less risk of the disease, i.e., has a protective effect. The smaller the relative risk, the more strongly it is associated with a lower frequency of disease. A relative risk of 1 suggests that there is no association between the exposure and the disease.

We have now developed the mathematical method for the first task of an epidemiologist in the search for a causal relation between an exposure and a disease. If the relative risk for exposed persons compared with unexposed persons is greater than 1, we will take this as evidence that exposure is associated with the disease. (Similarly, if the relative risk for exposed persons compared with unexposed persons is less than 1, there is evidence that the exposure is associated with the absence of the disease. However, for the purpose of this presentation, we will focus on exposures that may be associated with higher risk, not lower risk, of disease.)

An Example: Bedsores and Mortality

To illustrate some of what we have just learned, we will use an example of a recent study of bedsores in a group of elderly patients who fractured a hip. When older persons fall and break a hip, they are often unable to move for many hours or even days. This immobility can be caused by many different factors, including loss of consciousness, pain, medications, traction and surgery. Long periods of immobility in turn can result in the person's getting bedsores. Bedsores are skin wounds that occur when a person lies motionless for long periods of time. Some bedsores are fairly superficial, but some extend as far down as the muscle or the bone. They are painful and difficult to treat and can result in many serious complications, some of which are fatal. This study was done to examine the association between bedsores and death among elderly hip fracture patients. (By the way, this example is based on a real study, but the numbers have been changed slightly to illustrate more clearly our teaching points.)

In this study, 9,400 patients aged 60 and over were selected. To be eligible, patients had to have been admitted with a diagnosis of hip fracture to one of 20 study hospitals. The patients' medical charts were reviewed by research nurses to obtain information about whether they developed a bedsore during hospitalization and whether they died while in hospital. The results are shown in Table 2. Notice that this table is set up the same way as Table 1.

Table 2. Results of Bedsores Study, with Totals

	Died	Did Not Die	Total
Bedsore	79	745	824
No Bedsores	286	8,290	8,576
Total	365	9,035	9,400

7. What is the exposure in this example? What is the disease?
8. Try to fill in the blanks below, using the information from Table 2. To get you started, the first row has been completed for you.

Number of people with a bedsore who died	79
Number of people with a bedsore who did not die	
Total number of people with a bedsore	
Number of people without a bedsore who died	
Number of people without a bedsore who did not die	
Total number of people without a bedsore	
Proportion of people with a bedsore who died	
Proportion of people without a bedsore who died	

9. Insert the appropriate numbers in the formula to calculate the relative risk.

This calculation tells us that the probability of death was 2.9 times as high in people with bedsores as in people without bedsores. Therefore, it would seem that there is a fairly strong association between bedsores and death. Remember, if there was no association between bedsores and death, the risk of death would be the same in people with bedsores and people without bedsores, and the relative risk would be 1. Can we conclude, however, that bedsores *cause* people to die? Isn't it also possible that people with bedsores are more likely than people without bedsores to have some other characteristic, and that it is the other characteristic that is the cause of the higher death rate?

10. Can you think of any possible confounding variables? A confounding variable in this example would be a characteristic that is more common in people with bedsores than in people without bedsores and that is associated with a higher death rate.

The investigators suspected that people who have lots of medical problems aside from their hip fracture are more likely to get bedsores and are also more likely to die. If that were the case, the severity of medical problems would be a confounding variable.

11. Fill in Figure 3, using the bedsores, death and medical severity example. Look back at Figures 1 and 2 for guidance.

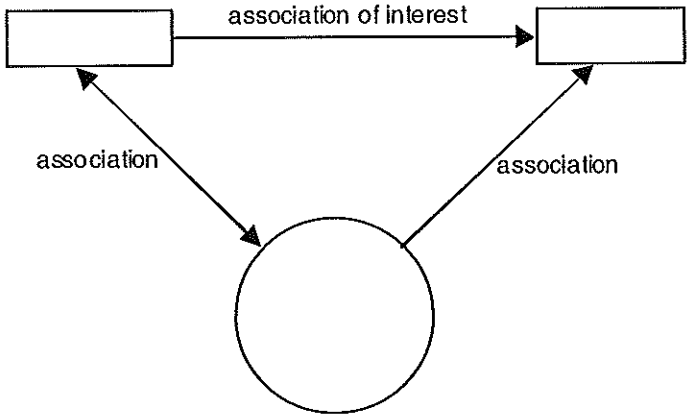

Figure 3. Associations of bedsores, death, and severity of medical problems

As a part of the design of the study, the epidemiologists who did this study obtained information about the patients' other medical problems. The information was summarized into a score based on information about the patients' diseases when they were admitted to hospital. To create two groups, the researchers classified everyone who had a score of 5 or more into the high medical severity group and everyone whose score was less than 5 into the low medical severity group. Using the data from their study, they were able to establish the following facts:

- Of the 79 people who had bedsores and died, 55 had high medical severity and 24 had low medical severity.
- Of the 745 people who had bedsores and did not die, 51 had high medical severity and 694 had low medical severity.
- Of the 286 people who had no bedsores and died, 5 had high medical severity and 281 had low medical severity.
- Of the 8,290 people who had no bedsores and did not die, 5 had high medical severity and 8,285 had low medical severity.

A simple way to organize all this information (and a first step in investigating whether medical severity is a confounding variable) is to create two 2 × 2 tables: one for people with high medical severity and one for people with low medical severity.

12. Use the numbers given above to complete Tables 3a and 3b.

Table 3a. Results of Bedsores Study, High Medical Severity Group

	Died	Did Not Die	Total
Bedsore			
No Bedsores			
Total			

Table 3b. Results of Bedsores Study, Low Medical Severity Group

	Died	Did Not Die	Total
Bedsore			
No Bedsores			
Total			

13. Look carefully at Tables 3a and 3b. How does it help to see the results displayed in this way?

As before, we can calculate relative risks. This time, we will calculate relative risks separately for the high and low medical severity groups (the two strata):

14. Calculate the relative risk in the high medical severity group.

15. Calculate the relative risk in the low medical severity group.

In answering the two preceding questions, you created strata based on categories of the suspected confounding variable and examined the exposure–disease association within each stratum. This procedure is called **stratification.** In each stratum, the association between bedsores and death cannot be explained by medical severity because in each stratum medical severity is held constant (as long as we can assume that all those with low severity have similar severity and all those with high severity have similar severity). By looking at the relative risks separately in the high and low medical severity groups, we have effectively adjusted for medical severity. That is, we have estimated the bedsores–mortality association in a way that eliminates the effect of medical severity on that association. Stratification is one of several ways to adjust for a confounding variable.

Now, let us look more closely at those stratum-specific results. Both relative risks are very close to 1. This means that the relative risk of death comparing those with and without bedsores, *and adjusted for medical severity*, is about 1. This is quite different from the original relative risk of 2.9 that we found when we looked at the overall 2 × 2 table. If the unadjusted and adjusted relative risks had been similar, it would mean that medical severity did not confound the association between bedsores and death. The fact that the unadjusted and adjusted relative risks are different means that there is confounding by medical severity.

16. We said above that the fact that the unadjusted and the adjusted relative risks are different means that there is confounding by medical severity. Can you think why this is so?

17. If the unadjusted relative risk were similar to the relative risk adjusted for medical severity, what would you conclude about whether medical severity confounds the association between bedsores and mortality?

Back to Definitions

As we said earlier, when there is confounding, the confounder is associated with both the disease and the exposure (Figure 2). Let's determine whether these two conditions are met in the bedsores study. The first part of the statement says that when there is confounding, the confounder is associated with the disease. If the confounder is associated with the disease, we would expect that people with high medical severity would have a higher probability of death than people with low medical severity.

18. Using the results in Tables 3a and 3b, determine what proportion of the high medical severity group died. Then determine what proportion of the low medical severity group died.

19. Is the probability of death in the high severity group similar or different from the probability of death in the low severity group? What does this suggest about the association between medical severity and death? To what part of Figure 1 does this conclusion correspond?

The second part of the statement says that when there is confounding, the confounder is associated with the exposure. If the confounder is associated with the exposure, we would expect that the proportion of people with bedsores would be higher among people with high medical severity than in people with low medical severity.

20. Using the results in Tables 3a and 3b, determine what proportion of patients with high medical severity had bedsores. Then determine what proportion of patients with low medical severity had bedsores.

21. Is the proportion of patients with bedsores among high medical severity patients similar to or different from the proportion of patients with bedsores among low medical severity patients? What does this suggest about the association between medical severity and bedsores? To what part of Figure 1 does this conclusion correspond?

22. Earlier we said that two conditions have to be met in order for confounding to occur: (1) there must be an association between the confounder and the disease and (2) there must be an association between the confounder and the exposure. Can you explain why both conditions have to be present? To help you answer this question using the bedsores example, think about what would happen if patients with higher medical severity were at higher risk of dying than those with lower medical severity (confounder–disease association), but patients with bedsores were not more likely to have higher medical severity than those without bedsores (no confounder–exposure association). Then, think about what would happen if patients with bedsores were more likely to have higher medical severity than those without bedsores (confounder–exposure association), but patients with higher medical severity were not at higher risk of dying than patients with low medical severity (no confounder–disease association).

We have shown that both conditions for confounding have been met, and we confirm that medical severity does in fact confound the association between bedsores and death. If medical severity was associated with bedsores, but people with high medical severity did not have a higher probability of death than people with low medical severity, there would not be confounding by medical severity. Similarly, if people with high medical severity had a higher probability of death than people with low medical severity, but medical severity was not associated with bedsores, there would be no confounding. Both conditions have to be present for confounding to be present.

The Bottom Line

Confounding occurs when a variable is associated with both the exposure and the disease that we are studying. The presence of associations between the confounder and the exposure, and between the confounder and the disease, makes it seem as though the exposure is the cause of the disease, but really the exposure is only guilty by association. When the effect of an exposure is mixed with the effect of another variable (the confounding variable), we may incorrectly conclude that the disease is caused by the exposure. We might then attempt to eliminate the exposure in the hope that the disease could be prevented. If, however, the association between the exposure and the disease is due to confounding and is not causal, elimination of the exposure will not have any effect on the incidence of the disease.

In our example, bedsores were associated with the probability of dying. The relative risk was 2.9, indicating a fairly strong association. However, when we controlled for the patient's medical severity, we found that the adjusted relative risk was about 1, quite a bit lower than 2.9. The fact that the adjusted relative risk was different from the unadjusted relative risk is evidence that there is confounding. Another symptom of confounding was identified by showing that there was an association both between bedsores and medical severity and between dying and medical severity.

Adjusting for medical severity in our example made the adjusted relative risks go down to 1, indicating no association between bedsores and mortality. In other words, the apparent association

suggested by the unadjusted relative risk of 2.9 was completely explained by confounding by severity. When we adjusted for severity by calculating relative risks separately in the strata defined by medical severity, there was no association between bedsores and mortality.

Note that confounding does not always work this way. Sometimes, adjusting for a confounder makes the relative risk go down but not all the way down to 1. In such situations, there may be other confounding variables that were not adjusted for. Or it is possible that there is a causal association between the exposure and the disease, but that this association is less strong than we might conclude based on the unadjusted relative risk. Adjusting for confounding can also cause the adjusted relative risk to be higher than the unadjusted relative risk. This occurs when there is a negative association between the confounder and the disease or between the confounder and the exposure.

When we conclude that there is confounding, does this mean that the association between exposure and disease is not real? In our example, there is confounding by medical severity but does that mean that the association between bedsores and dying is not real? The answer is no. Patients with bedsores really do have a higher risk of dying, but it is not because they have bedsores. The observed association came about because people with high medical severity were overrepresented in the group of people with bedsores and these people with high medical severity bring with them a higher risk of dying. This is what makes the group with bedsores look as though they are at higher risk of dying—bedsores are *guilty by association.*

Let us return briefly to the case of smoking and lung cancer. At first the tobacco companies argued that the apparent association between smoking and lung cancer arose because of some factor that could cause lung cancer and that was more common in smokers than nonsmokers. In other words, they claimed that there might be one or more variables that confounded the association between smoking and lung cancer, and that smoking was only guilty by association. However, in the years since the first observations of an association between smoking and lung cancer, numerous studies have shown that confounding was not the explanation. Using different study designs and different study populations, epidemiologists and other scientists painstakingly proved that smoking does cause lung cancer.

23. If there is an association between an exposure and a disease, but the association is entirely due to confounding, what will happen if you develop an intervention to eliminate the exposure? Will you have an impact on the prevention of the disease?

If epidemiologists did not take great care to identify and control for confounding, incorrect conclusions would be drawn, and time and resources would be unnecessarily expended with little hope of improving the well-being of the population.