The Logic of Inference in Science

Epidemiology is the science of studying health-related events that affect populations. Like all science, it is built on the fundamental belief that precise observation and measurement, combined with careful reasoning in the light of existing knowledge, is the most effective way to proceed with that study. Epidemiologists are concerned with the origins of health problems and in particular problems related to nutrition, environmental dangers and risky behaviors of humans. Generally the epidemiologist gathers information in a community and through analysis of the data seeks to uncover risk factors for health problems, especially those risk factors that could be altered by governmental, medical or educational intervention in the population.

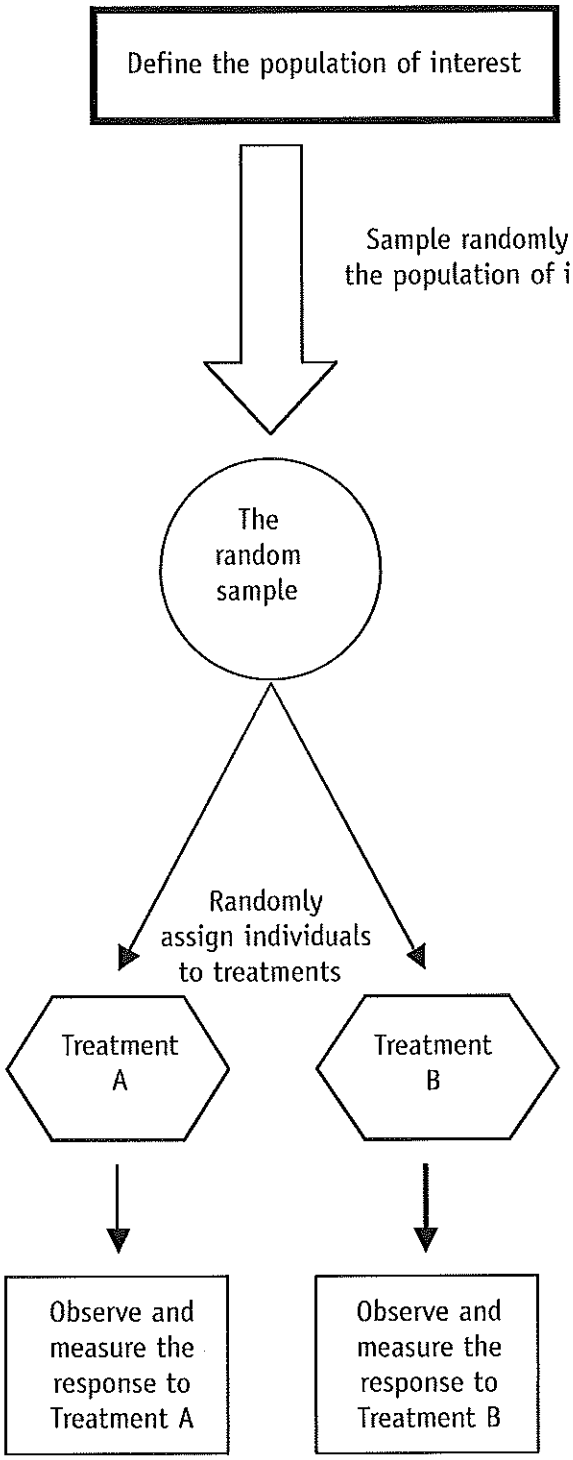
From the traditional scientific point of view the questions epidemiologists ask are elementary: Given that some individuals in a population will come down with a cold and some won't, why is this so? Is catching the common cold a completely random event, or might this malady be prevented or at least minimized by some sort of intervention, such as drinking orange juice? In the customary language of the community, one might ask: Does drinking orange juice prevent or at least minimize the likelihood of catching a cold? In epidemiologic science-speak, the question might be formed like this: Does drinking orange juice protect one from catching a cold? And in the formal logic of science, the question would be as follows: Is the ingestion of orange juice one possible cause of a failure to catch a cold?

In modern science such a question might be answered by performing an experiment. The logic of experimentation is fairly clear, and we will summarize it here. Generally, to determine that X causes Y—in this case that drinking orange juice causes a lack of cold—one would have to demonstrate the following important statements:

1. Drinking orange juice is associated with not getting a cold.
2. Drinking orange juice precedes exposure to the cold-causing virus.
3. The more orange juice one drinks, the less likely one is to get a cold.
4. There is no other variable that is associated with drinking orange juice that could explain the protection against a cold.
5. The mechanism of causation can be explained consistently with accepted science.

The logic of demonstrating causation with an experiment is tied to the above requirements, and there are necessary steps involved in the construction of an experiment. These steps—and their reasons—are diagrammed below.

A Schematic Diagram of the Randomized Controlled Experiment



Defining the population of interest communicates to other scientists what population you are studying. Perhaps orange juice (OJ) doesn't guard against colds for all people.

Sampling randomly is the only way we have for logically asserting that the individuals in our sample are representative of the individuals in the population. Because of this random sampling we can generalize the results from the sample back to the population of interest.

The random sample of individuals is usually a small fraction of the population that will be studied. The randomness of the sampling procedure will tend to make the sample similar to the population and its individual characteristics such as proportion of healthy people, average height and weight, etc. This characteristic is usually termed **representativeness.**

The purpose of random assignment to treatments is perhaps the least understood aspect of doing an experiment. It is usually thought that random assignment to treatments preserves the representativeness of the sample in each of the treatment groups. This is true, but there is another important aspect of randomly assigning subjects to treatments—the destruction of an association

between treatment membership and any other potential cause of protection against getting a cold. That is, there is no other known or unknown factor (for example, risk of exposure to the cold virus) that "rides along" with the OJ. Thus the OJ must be the cause of any observed protective factor.

If there is an effect caused by the OJ, it should show up as a difference in the average response, or the proportion of individuals getting a cold in the two treatments.

Unfortunately for the epidemiologist—and for all of us who would like to lead healthy lives if we only knew how—the complete logical requirements for an experiment are often impossible for the epidemiologist. As can be imagined, this presents extra challenges for them when they try to make sense of observational data with about a hand and a half tied behind their backs. Let's consider some of the limitations imposed on the epidemiologist that are not particular problems in other scientific endeavors.

Ethical Standards Severely Limit Random Assignment to Treatments

Recall that the concern of the epidemiologist is isolating causes of illness or even death in human populations. As an example suppose that legislation were being considered to require graduated drivers' licenses among teenage drivers. Drivers at age 16 might be allowed to drive only during the day and alone. Not until drivers were 17 would they be allowed to drive at night, and not until 18 would they be able to drive anyone other than parents or siblings. From an epidemiologic standpoint, one question might be, is driving at night (as opposed to during the day) a risk factor for death due to traffic accidents? It is clearly not ethical to assign teens to drive at night if that is felt to be a risk to health and life. It would be like assigning some people to smoke two packs of cigarettes every day!

Are high-power lines implicated in causing cancer? Are oral contraceptives causing an increased risk of heart disease? These are questions that cannot be addressed in an experimental situation because of ethical concerns.

This limitation alone is incredibly crippling to the logic of establishing causation. Without random assignment to treatments there is no easy logical way to establish the unlikelihood of competing causes. For example, one might argue that smoking isn't really what causes cancer. It's the pressure of work. People under pressure tend to smoke, but it's the pressure that degrades their nervous system and makes them at increased risk of all sorts of health problems, including cancer. If we were able to randomly assign people to treatments, there would be just as many nervous people in the two-pack-per-day group as in the zero-pack-per-day group, and then nervousness would be eliminated as an alternative explanation for cancer, because it would not be associated, *could* not be associated, with the smoking variable.

Random Sampling Is Very Difficult in the Epidemiologic World

Strictly speaking, random sampling from existing populations is very difficult in the real world and is a goal often not fully complied with. Usually a sample is taken in a manner that can be argued is morally equivalent to random sampling. For the epidemiologist there are added problems that sometimes make it difficult to fully comprehend what population is being studied. It may be relatively easy to study the population of teenage drivers in a particular state. They are required to have and carry driver's licenses, and theoretically at least a list could be made of these young people and their driving habits studied. However, suppose there is an outbreak of a severe skin rash of some sort, one that has not before been identified. Is the cause possibly nutritional? Then perhaps the population of concern—called the **target population**—is that group of individuals who have eaten a specific food. Or might it be the population of individuals who shopped at a certain store? Or might it be those who did not prepare the food according to instructions? Possibly the rash is the result of a bug bite. Is, then, the target population people who live in or near woods? Or is it those individuals who own pets? Are all individuals in a neighborhood exposed to these bugs? Are all individuals susceptible to these rashes or only those who eat lots of carbohydrates? Without knowing what the at-risk population is, it is not very easy to take a sample, and without that sampling strategy, it is very difficult to know to whom the results of an epidemiologic study can be generalized.

Information May Be Faulty

As we shall see below, epidemiologists have basically two observational problems if they wish to show that exposure to a risk factor (X) in some sense causes the impact of interest (Y). Those observational problems are:

1. Was an individual in fact exposed to the risk factor?
2. Has the impact occurred or not?

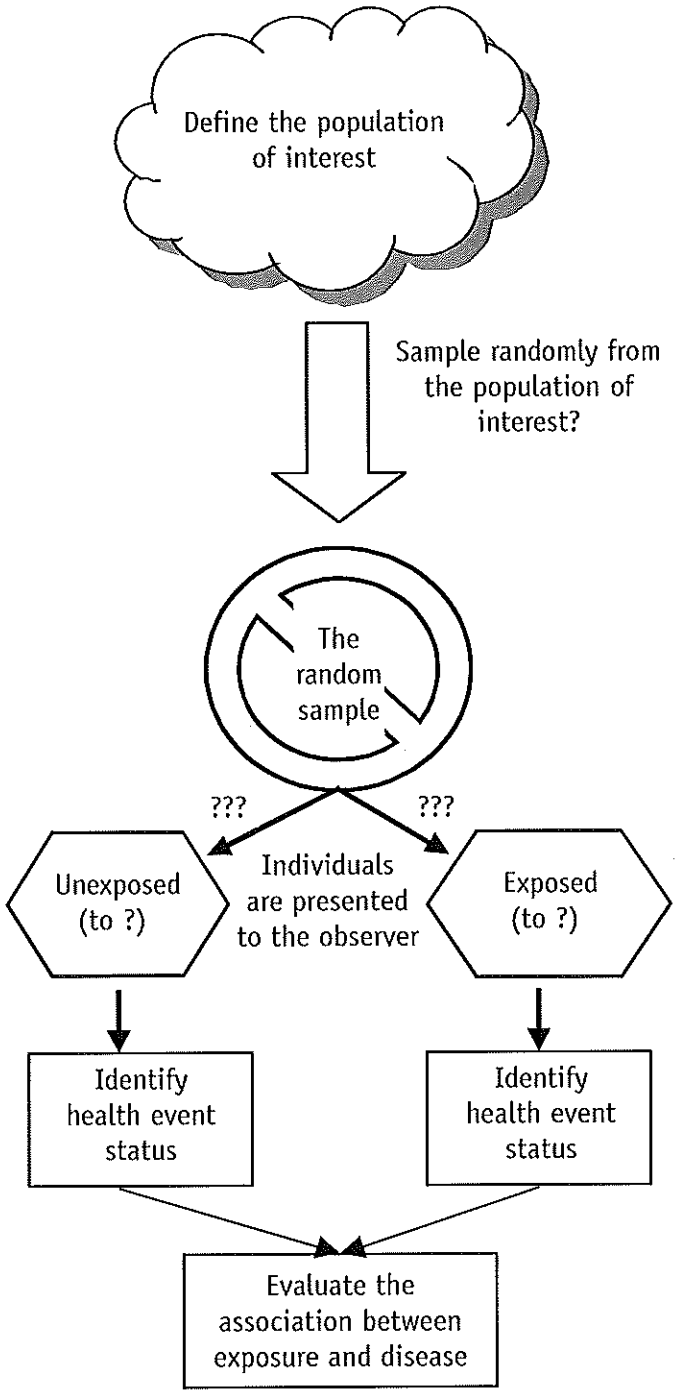
At first blush these observations would seem to be relatively easy. Such is not necessarily the case. Exposure to a risk factor is frequently ascertained only through surveying individuals, individuals with possibly good reasons to provide faulty information (Are you using illegal drugs?) or shade the truth a bit (Are you flossing every day?) about the nature or amount of their exposure. Also in some cases the information sought is in the past, and human memory, always a problem, is the only source of information.

And what about impact? Suppose that we have perfectly reliable information about the use of oral contraceptives by a sample of individuals. We suspect that using oral contraceptives is a risk factor for breast cancer, but how do we know that an individual has breast cancer?

Mammograms are notoriously unreliable in young women, and breast cancer may exist undetected for some time. If a young woman has breast cancer, unknown to anyone, but dies in a traffic accident, her cancer status would be recorded incorrectly as cancer free.

Stripped of the capability of intervention by the investigator and specifically random assignment of subjects to treatments (i.e., the capability to assign exposure), a study is known as an **observational study.** As the name implies, the investigator is on the sidelines, observing events as they unfold. The logic and methods of an observational study are somewhat different from those of an experiment. A diagram of this logic appears below.

A Schematic Diagram of the Remains of a Randomized Controlled Experiment (An Observational Study!)



The population of interest may be well defined, such as young bicycle riders, but rarely is there a reliable list of individuals in the population. When investigating an outbreak of a disease in search of a cause, the population of interest is unclear, at least in the beginning.

Deciding who and where to sample is a serious problem. Epidemiologists must infer a population from studying who is affected and then try to sample from that ill-defined population. Often they do not formally choose a random sample.

Therefore if the sample is not randomly chosen, it is not possible to generalize results to *any* recognizable population, and if a random sample is chosen, it may not be representative of the appropriate population.

Individuals may be identified as exposed or not and then followed up to observe their future disease status. Or individuals with a disease may be identified, and their prior exposure to possible risk factors is sought.

Diagnoses may be difficult to determine and may possibly be in error, or patient records may be incomplete.

Association alone does not imply cause.

The Logic of Observational Studies and the Problem of Bias

As we have already seen, capitalizing on the logic of experimentation will not be possible in most epidemiologic studies. Nevertheless it is still possible to make an argument for cause and effect, though the usual procedures of statistics and laws of probability conferred by randomization cannot be depended on. In the face of a lack of random sampling and random assignment to treatments, epidemiologists must be very careful in interpreting the results of their study, recognizing the potential for error. This means that epidemiologists (as well as all who engage in observational studies) must be on the lookout for problems that lurk in the design and execution of a study and must recognize the great potential for error. One significant class of errors is known as bias. **Bias,** as defined in epidemiology, is an error in design or execution of a study, which produces results that are consistently distorted in one direction because of nonrandom factors. Bias can occur in randomized controlled trials but tends to be a much greater problem in observational studies.

So that we can better understand the nature of bias, recall that the goal of the epidemiologist is to establish that exposure to a particular risk factor is responsible for causing or partially causing a health problem. Greatly simplifying the analytical work of the epidemiologist, we will suggest that observational studies as performed in epidemiology reduce to the following problems:

1. Estimate the proportion of people in a population who are exposed (E) to a risk factor
2. Estimate the proportions of people in a population who, having been exposed to a risk factor or not, subsequently develop a health problem or disease (D).
3. Estimate the association between the risk factor and the disease.

Estimating these proportions in a population is essentially a statistical problem of sampling and a methodologic problem of correct classification of both exposure and disease status. A bias in this context would be any procedure that leads to a systematic over- or underestimation of the association between the risk factor and the disease. Because the association between risk factor and disease is calculated from the proportions, distortion of either of the proportions may lead to a distortion of the estimate of the association.

In the schematic diagram (above) of the observational study, we can see some aspects that are present in a randomized experiment. First, an epidemiologist will have to be concerned about sampling, so that his or her results can be generalized to a target population. Second, correctly ascertaining the exposure status of an individual is analogous to an individual's being in one of two treatment groups: the exposed (E) and the unexposed (not E) treatments.

Third, the effects of being in the exposed or not exposed groups will be compared to assess the effect of being exposed to the risk factor. Epidemiologists are confronted with complex health situations, and therefore their mathematical procedures can be complicated. A wide variety of measures and calculations are used in actual practice, in response to the wide variety of individual situations that may be studied. However, to focus on the logic of observational studies and sources of bias, we will use some elementary formulas. We will suppose that in a target population there are individuals who are exposed to a risk factor and that of those individuals a certain number will subsequently develop the health problem of interest, disease D. In the following tables we distinguish the target population, the sampling frame and the actual sample. The target population is the "real" population, those individuals who are in a population and may be at risk for exposure to a risk factor. The **sampling frame** is a statistical term and conceptually denotes those individuals who are available for sampling.

The target population and the sampling frame really should be in good agreement, but this situation is not necessarily so. One of the most famous polling fiascos in history is thought to have been the result of a sampling frame that differed from the target population. In the 1936 presidential election, a magazine called the *Literary Digest* decided to sample voters to see whom they were supporting. The target population, then, was those 1936 voters. The *Digest* used a sampling frame composed of telephone books and automobile registrations. In 1936 those who owned cars and had telephones were not just different from the population of voters—they were very different! By far a larger proportion of Republicans owned cars and had telephones. Thus the proportion in the sample who reported they were supporting the Republican candidate was seriously distorted from the true value in the target population.

We will conceptualize the target population as consisting of combinations of folks who are exposed or not exposed, and diseased or not diseased. (Remember, epidemiologists are not only concerned with infectious disease but also with health risks in general. The term disease is shorthand for the health risk of interest.) The sampling frame and the actual sample are analogous.

The target population consists of individuals classified according to the following table. The variables α, β, γ and δ denote the proportions of the *population* that are in each group.

The Target Population

Exposure Status	Developed Disease	Did Not Develop the Disease
Exposed	α	β
Not Exposed	γ	δ

The sampling frame consists of *accessible* individuals classified according to the following table. The variables α', β', γ' and δ' denote the proportions of the *accessible* population that are in each group.

The Sampling Frame

Exposure Status	Developed Disease	Did Not Develop the Disease
Exposed	α'	β'
Not Exposed	γ'	δ'

The actual sample results in a certain number of individuals who have been surveyed, examined and diagnosed. The variables, a, b, c and d denote the sample results.

The Sample

Exposure Status	Developed Disease	Did Not Develop the Disease	Total
Exposed	a	b	$a + b$
Not Exposed	c	d	$c + d$

The measure we will use for our discussion of association of exposure to disease is known as the relative risk. In some studies other calculations are used to estimate the relative risk—recall that we are keeping the mathematics simple. Essentially the **relative risk** is a number that compares the risk of disease for an exposed group with the risk of disease for an unexposed group, using a ratio:

$$\text{Relative risk} = \frac{\text{proportion of those exposed who develop the disease}}{\text{proportion of those unexposed who develop the disease}}$$

$$= \frac{a/(a+b)}{c/(c+d)}$$

The relative risk is an easy statistic to interpret. If exposure to the risk factor elevates the probability of getting the disease, the proportion of exposed people who subsequently develop the disease should be greater than the proportion of unexposed people who subsequently develop

the disease. If that is so, the numerator of the relative risk should be greater than the denominator, and the relative risk is therefore greater than 1.0. There is also the possibility that exposure to a factor, such as a vaccination, will decrease the risk. In that case the relative risk would be less than 1.0.

In summary:

1. If the relative risk is equal to 1.0, no association is indicated.
2. If the relative risk is less than 1.0, a risk factor is indicated.
3. If the relative risk is greater than 1.0, a protective factor is indicated.

The relative risk is a sample statistic and is used to estimate the corresponding population relative risk:

$$\text{Relative risk} = \frac{\alpha/(\alpha + \beta)}{\gamma/(\gamma + \delta)}$$

Characteristics of the Relative Risk When Random Sampling . . . and Not

When either experimental or observational studies are undertaken to estimate a characteristic of a population, it is not very likely that a sample result will exactly equal the corresponding population value. In the case of our observational studies it is not very likely that in our samples either

$$a/(a+b) = \alpha/(\alpha+\beta)$$

or

$$a/(c+d) = \gamma/(\gamma+\delta)$$

is true, and therefore it is not very likely that our estimate of the relative risk will be exactly equal to the population value. However, when we randomly sample, the laws of probability guarantee that both proportions are on average equal and thus that the estimate of relative risk will on average equal the population value. The laws of probability also guarantee that even when the sample statistics are not exactly on target, they will be pretty close if our sample sizes are large enough. When a statistic on average lands on the true population value, it is said to be **unbiased.** When this is the case, we are relieved to know that if everything else goes right, we should get sample results that mirror the population. However, there is a slight problem with all this that you will no doubt recall: In an observational study we do not have the opportunity to randomly allocate exposure and we may have difficulty randomly sampling from the population. This presents serious difficulties when epidemiologists attempt to estimate the association between a risk factor and a disease.

We are now in a somewhat uncomfortable position from the standpoint of methodology. If we do an excellent job of random sampling from the target population, it is entirely possible that we will either misclassify the exposure of individuals to a risk factor or misdiagnose the disease status or both. If we are not randomly sampling, it is quite possible that our methods for estimating the association will give incorrect results even if the sample has no errors of classification of exposure or disease status! In each of these situations there is a serious risk of bias in the estimation of the association between exposure and disease.

Generally three types of bias are distinguished in epidemiology: confounding, selection bias and information bias. Confounding is distinguished from selection and information bias in that when it appears, advanced mathematical methods (which, thank goodness, we will *not* get into!) can be used to correct the biased estimates of association between exposure and disease. For selection bias and information bias, however, there is no way to undo the effects. Thus we need to be extra careful at the design and execution stages of an observational study.

Types of Bias

Confounding is a bias that results when the risk factor being studied is so mixed up with other possible risk factors that its single effect is very difficult to distinguish. For example, it might be thought that smoking is a risk factor for heart disease, because people who are exposed to smoking have a higher occurrence of heart disease. However, the case is not quite so clear as it might appear. It turns out that people who smoke also drink alcohol—so is it the smoking, the alcohol or both that are responsible for heart disease? Unless these tangled effects are untangled with advanced mathematical methods (which, remember, we are not getting into), the association between smoking and heart disease, as measured using the relative risk formula we have, is probably too high or too low—that is, it is biased.

Selection bias is a distortion in the estimate of association between risk factor and disease that results from how the subjects are selected for the study. Selection bias could occur because the sampling frame is sufficiently different from the target population or because the sampling procedure cannot be expected to deliver a sample that is a mirror image of the sampling frame.

Information bias is a distortion in the estimate of association between risk factor and disease that is due to systematic measurement error or misclassification of subjects on one or more variables, either risk factor or disease status. It is important to realize that these errors are part of being human and they are not occurring because the physicians or researchers are not being sufficiently careful. It is not so much the random mismeasure or misdiagnosis of an individual that is problematic (although random errors in diagnosis will tend to bias the association toward a relative risk of 1.0, because the true association is diluted with noise). It is the *method* of measurement or classification that is the greater problem, because it systematically exerts an effect on each of the individual measurements in the sample.

Our discussion up to now has been somewhat abstract and a little mathematical. Let's see if we can fill out the discussion with some examples from out in the field of epidemiology.

Selection Bias

Recall that selection bias occurs whenever the manner of selection of study participants creates a deviation between the measurement of the association in the study and the real magnitude of the association between factor and disease in the population.

Nonresponse bias occurs because individuals who do not respond to a call to participate in research studies are generally different from those who do respond. For example, respondents tend to have healthier lifestyle habits, with lower smoking and mortality rates. Because of this they tend to be different from the target population. To illustrate, suppose we would like to conduct a case–control study of the association between liver cancer and smoking. Cases (those identified as having liver cancer) could be all available individuals in all the hospitals in town during the year of the study. Controls (individuals without history of liver cancer) would be recruited by local mass media advertisements—hence they would be volunteers. The study results would most probably show a strong association between smoking and liver cancer, not necessarily because smoking and liver cancer are related, but because the selection process was different for cases and controls. Although the cases were arguably sampled from the population at large, the controls were sampled from a population of volunteers! Several studies have shown that volunteers have lower mortality rates and are less likely to engage in high-risk behaviors such as smoking, when compared with nonvolunteers. The effect in this situation would be to overestimate the numerator and underestimate the denominator in the formula for relative risk because volunteers will have a lower proportion of smokers compared with nonvolunteers, thus biasing the estimate of the relative risk to the high side. (Readers who may be familiar with the methods of epidemiology may question the calculation of the relative risk rather than the odds ratio from a case–control study. However, in this example the total population of the town is known, so the relative risk can be estimated.) Notice that in this example there are two potential sampling biases. First, the target population (everybody in the communities) has been replaced by the sampling frame of those who are reachable by mass media. Second, the two actual samples of cases and controls differ in significant ways, so comparing them leads to complications in the interpretation of relative risk.

Hospital admission rate bias is a selection bias that rears its head when hospital-based studies, especially case–control studies, are undertaken. In **Berkson's bias**, one form of hospital admission bias, the problem is that hospitalized individuals are more likely to suffer from many illnesses, as well as more severe illnesses, and engage in less than healthy behaviors. Thus they are probably not representative of the target population, i.e., the potential patients in the community served by the hospital. In case–control studies, controls are often selected from the same hospital where cases were found. Such controls are conveniently accessible for purposes of the study. To illustrate another form of hospital admission bias, suppose that we would like to assess the association between low socioeconomic status and asthma by using a hospital-based

case–control study. Suppose further that this hospital is located in a low-income area of the city and is famous for its expertise in asthma. Because of that expertise individuals with asthma (cases) from all over the state and region elect to go to that hospital to get care. If the hospital is not as well known for other medical conditions and specialties, and controls are chosen from that hospital, there would be a difference between the two populations sampled. Cases would come from all over the region, and controls would be mostly local low-income individuals. This discrepancy would result in an underestimate of the true association between low income and asthma.

Exclusion bias occurs when in certain circumstances epidemiologic studies exclude participants to prevent confounding. If the exclusion criteria are different for cases and controls or different for the exposed and nonexposed, an exclusion bias may be introduced. In 1974 researchers published the results of a case–control hospital-based study in which breast cancer was associated with the use of reserpine, at that time a popular treatment for high blood pressure. To make the controls more closely resemble the study population, women who had medical conditions that would lead to the prescribed use of reserpine were excluded from the control group. However, the same exclusion criteria were not used for the cases. Therefore, an overestimation of the association between breast cancer and reserpine was found. Investigators replicated the study in 1985 and performed two different analyses of their data. First, they included all women in their analysis. Then they reanalyzed the data after excluding controls with cardiovascular disease. Their findings showed no association when all women were included, but when they excluded the women with cardiovascular diseases from the control group, their data showed a strong association.

Publicity bias (also called **awareness bias**) occurs when media attention is drawn to a particular illness. Thus if a government official or movie star is widely reported to have a particular illness, this stimulates individuals to wonder if they might have the same illness, resulting in an increase in the reporting of the disease. Publicity bias can also occur from news reports not related to individuals. In a 1981–1982 survey of individuals near two hazardous waste disposal sites in Louisiana, people were asked about various symptoms. Air and water quality data showed little evidence of hazardous concentrations of chemicals, but there had been extensive media coverage at the time of the survey. Respondents living near the sites were two to three times as likely to report symptoms as respondents in an unexposed community because of the influence of the publicity at that moment in time.

Information Bias

When the information obtained from study participants is systematically inaccurate regarding the disease or exposure under study, information bias may occur. Whenever the accuracy of the information about exposure is different in cases and controls, a differential information bias occurs. Such bias could result in an over- or underestimate of the real association depending on the circumstances. For example, if exposure is underrecognized in cases, an underestimation of a positive association will occur, but if exposure is underrecognized in controls, an overestimation of a positive association will occur. There are several forms of information bias; we will present the most common ones.

Medical records are often used in epidemiologic studies to **abstract data.** However, careful consideration should be given to the quality of the data because medical records are made for diagnostic and treatment purposes, not for research. For example, data are usually more complete when a clear diagnosis has been established. Better quality on the diagnosis is usually achieved in patients with severe disease. Therefore more complete information about exposures would be more frequently found in patients with severe disease. This difference in the accuracy of medical records will tend to produce an overestimation of the association under study.

Biases also occur when **interviewing.** Interviewers tend to be more accurate when interviewing cases than when interviewing controls. When interviewers have knowledge that a respondent is a "case," they will tend to assertively find exposure, as well as classify vague or indeterminate responses as indicating exposure. This more precise questioning improves the quality of the data from cases. However, since the precise questioning tends not to happen with a control respondent, the information about exposure in controls will tend to be underestimated. Thus an overestimation of the association will be the consequence. (Note: This also happens with abstracting bias if one thinks of the records as analogous to respondents.)

Recall bias is one very common form of information bias. Cases (individuals identified as having the disease under study) tend to better recall past exposures than controls. For example, women who have had a baby with a malformation will remember better any events during pregnancy than mothers of infants with no malformations. It seems probable that this is true because individuals with a disease are more concerned about remembering potential causes. Therefore recall bias will tend to overestimate the association of the outcome with exposure to a risk factor.

Reporting bias occurs when a case emphasizes the importance of exposures that he or she believes to be important. Sometimes this report bias may be related to occupational exposures that the patient wants to underscore as a result of worker's compensation or any other benefit, thus producing an overestimation of the association under study as it does not occur in controls.

Conclusion

Our discussion and examples above have shown that there are many possible sources for error that can result in systematic distortions of study results. These distortions are a problem especially when the epidemiologist is estimating the association between a risk factor and a health problem. Whether a risk factor or a protective factor goes undetected, or a behavior or condition is misidentified as a risk or protective factor, the implications can be serious for the public. A risk factor that goes unidentified is one about which information cannot be used to alter the public's behavior and will result in sickness or death for individuals. An erroneously identified risk factor may cause unneeded pain and worry among the public or perhaps an unnecessary diversion of research funds. Epidemiologists conducting observational studies (cohort, cross-sectional and especially case-control) need to be aware of the potential for biases and exert extra care to eliminate or lessen their effect. As interpreters of studies we members of the public need to be aware of the possible biases in such studies when we evaluate their conclusions as reported by the mass media.