Concepts in Cancer Epidemiology and Etiology

PAGONA LAGIOU, DIMITRIOS TRICHOPOULOS, AND HANS-OLOV ADAMI

Epidemiology has been a powerful tool in the identification of causes of infectious diseases and the elucidation of the conditions underlying epidemic outbreaks that are frequently, but not always, of infectious etiology. Around the middle of the twentieth century, first in the United Kingdom (Doll and Hill, 1950) and later in the United States and the rest of the world (Wynder and Graham, 1950; Clemmesen and Nielsen, 1957; MacMahon, 1957), epidemiology expanded in scope by focusing also on the etiology of chronic diseases, irrespective of the nature of the causal agents. Since then, epidemiology has developed and matured to become a rich and powerful toolbox for the study of biologic phenomena in humans. With a number of fine textbooks nowadays available to students of epidemiology (for instance Miettinen, 1985; Hennekens and Buring, 1987; Walker, 1991; MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998; Rothman, 2002; and several others), this chapter is not intended to expand on methods or quantitative considerations. For the purpose of better understanding the logic underlying cancer epidemiology, however, central concepts in epidemiology—the study of disease etiology—will be reviewed. We examine cohort and case-control studies (with special reference to studies of genetic epidemiology), we consider the impact of chance and systematic errors (confounding and bias), and we trace the process of causal reasoning. Familiarity with these concepts is essential for critical reading and understanding of the chapters on specific cancers. A glossary found at the end of the chapter provides a summary of definitions for words in italics.

ETIOLOGY

Causality
The definition of a cause should apply to all diseases, whether defined on the basis of a particular exposure, such as many infectious and occupational diseases, or documented by a constellation of clinical and/or laboratory findings—for example, malignant...
tumors, connective tissue disorders, or psychoses. In terms of a particular individual, exposure to a cause of a disease implies that the individual is now more likely to develop the disease, although there is no certainty that this will happen. The complexity of biological phenomena and our ignorance or limited understanding of many of the underlying processes hinder a deterministic, logically unassailable, explanation of disease causation. Hence, causation of disease can only be conceptualized in a probabilistic (stochastic) sense that involves statistical terms and procedures. For instance, while heavy smokers are much more likely to develop lung cancer than nonsmokers, most smokers never develop lung cancer and some nonsmokers do.

In epidemiology, there are several models of causality that have been applied to help clarify the role of various exposures in the etiology of disease. The causal pies presented by Rothman (1976) provide perhaps the most coherent approach to conceptualizing causality in a variety of epidemiologic settings (Rothman, 1986). Each of these pies describes a set of exposures that work together on the same pathway to cause disease (Fig. 6–1). Different exposures may occur within a short time span, or may happen decades apart. Once every exposure in a causal pie has occurred, that is the pie is complete, disease is, in a deterministic context, inevitable. Table 6–1 provides a summary of the attributes of the causal pie model.

Causality is rarely, if ever, characterized by a simple one-to-one correspondence between a particular exposure and a specific disease. If so, the presence of the exposure would be both necessary and sufficient for the occurrence of the disease. By necessary we mean that the disease cannot occur without the presence of that exposure (although other exposures may be required for the occurrence of the disease). By sufficient we mean a set of exposures that inevitably produce disease. There may of course be different ways by which one could get disease, and thus sufficient causes may not be necessary.

In cancer epidemiology, the only known examples of exposures that are sufficient to cause disease refer to the genetic origin of some rare cancers due to dominant genes with complete penetrance. In this instance, the causal pie would require only one factor for the pie to be complete and this would be the way that carriers would get the specific cancer. Also rare is the existence of single factors that are in and by themselves sufficient (although not necessary) for the causation of a certain disease. Even powerful exogenous factors, such as life-long heavy smoking, cause cancer in very rare cases. In individual cases, certain exposures may be necessary, so necessary means cancer has to be present, the absence of which will lead to cancer. Malignant tumors of the cervix are an example of this, whereas, for example, cancers of the bladder and lung are not.

Figure 6.1. The causal pie model describes a set of exposures that work together in the same pathway to cause disease. These are hypothesized ways in which a series of exposures could interact biologically over time to cause disease. This figure provides an example of sufficient causes from cancer epidemiology. Tobacco is an established component cause in many cases of oral cancer. However, tobacco use by itself is not enough for the disease to occur; in addition, oral cancer can occur among people who have never used tobacco. In a given causal pie, the complementary exposures can occur simultaneously, or many years apart. If even one of the component causes did not occur, disease would be prevented by this pathway, although a person could develop the disease by another mechanism (a different causal pie).
CONCEPTS IN CANCER EPIDEMIOLOGY AND ETIOLOGY

Table 6-1. Attributes of the causal pie

<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inevitability</td>
<td>Completion of a sufficient cause (causal pie) is synonymous with eventual occurrence</td>
</tr>
<tr>
<td></td>
<td>(though not necessarily diagnosis) of the disease.</td>
</tr>
<tr>
<td>Causality</td>
<td>A component cause (piece of a causal pie) can involve presence of a detrimental</td>
</tr>
<tr>
<td></td>
<td>exposure or absence of a preventive exposure.</td>
</tr>
<tr>
<td>Burden of disease</td>
<td>The amount of disease caused by a sufficient cause depends on the prevalence</td>
</tr>
<tr>
<td></td>
<td>of all complementary component causes.</td>
</tr>
<tr>
<td>Temporality</td>
<td>Component causes can act far apart in time</td>
</tr>
<tr>
<td>Interaction</td>
<td>Component causes in the same pie interact biologically to cause disease</td>
</tr>
<tr>
<td>Attributable fraction</td>
<td>Different component causes are responsible for more than 100 percent of disease cases.</td>
</tr>
<tr>
<td>Disease prevention</td>
<td>Blocking the action of any component cause prevents completion of the respective</td>
</tr>
<tr>
<td></td>
<td>sufficient cause and therefore prevents disease by that pathway.</td>
</tr>
</tbody>
</table>


Disease is, in a determinate. Table 6–1 provides the attributes of the if ever, characterized e correspondence between disease and a specific exposure and a specific exposure necessary for disease. By necessary disease cannot occur of that exposure (alleles may be required for disease). By sufficient osures that inevitably re may of course be ich one could get distinct causes may not be

logically, the only known s that are sufficient to the genetic origin of te to dominant genes. In this instance, require only one factor te and this would be would get the specific he existence of single d by themselves sufficient for the causality. Even powerful ich as life-long heavy smoking, and strong genetic influences, like those conveyed by dominant breast cancer genes, do not always cause disease in an individual.

Certain exposures are by definition necessary (although not sufficient) for the occurrence of a particular disease. For example, chronic lead disease cannot occur in the absence of lead exposure, and a motor vehicle injury requires the involvement of a motor vehicle (MacMahon et al., 1960; Hill, 1965; Rothman, 1976; Susser, 1991; MacMahon and Trichopoulos, 1996). Again, while these represent necessary causes, there are additional cofactors that must work in concert before disease is inevitable. Most human cancers can occur via several pathways, so it is hard to define any single necessary cause. Asbestos, in relation to mesothelioma (cancer of the pleura), and human papillomavirus infection, in relation to cervical squamous cell cancer, are close to being necessary. However, cases of these cancers do arise without the exposure being documentable, either because the exposure occurred but could not be identified, or because these exposures are not necessary for all cases.

For most diseases, there is no one necessary cause. Indeed there may be numerous causal pies by which disease can occur. Such an example is illustrated in Figure 6–1, with suggested sufficient causes of oral cancer. In the first example, exposure to tobacco and alcohol over time are contributing factors (component causes) in the etiology of oral cancer. However, the oral cancer would not have occurred in the presence of a dental visit that could have treated precancerous lesions and might have prevented the disease. While smoking is a component cause in many causal pies for oral cancer, people can get oral cancer without smoking, as shown by the second causal pie in this figure.

Interventional Epidemiology

How do we design a scientific study to evaluate whether a particular exposure (for example, asbestos) is a cause of a specific disease (for example, lung cancer)? To understand the most appropriate design in practice, it is useful to begin by describing the ideal scientific study. Imagine for a moment that we have access to a time machine.

In an imaginary study, we follow a group of individuals from birth to death, where everyone is exposed to asbestos, and we observe whether they develop lung cancer. We then send everyone back in a time machine, to live the exact same lives they lived, except that we completely remove asbestos from the environment so that no one is exposed. We then compare whether there are changes in the frequency of occurrence of lung cancer before and after use of the time machine. Since the same people live identical lives but for the presence/absence of asbestos, any difference in the frequency

1. smoking, alcohol
2. exposure, and some specific
3. genetic influences
4. disease.
5. In this instance, require only one factor case, and this would be would get the specific existence of single d by themselves sufficient for the causality. Even powerful such as life-long heavy smoking, and strong genetic influences, like those conveyed by dominant breast cancer genes, do not always cause disease in an individual.

Certain exposures are by definition necessary (although not sufficient) for the occurrence of a particular disease. For example, chronic lead disease cannot occur in the absence of lead exposure, and a motor vehicle injury requires the involvement of a motor vehicle (MacMahon et al., 1960; Hill, 1965; Rothman, 1976; Susser, 1991; MacMahon and Trichopoulos, 1996). Again, while these represent necessary causes, there are additional cofactors that must work in concert before disease is inevitable. Most human cancers can occur via several pathways, so it is hard to define any single necessary cause. Asbestos, in relation to mesothelioma (cancer of the pleura), and human papillomavirus infection, in relation to cervical squamous cell cancer, are close to being necessary. However, cases of these cancers do arise without the exposure being documentable, either because the exposure occurred but could not be identified, or because these exposures are not necessary for all cases.

For most diseases, there is no one necessary cause. Indeed there may be numerous causal pies by which disease can occur. Such an example is illustrated in Figure 6–1, with suggested sufficient causes of oral cancer. In the first example, exposure to tobacco and alcohol over time are contributing factors (component causes) in the etiology of oral cancer. However, the oral cancer would not have occurred in the presence of a dental visit that could have treated precancerous lesions and might have prevented the disease. While smoking is a component cause in many causal pies for oral cancer, people can get oral cancer without smoking, as shown by the second causal pie in this figure.

Interventional Epidemiology

How do we design a scientific study to evaluate whether a particular exposure (for example, asbestos) is a cause of a specific disease (for example, lung cancer)? To understand the most appropriate design in practice, it is useful to begin by describing the ideal scientific study. Imagine for a moment that we have access to a time machine.

In an imaginary study, we follow a group of individuals from birth to death, where everyone is exposed to asbestos, and we observe whether they develop lung cancer. We then send everyone back in a time machine, to live the exact same lives they lived, except that we completely remove asbestos from the environment so that no one is exposed. We then compare whether there are changes in the frequency of occurrence of lung cancer before and after use of the time machine. Since the same people live identical lives but for the presence/absence of asbestos, any difference in the frequency
may be attributed to alterations in the exposure to asbestos, which leads to the definition of cause.

How then can we develop the time machine analogy into a realistic epidemiologic approach? We could study two groups of people who are comparable on every characteristic, except that one group had exposure and one did not. The randomized controlled trial closely approximates this goal. By randomly allocating who receives an exposure, for example treatment, and who does not, the exposure occurs only because the investigator has assigned it. For example, an investigator randomly assigns one group of people to receive vitamin E supplements (exposed), while the other group receives a placebo (unexposed). Study participants are then followed forward in time to see whether they develop cancer. Whether someone receives vitamin E then does not depend on whether or not the subject, for example, smokes, drinks, eats a high-fat diet, or has a certain genetic susceptibility.

In this way, the randomization in a trial makes the two (or more) groups, those exposed and those unexposed, comparable on other study factors that might cause the disease. Hence, the unexposed group is a proxy of what would have happened to the exposed group if they had been unexposed—that is if we could have sent them back in time. Comparability is essential in order to ascribe any changes in the frequency of disease to alterations in the exposure.

While some researchers describe the randomized controlled trial as the gold-standard of scientific studies, this design is impractical in the majority of epidemiologic situations. For one thing, most exposures we study are detrimental. If we want to study the impact of asbestos on lung cancer, we cannot ethically randomize people to live in a house with asbestos. But even for exposures that are not necessarily detrimental, randomization may be difficult or impractical. For instance, it is very difficult and expensive to randomize a large group to eat a low-fat versus a normal diet, and have everyone comply with this allocation over the course of many years. Most trials are thus only conducted for no more than a few years, an unrealistically short period to test the effect of most exposures because of the long latency between exposure and diagnosis of cancer. Furthermore, in many randomized trials, subjects become noncompliant over time—that is people allocated to the intervention arm stop taking the intervention, and those in the original placebo or usual care arm may adopt the intervention (a phenomenon called cross-over). This diminishes the contrast between the original randomized groups, reducing the power to detect a difference in disease rates between the groups.

Because of the limitations of the randomized controlled trial, the observational cohort and case–control designs are extensively utilized in epidemiology. As will be discussed later in the chapter, attention to both the design and analysis of these studies may allow us to approximate the standards of comparability, necessary to validly evaluate the effect of an exposure on the frequency of a disease.

Observational Epidemiology

The essence of observational epidemiology is the noninterventional investigation of disease causation in human population groups. The argument is that only by studying humans is it possible to draw confident conclusions about normal or pathological processes concerning humans (MacMahon, 1979; MacMahon and Trichopoulos, 1996). In vitro studies, such as those involving cell cultures, and studies in laboratory animals are valuable. They are indeed indispensable when toxic exposures or invasive procedures like repeated biopsies are needed for the study of physiologic or pathologic processes, such as carcinogenesis. However, in vitro systems are frequently artificial, and there are physiologic and metabolic differences between humans and laboratory animals that hinder interspecies analogies. These analogies are further complicated by the unavoidably limited number of animals used in laboratory studies and the relatively short life span of these which impose the adn doses of suspected agent a sufficient number sequently, questionable olation to humans hav.

Even when experimental randomized control ethical, they are, with practical because most their latent period, that exposure to a cause and clinical disease, is long, essay to enroll unreal bers of compliant vol period (Hennekens a MacMahon, 1979; M chopoulous, 1996).

Observational, that i studies represent the ma epidemiology. Such stu ment causal relations o cations between partic cancer or other disease sation on the basis of when the association biologically credible cancer, or hepatitis B, are striking exampl difficult when the associ compelling but the ep erence weak—for exampl level ionizing radiati passive smoking and l interpretation also be when the epidemiologic convincing but the bi uncertain, as it is with and colorectal cancer o cancer. When an epider is weak, is derived fro tionable quality, and fl vacuum, inferring caus: CON

STUDY D

Descriptive Studies

It is possible to distin epidemiological studies analytic. In descriptive s
short life span of these animals, both of which impose the administration of high doses of suspected agents in order to generate a sufficient number of outcomes. Consequently, questionable quantitative extrapolations to humans have to be undertaken.

Even when experimental studies, such as randomized controlled trials, in humans are ethical, they are, with few exceptions, impractical because most diseases are rare and their latent period, that is, the time between exposure to a cause and the appearance of a clinical disease, is long. This makes it necessary to enroll unrealistically large numbers of compliant volunteers for a very long period (Hennekens and Buring, 1987; MacMahon, 1979; MacMahon and Trichopoulos, 1996).

Observational, that is nonexperimental, studies represent the mainstream of modern epidemiology. Such studies seek to document causal relations on the basis of associations between particular exposures and cancer or other diseases. Inference of causation on the basis of association is easy when the association is both strong and biologically credible—smoking and lung cancer, or hepatitis B virus and liver cancer, are striking examples. It becomes more difficult when the association is biologically weak—for example, in studies of low-level ionizing radiation and leukemia or passive smoking and lung cancer. Causal interpretation also becomes problematic when the epidemiologic association is fairly convincing but the biological rationale is uncertain, as it is with respect to red meat and colorectal cancer or alcohol and breast cancer. When an epidemiologic association is weak, is derived from a study with questionable quality, and floats in a biological vacuum, inferring causation is perilous.

STUDY DESIGN

Descriptive Studies

It is possible to distinguish observational epidemiological studies into descriptive and analytic. In descriptive studies the frequency of occurrence of a disease (incidence)—or of death from a disease (mortality)—is estimated in a population, by routinely available time, place, and/or group characteristics. Descriptive studies are essentially exploratory and hypothesis generating. For instance, descriptive studies that documented the increasing trend of lung cancer incidence among men, but not among women, in the early part of the twentieth century pointed to tobacco smoking as a likely cause of this disease. In contrast, the objective of analytic studies is to document causation from the pattern of association in individuals between one or more exposures on the one hand, and a particular disease on the other.

Ecologic Studies

Ecologic studies in epidemiology occupy an intermediate position between descriptive and analytic investigations, in that they share many characteristics with descriptive studies, but serve etiologic objectives. In ecologic studies, the exposure and the disease under investigation are ascertained not for individuals but for groups or even whole populations (Morgenstern, 1982). Thus the prevalence of hepatitis B virus (HBV) in several populations could be correlated with the incidence of liver cancer in these populations, even though no information could be obtained as to whether any particular individual in these populations was or was not an HBV carrier and has or has not developed liver cancer. Associations from ecologic studies are viewed with skepticism, because these studies are susceptible to unidentifiable and intractable confounding as well as to several other forms of bias (Morgenstern, 1982; Greenland and Robins, 1994).

When an exposure is fairly common, for example, smoking, or even prevalence of HBV carriers, ecologic studies can provide useful evidence on the possible effects of these exposures. For instance, following the increase in tobacco consumption, the incidence of lung cancer increased sharply over time, and the incidence of primary liver cancer is higher in populations with higher...
prevalence of HBV. As a corollary, lack of an association in ecologic studies between a widespread exposure and the incidence of a disease allegedly caused by this exposure, does not support a strong causal relation.

Analytic Studies
Analytic epidemiologic investigations ascertain exposure and disease outcome in individuals and are usually distinguished into cohort and case–control studies, although there are also several variants of these prototype designs (MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998). The objective of analytic epidemiologic studies is to ascertain whether a particular exposure, such as a physical, chemical, or biological agent, and a specific cancer or other disease are unrelated (independent) or associated. An association does not necessarily indicate causation. Chance, bias, and confounding (see following) can also generate associations, and they frequently do. Causation is unlikely when there is no association observed. Even if a causal relation does exist, however, it may sometimes be difficult to document it, particularly when the association is weak, the study has limited statistical power, or the exposure is seriously misclassified.

Person-time and Study Base
The concepts of person-time and study base are fundamental to the design and analysis of epidemiologic studies. As the name implies, there are two key components in our description of the person-time, namely the number of people and the time they are followed. To illustrate this, we could ask how many brain cancer cases we would expect if we followed one million people exposed to x-rays for zero seconds. Conversely, how many cases would we expect if we followed zero people for one million years? The answer in both instances is, of course, zero. Hence, neither people nor time alone provides adequate information about the disease experience of a population, and thus both should be taken into account. Person-time is the sum of all the time contributed in a study by subjects at risk of a disease.

Theoretically, an ambitious investigator might wish to include the entire world population in an epidemiologic study during many decades. Needless to say, such a study would provide marvelous opportunities to evaluate many different exposures in relation to many diseases. Millions of person-years would be generated even within a few weeks. In real life, however, any investigator has to restrict the person-time from which information is harvested. This specified person-time is called the study base. Defining the person-time to be included in the study base may include geographic restrictions, defined time periods, and certain age limits. Personal characteristics such as gender, ethnicity, and occupation may further specify the study base. For example, the study base may be comprised of all British doctors who answered a questionnaire in 1951 (Doll and Hill, 1956), or by all Swedish women who were aged 50 to 74 between 1994 and 1995 (Weiderpass et al, 1999), and who generated person-time until they died or until the follow-up was completed.

Thus, the study base is simply the person-time of a population of individuals at risk of a disease under study. Defining the study base is a crucial step in the design and conduct of an epidemiologic study. There are three central considerations. One is to accommodate realistic goals with regards to feasibility and resources, as certainly no investigator is independent of time and money. A second goal is to make the study efficient. For example, it would make little sense to study the association between smoking and cancer in a population where very few are smokers. Likewise, a study of diet and prostate cancer would be inefficient among men younger than 40, since virtually no cases arise among such young people. The final challenge is to identify a study base that allows valid inferences concerning associations between exposure(s) and a particular disease—that is, a study base that does not impose intractable confounding or raise insurmountable obstacles of other biases.
Person-time is the source of any event we want to investigate, for example the occurrence of cancer. To help set a foundation for better understanding person-time in a study base, we will use the example of x-rays and risk of brain cancer (Fig. 6-2). In this population, five people have been exposed to x-rays, and another five have not been exposed and remain unexposed during the study period. While in real life the study populations are much larger, we use this elementary example to illustrate the principles.

Among the people exposed to x-rays, persons 3 and 5 were followed from the time they were exposed to x-rays till the end of the study period—a total of 5 years each. Persons 1, 2, and 4, however, developed brain cancer at the end of years 1, 4, and 2, respectively. Once these individuals develop brain cancer, they are no longer at risk of the disease, and thus no longer contribute information to the study base. The person-time among those exposed to x-rays is estimated by summing up the person-time of all the individuals while at risk for the disease, that is:

\[(2 \text{ persons} \times 5 \text{ years}) + (1 \text{ person} \times 1 \text{ year}) + (1 \text{ person} \times 4 \text{ years}) + (1 \text{ person} \times 2 \text{ years}) = 17 \text{ person-years}\]

We can similarly sum up the person-time among the group of five individuals who were not exposed to x-rays:

\[(4 \text{ persons} \times 5 \text{ years}) + (1 \text{ person} \times 2 \text{ years}) = 22 \text{ person-years}\]

Later on when we discuss analysis of epidemiologic studies, we will see how the
person-time data will help us to compare disease incidence between exposed and unexposed people.

Cohort studies

The word *cohort* derives from the similar Latin word, which identified one of the ten divisions in a Roman legion. In epidemiology, cohorts are groups of individuals, which can be followed over time. In cohort studies, individuals are classified according to their exposure and are observed for ascertainment of the frequency of disease occurrence or death in the various exposure-defined categories (Fig. 6-3A). In each category the frequency of occurrence is calculated either as risk or as incidence rate. Risk describes the number of those who developed the disease under study among all individuals in this category. Rate describes the frequency of occurrence among those who developed the disease divided by the person-time during which the individuals in this category have been under observation. Cohort studies have the following defining characteristics.

*Cohort studies are exposure-based.* The groups to be studied are selected on the basis of exposure. In special exposure cohorts, the groups are chosen on the basis of a particular exposure. In general population cohorts, groups offering logistical advantages for follow-up are initially chosen and the individuals are classified according to their exposure status. Special exposure cohorts may be necessary when rare exposures need to be studied, such as those encountered in the occupational setting. For example, to study efficiently the effect of vinyl chloride on liver angiosarcoma, or aromatic amines on bladder cancer, epidemiologic studies have been conducted in cohorts of workers in the plastic and dyestuff manufacturing industries, respectively.

The general population cohort is appropriate when the exposure under consideration is fairly common. Classical examples of general population cohorts, in which the profession facilitated accessibility of cohort members rather than being a study factor, include the British Doctors Study and the Nurses Health Study. The British Doctors cohort, established in 1951, consisted of more than 30 000 doctors from Great Britain. In this landmark study, Doll and colleagues prospectively followed the cohort and collected updated information on multiple exposures, particularly smoking, over several decades. Indeed, prospective data from the British doctors were among the first to demonstrate convincingly the role of tobacco in the etiology of lung cancer (Doll and Hill, 1956). More than four decades later, data from the British Doctors have continued to provide insight into the etiology of cancer (Doll et al, 2005).

Another notable cohort is the Nurses Health Study, which began in 1976 with over 120 000 US registered nurses. This cohort was assembled initially to evaluate prospective the effect of oral contraceptives on the risk of breast cancer (Hennekens et al, 1984). Subsequently, diet and many other exposures have been studied in relation to the risk of cancer as well as other chronic conditions (Zhang et al, 2003). Information on these diverse exposures has been collected biennially through questionnaires. Moreover, blood samples have allowed researchers to explore biomarkers and genetic factors. For example, prospective data from the Nurses Health Study has provided insight into the role of both exogenous and endogenous estrogens in breast cancer etiology. A particular characteristic of these types of cohorts is that the individuals can be followed almost completely over time, due to their membership in groups with a high interest in health studies and registration requirements that facilitate initial contact and long-term follow-up.

*Cohort studies are patenty or conceptually longitudinal.* The study groups are observed over a period of time to determine the frequency of disease occurrence among them. The distinction between retrospective and prospective cohort studies depends on whether the cases of disease occurred in the cohort at the time the study began. In a retrospective cohort study, exposures and health outcomes occurred before the investigation started. These are typically assembled from pre-existing records of a population over time—employment histories of linked to recorded he- mation of the worker cohort study, the rele may not have acted anc cerntainly have not ye- following identification the investigator must w appear among cohort r Methodologically, th cohort studies: closed c open or dynamic coh- are frequent in occupa and the study of outbr eohorts dominate canci form the conceptual b con- control studies. The key open and closed coh}
in 1951, consisted of doctors from Great mark study, Doll and vely followed the co- updated information on particularly smoking.

Indeed, prospective h doctors were among trate convincingly the etiology of lung can-

More than four om the British Doctors rovide insight into the Doll et al, 2005).

cohort is the Nurses h began in 1976 with stered nurses. This co-

initially to evaluate pro-

oral contraceptives ancer (Hennekens et al, , diet and many other studied in relation to ; as well as other chronic al, 2005). Information posures has been col-

rough questionnaires. niples have allowed re-

biomarkers and genetic , prospective data from study has provided in-

of both exogenous and is in breast cancer eti-

characteristic of these at the individuals can be
dpleted over time, due in groups with a high
studies and registration acilitate initial contact

patently or conceptu-
estudy groups are ob-
d of time to determine ease occurrence among on between retrospec-

as the study began. In \textit{rt study, exposures and occurred before the in}. These are typically as-

population over time—for example, the em-

ployment histories of a factory can be
linked to recorded health-outcome informa-
tion of the workers. In a prospective
cohort study, the relevant causes may or may not have acted and the cases of disease certainly have not yet occurred. Hence, following identification of the study cohort, the investigator must wait for the disease to appear among cohort members.

Methodologically, there are two types of cohort studies: closed or fixed cohorts, and open or dynamic cohorts. \textit{Closed cohorts} are frequent in occupational epidemiology and the study of outbreaks, whereas \textit{open cohorts} dominate cancer epidemiology and form the conceptual basis for most case-control studies. The key distinction between \textit{open} and \textit{closed} cohorts is how member-

ship in the cohort is determined. In a closed cohort, it is determined by a membership-defining event that occurs at a point in time. For example, people who were living in Hiroshima and Nagasaki when the atomic bombs were dropped in 1945 are part of a cohort whose membership began on the date of the bombing. These subjects remain in the cohort until they die.

Open cohorts are composed of individuals who contribute person-time to the cohort only as long as they meet the criteria for a membership-defining state (Fig. 6-3A). Examples of such criteria include place of residence, age, and health status. Once individuals can no longer be characterized by the defining state(s), they cease to contribute person-time to the open cohort and are no longer members. Open cohorts are

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.3a.png}
\caption{A cohort study comprises individuals who are either exposed or unexposed to the factor(s) of interest. When these people are followed over time, they generate person time. Newly diagnosed cases of a particular disease, that occur while person-time is accumulated are recorded. The exposure status of a person can change. A person could be smoking high tar cigarettes for five years, then switch to light cigarettes for fifteen years, and then quit. Consequently, each person can contribute to person-time in different exposure groups. A case is considered exposed, if the disease occurred when the person who developed the disease was accumulating exposed person-time. A case is non-exposed if it occurred while the person was accumulating non-exposed person time. The example assumes, for simplicity, zero latency. The total amount of exposed and non-exposed person time and the number of exposed and non-exposed cases can be calculated. After that, one can determine whether more cases occurred in the exposed or non-exposed group per unit of person-time, that is, one can calculate the incidence rate ratio. This ratio will indicate whether there is a relationship between the exposure and the disease of interest.}
\end{figure}
used, for example, in cancer epidemiology studies based on registry data (Hansson et al, 1996). A person could be a member of the cohort, for example, only as long as he or she was a resident of Sweden and was not diagnosed with the cancer under study. If the person emigrated from Sweden to another country, he or she stopped contributing person-time to the cohort at that time. Similarly, if someone born outside of Sweden immigrates there later in life, he or she would begin contributing person-time to the cohort at that time. In studies based on open cohorts it is not possible to directly measure risk, otherwise referred to as cumulative incidence. Analyses are based on person-time using incidence rate measures.

As an example, assume that in a closed cohort among 5000 nonsmoking men followed for an average period of 10 years (\(P_0 = 50,000\) person-years), \(x_0 = 25\) were diagnosed with lung cancer, and among 10,000 smoking men followed for an average period of 8 years (\(P_1 = 80,000\) person-years), \(x_1 = 600\) were diagnosed with lung cancer. In this example the incidence rate among nonexposed would then be \(50 \text{ per 10}^5 \text{ person-years}\) and among exposed \(750 \text{ per 10}^5 \text{ person-years}\). The relative risk (incidence rate ratio) would be \(25/750 = 0.033\) or 15. The conclusion is that there is a 15-fold increase in lung cancer occurrence from smoking.

**Case-control Studies**

In case-control studies, patients diagnosed with the disease under consideration form the case series. As in cohort studies, their exposure to the factor under investigation is ascertained, for example, through questionnaires, interviews, examination of records, undertaking of laboratory tests in biological samples, and other means (Fig. 6.3B). Using the same methods, the pattern of exposure to the study factor(s) is then estimated in the population, or more strictly in the person-time from which the case series arose. This is done among control subjects selected as a sample of the study base from which the cases arose. If only two categories of exposure are relevant (exposed and unexposed), the relative risk can be estimated by dividing the odds of exposure among cases with the corresponding odds among the controls, the odds ratio.

Thus, if among 200 male patients diagnosed with lung cancer (cases), \(a = 150\) were smokers and \(b = 50\) nonsmokers, whereas among 300 men similar in age to the cases but without lung cancer (controls), \(c = 50\) were smokers and \(d = 250\) were nonsmokers, the odds ratio would be \(150/50 = 3\). This measure is a good approximation to the relative risk (or risk ratio, or rate ratio). Hence, similar to the cohort study, these data from a case-control study show a 15-fold excess of lung cancer among smokers.

There are some features of case-control studies that make this design susceptible to bias (see following). A well-designed case-control study, however, is a valid and cost-efficient approach to the study of the etiology of cancer and other conditions.

**Nested case-control studies**

Some case-control designs are methodologically superior to others. The best example is the nested case-control design. The definition of this study design is still somewhat ambiguous (Walker, 1991; Rothman and Greenland, 1998). A definite requirement, however, is that controls are chosen from the clearly defined person-time from which all cases have arisen. In other words, if one of the controls had developed the disease under study, he or she would have definitely been included among the cases. Defining the underlying person-time from which a series of cases—for example, lung cancer cases presenting at a referral hospital—arose can be difficult. Sampling controls from a cohort different from the one that gave rise to the cases often results in selection bias.

According to a more strict definition, the term nested case-control study is used only when the underlying cohort and the corresponding person-time have been previously enumerated and the exposure information was collected prior to the diagnosis. In other words, the controls are selected from exactly the same person-time that gave rise to the cases, the study base being due to selective differences in recall, the validity of a study. Case-control studies are used to bias due to selective differences in recall, the serves the validity of a study. Case-control studies are being cost efficiency when an occupational cohort case defined whereas abstract exposure information f requires substantial work to cost to investigate occupational cohort case defined whereas abstract exposure information f is the controls. Nowa-case-control studies are used p.
The relative risk can be estimated from the corresponding odds ratio. For example, in a study of 150 male patients diagnosed with lung cancer (cases), 75 of whom were smokers and 75 of whom were nonsmokers, the odds ratio would be 15. Similar measures are the marginal and conditional odds ratios.

A well-designed case-control study is a valid and cost-effective study of the etiology of diseases. Another advantage of the case-control design is that it is susceptible to selection bias. Hence, similar to the relative risk (or relative risk ratio), the odds ratio is a measure of association between disease and exposure. The odds ratio is a measure of association between disease and exposure. The odds ratio is an unbiased estimator of the incidence rate ratio and so indicates whether there is an association between the disease and the exposure of interest.

A nested case-control study is frequently used in occupational epidemiology (Rothman and Greenland, 1998). The occupational cohort can often be readily defined whereas abstraction of detailed exposure information from existing records requires substantial work. Hence, it is more efficient to investigate only the cases of interest and a sample from the cohort that is the controls. Nowadays, nested case-control studies are used routinely when exposure information is derived, often through expensive laboratory procedures, from biologic samples such as blood or blood products, tissue, urine, or nails.

One such example is a study of selenium status and breast cancer risk in the Nurses’ Health Study (Hunter et al., 1990). On the basis of prior evidence that selenium intake may influence breast cancer risk and since selenium levels in toenails are a reliable source of selenium exposure over several months, the participating women were asked to provide toenail clippings in 1982. After 4 years of follow-up, there were 434 cases of breast cancer. It would have been very expensive and inefficient to get exposure information for all the 62 000 nurses who had returned toenail samples at the start of follow-up. Hence, 434 controls without breast cancer were sampled from the cohort. Using this design meant that only 868 rather than 62 000 samples had to be sent to the laboratory for selenium analyses (Hunter et al., 1990).
Matching in case-control studies
Occasionally, case-control studies are matched. This means that controls are chosen so as to match particular cases with respect to gender, age, race, or any other factor that is likely related to the disease under investigation but not intended to be analyzed in the particular study. Matching is not strictly necessary, nor does it increase the validity of results. But it improves statistical efficiency and, thus, the ability to substantiate a true association (Rothman and Greenland, 1998). What is necessary, however, is that, whenever matching has been used in the enrollment of cases and controls, the statistical analysis should accommodate the matching process. This can be done through either a matched analysis (for example, conditional modeling) or unmatched analysis with explicit control for the matching factors (proper application of unconditional modeling).

Studies of the Genetic Epidemiology of Cancer
Genetic epidemiology of cancer is considered in more detail in a distinct chapter (Chapter 4). Here, we refer briefly to such studies, to provide an integrated picture of epidemiologic designs available for the study of cancer etiology. Two main types of epidemiologic studies are used for the identification of genes predisposing to cancer: genetic linkage studies and genetic association studies.

Genetic linkage studies are generally undertaken in families with a high cancer burden and rely on the principle that two genetic loci, or a cancer and a particular locus, are linked when they are transmitted together from parent to offspring more often than expected by chance. Linkage extends over large regions of the genome and refers to a locus, rather than specific alleles in that locus, which can vary from study to study and from family to family (Teare and Barrett, 2005). Such studies have led to the identification of genes that have substantial impact on the occurrence of breast cancer and colorectal cancer, but these genes are generally rare, presumably because of natural selection pressure.

Genetic association studies can be of either cohort or, more frequently, case-control design. They are frequently undertaken in the general population, rather than in families, and are conceptually similar to traditional epidemiological investigations. The difference is, however, that instead of focusing on environmental factors, like smoking or diet, genetic association studies evaluate as “exposures” specific alleles (rather than loci) of genetic polymorphisms, usually single nucleotide polymorphisms (SNPs). The specific alleles may be etiologically related to cancer or, much more frequently, very closely linked to the truly etiological allele which may not be known. The actually investigated allele and the true etiological allele are said to be in linkage disequilibrium—that is, they are so closely linked that they tend to be inherited together. Two loci in linkage disequilibrium are obviously linked, but two linked loci may not be in linkage disequilibrium if they are sufficiently apart in the chromosome to be separated, sooner or later, by the frequent cross-over process in the meiosis phenomenon during the generation of gametes. In other words, linkage covers longer genetic regions than linkage disequilibrium (Cordell and Clayton, 2005; Teare and Barrett, 2005). The specific allele may be chosen to study because the corresponding locus is thought to be involved in the etiology of the cancer under investigation (eg, a candidate gene). Many SNPs over large parts of the genome, or even over the whole genome, may also be evaluated, with little or no prior evidence that most of them are etiologically relevant or are in linkage disequilibrium with etiologically relevant genes. In the latter situation, most statistically significant findings are likely to be false positive and special procedures are recommended to delineate which ones among the apparent associations are probably genuine (Wacholder et al, 2004). Genetic association studies have not been very successful to date in identifying genes or polymorphisms involved in cancer etiology, possibly be-
n studies can be of psychological processes. Genetic association "exposures" specific in etiological or population-based studies are frequently used to evaluate the genome simultaneously—rather than at a limited number usually selected on the basis of weak prior probabilities of being truly associated—are only just becoming available.

THE ROLE OF CHANCE

Before an epidemiologic association could be considered true and therefore deserve interpretation in causal terms, the role of chance and systematic errors should be considered.

The P-value

Our daily lives are full of highly unlikely events and coincidences. At the extremes, thousands of people have become wealthy from lotteries; many more have died in strange accidents, even though the probabilities for the respective events are extremely small—say one in 100 000 or smaller. The lesson is simple: Highly unlikely events happen by chance all the time. Chance does not operate differently in scientific research and everyday life. In science, however, sufficient quantification and judgment, relying on sound substantive knowledge, are necessary before considering chance as an unlikely explanation for a phenomenon.

Let us take, as an example, tossing a fair (unbiased) coin that has a 50% or 0.5 probability of turning up heads and an identical probability of turning up tails. Tossing the coin three times and getting three heads in a row is somewhat unusual but it can hardly be taken as an indication that the coin is systematically biased. The p-value in this instance is 0.0525 and is calculated by multiplying $0.5 \times 0.5 \times 0.5 = 0.125$, and then doubling 0.125, because the symmetrically opposite outcome, three tails in a row, is as extreme as three heads in a row. Getting five heads or five tails in a row generates some suspicion ($p = \frac{1}{2^5} \times 2 = 0.0625$). But if 100 people have tossed a fair (unbiased) coin five times each, it should be expected that about six ($100 \times 0.0625$) among them would have obtained either five heads or five tails in a row.

It must be realized that stochastic (probabilistic), in contrast to deterministic, processes always have built-in uncertainty. In their research, all investigators want to reduce chance-related uncertainty as much as possible in order to allow more reliable conclusions. This can be achieved mainly by enrolling progressively larger numbers of individuals in a study. The remaining uncertainty can always be assessed by utilizing statistical procedures that generate a number of summary statistics, including the p-value.

The true meaning of the p-value, however, is poorly understood and the concept itself is widely misused. Surprisingly, this misunderstanding and misuse is quite common even in scientific research. Traditionally, p-values are expressed as numerical fractions of 1. For example, a p-value of 0.1 for a particular positive association (or difference) indicates that there is a 10% chance that such an association or a more extreme one (or a symmetrically opposite one—that is an inverse association) would appear by chance, even if there were in reality no association at all.

In essence, the p-value is interpretable as such when only one comparison or one test is performed. When multiple comparisons or multiple tests are carried out the set of the respective p-values loses its collective interpretability. Various procedures for adjusting p-values according to the number of comparisons undertaken or tests performed have been proposed (Wacholder et al, 2004).

A p-value of 0.05 or smaller is traditionally—and indeed arbitrarily—treated in medical research as evidence that an observed association may not have arisen by chance. For example, the proportion of long-term smokers is found to be larger among lung cancer patients than among individuals without the disease and the p-value for this difference is, say, 0.05. This implies that the probability of finding a difference of this magnitude or larger (in absolute terms) is 5% if smoking were
In order to integrate information about the strength of an association (as reflected in the relative risk-effect measure, described later on) and its statistical significance, the concept of confidence interval has been developed. Most common are 95% confidence intervals. With a 95% confidence interval, one can be 95% confident that the interval covers the true measure of association (for example the relative risk). But in 5 times out of 100, the true measure is not included. The confidence interval is closely linked to the p-value. The width of the confidence interval is determined primarily by the desired level of confidence and the sample size. Hence, the interval is wider if it includes the true value with 95% confidence than with, for example, 80% confidence. Likewise, smaller studies create wider confidence intervals—that is, greater uncertainty about the true value—than larger studies.

SYSTEMATIC ERRORS

The Experimental Study

The chance-related issues apply to all types of studies, observational as well as experimental. As discussed earlier, experimental studies undertaken under optimal conditions are methodologically superior to observational studies. With randomization of exposure, complete follow-up of study subjects, and double-blind assessment of outcome, they are not as liable to the pitfalls of typical observational studies—that is confounding and bias (Miettinen, 1985; Hennekens and Buring, 1987; MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998; Rothman, 2002). Proper evaluation of the association between a particular exposure and a specific disease presupposes that every other factor that could influence disease occurrence is either constant among subjects studied or distributed equally between exposed and unexposed subjects. In other words, an experimental study uses random allocation of study subjects into those who will be exposed and those who will not. Thus, the two or more groups will tend to be similar in distribution to known as well as unknown factors that may influence the results. In some studies, blinding of researchers through the use of app and devices (for example, inert pills, the so-called ther assure that every fi disease occurrence, other under study, is kept at a between the exposed anc Experimental studies Latin dictum ceteris p being equal). However, timal conditions that c confounding and bias a even in randomized con over, as already indicate fully control the inher role of chance, except large numbers of study alistic objective in man The randomized cont methodological advan experimental research In humans, however, experiments faces seri most important of wh obviously not acceptabl intentionally to a poten agent in order to asce tion. For this reason, controlled trials in hun formed to evaluate tre and occasionally to det tive potential of vaccine. supplements. In most in disease etiology has to mal models—with inhu
measure, described later in significance, the con-
terval has been devel-
95% confidence interval,
ident that the interval
ure of association (for risk). But in 5 times out
asure is not included.
d closely linked to
楇d primarily by the de-
ider and the sample size.
ider if it includes the
idence with confidence. Likewise, a wider confidence in-
ter uncertainty about larger studies.

IC ERRORS

issues apply to all types
onal as well as experien-
dermal, experimental
nder optimal condi-
tions. With randomization of
ollow-up of study sub-
and assessment of out-
iable to the pitfalls of
udies—that is confi-
ientiin, 1983; Hen-
nian and Mac-
man and Gre-
, 2002). Proper evalu-
 between a particular
c disease presupposes
that could influence
either constant among
distributed equally be-
exposed subjects.
in experimental study
ion of study subjects
be exposed and those
he two or more groups
ilar in distribution to
unknown factors that
ults. In some studies,
blinding of researchers and study subjects
through the use of appropriate procedures
and devices (for example, indistinguishable
 inert pills, the so-called placebos) may fur-
ther assure that every factor that can affect
disease occurrence, other than the exposure
under study, is kept at about the same level
t between the exposed and unexposed groups.

Experimental studies aim to fulfill the
Latin dictum ceteris pariba (other things
being equal). However, in humans, the op-
timal conditions that completely eliminate
founding and bias are difficult to create
even in randomized controlled trials. More-
over, as already indicated, there is no way to
fully control the inherently unpredictable
role of chance, except by the use of very
large numbers of study subjects—an unreal-
istic objective in many studies.

The randomized controlled trial, with its
methodological advantages, dominates ex-
perimental research in laboratory animals.
In humans, however, the undertaking of
experiments faces serious obstacles, the
most important of which are ethical. It is
obviously not acceptable to expose humans
intentionally to a potentially carcinogenic
agent in order to ascertain cancer causation.
For this reason, most randomized
controlled trials in humans have been per-
fomed to evaluate treatment effectiveness
and occasionally to determine the prevent-
tive potential of vaccines, vitamins, or other
supplements. In most instances, research on
disease etiology has to rely either on ani-
mal models—with inherently dubious as-
sumptions about interspecies similarities
and exposure dose extrapolations—or on
epidemiologic studies with an observational
design.

Epidemiologic studies have indeed gen-
erated most of what is currently known
about the etiology of human diseases in
general, and cancer in particular. At the
same time, however, epidemiologic studies
have also generated conflicting results, un-
warranted concern about everyday expo-
sures, and considerable confusion over the
rational ranking of public health priorities
(Taubes, 1995). The problem arises because
epidemiologic studies must confront not
only the vagaries of chance but also the
problems of systematic errors that under-
mine their validity.

Confounding

Confounding is the systematic error gener-
ated when another factor that causes the
disease under study, or is otherwise related
to it, is also related to the exposure un-
der investigation (Fig. 6-4A). Thus, if one
wishes to examine whether hepatitis C virus
(HCV) causes liver cancer, hepatitis B virus
(HBV) would be a likely confounder. Con-
founding arises because HBV causes liver
cancer and carriers of HBV are more likely
to also be carriers of HCV (because these
two viruses are largely transmitted by the
same routes). Hence, if the confounding
influence of HBV is not accounted for in
the design (by limiting the study to HBV-
negative subjects) or in analyses of the data,

![Figure 6.4A](image_url)

**Figure 6.4A.** Infection with hepatitis C virus (HCV), a cause of liver cancer, is (positively)
confounded by hepatitis B virus (HBV) infection, another cause of liver cancer. If this
confounding is disregarded, the strength of the association between HCV and liver cancer will be
overestimated.
then the strength of the association between HBC and liver cancer would be overestimated (Fig. 6-4A).

A more trivial example is the strong association between carrying matches or a cigarette lighter and developing lung cancer. Obviously, neither matches nor lighters cause lung cancer and their association to the disease is due entirely to confounding by cigarette smoking. The confounding factor, cigarette smoking, is the true cause of lung cancer and the dependence of cigarette lighting on matches or lighters generates the confounded, entirely spurious association of the latter two factors with the disease (Fig. 6-4B).

There are several ways to deal with confounding: some simple, others more complicated. They all assume that two conditions are satisfied: (1) that all the confounders have been identified or at least suspected, and (2) that the identified or suspected confounders can be adequately conceptualized and accurately measured. When the study is fairly large, it is always possible to evaluate all suspected confounders in the analysis. However, the ability to conceptualize and accurately measure all of them is frequently beyond the control of any investigator. The result is what has been termed residual confounding, that is, confounding left unaccounted for (MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998).

Bias

Compounding the problems of epidemiologic studies is that the data are almost never of optimal quality. Data collection relies on the recollection of exposures and their accurate reporting by study participants, laboratory procedures, or existing records. These sources are rarely perfect. For example, studies on diet rely on individuals' imperfect recall on how frequently they eat specific foods, or on serum markers of nutrients that are far from perfect indicators of long-term consumption. Such misclassification, or information bias, can influence the relative risk in any direction and, thus, entails exaggeration, underestimation, or even reversal of the true associations.

In case-control studies, the ascertainment of exposure occurs after the occurrence of disease. Therefore, this study design is particularly subject to information bias. In particular, cases may be likely to remember their exposures differently than controls—a form of information bias called recall bias. For example, a reasonable concern is that cases, or their relatives, are inclined to ruminate about the disease and identify a particular exposure as the causative agent, either for conscious or subconscious reasons. Cases may also try harder than controls to recall relatives with the disease of interest, leading to a biased estimate of the effect of (Chang et al, 2006).

A well thought-out and executed procedures, and careful measures can reduce some quantification of information bias. However, complete ascertainment of confounders cannot be eliminated can never be addition, the reliance of controls on a control series that has to meet criteria of eligibility to the case series, and general practitioners susceptible to selection bias-and direction and magnitude of the population arise when eligible controls are unavailable in the person-time that goes into the analysis (Wacholder et al, 1992b, 1992b; Wacholder et al, 1992b).

Assume as in the same that controls refuse to participate if they are smoker nonsmokers. We would underestimate both the cases and controls and the controls enrolled through phone lists have their own biases, issues of chance are equally relevant to cohort investigations (Hennek, 1987; MacMahon and Greenland, 1998).

In contrast to selection biases, issues of chance are equally relevant to cohort investigations (Hennek, 1987; MacMahon and Greenland, 1998).

**ANALYSIS OF EPIDEMIOLOGIC STUDIES**

**Effect Measures**

The underlying goal of determining the magnitude of the frequency caused by the condition of interest, leading to a biased estimate of the effect of these exposures different from controls—a form of information bias called recall bias. For example, a reasonable concern is that cases, or their relatives, are inclined to ruminate about the disease and identify a particular exposure as the causative agent, either for conscious or subconscious reasons. Cases may also try harder than controls to recall relatives with the disease of interest, leading to a biased estimate of the effect of information.
mate of the effect of the family history (Chang et al, 2006).

A well thought-out protocol, standardized procedures, and built-in quality control measures can reduce bias and allow some quantification of its potential impact. However, complete assurance that bias has been eliminated can never be achieved. In addition, the reliance of case-control studies on a control series that simultaneously has to meet criteria of compliance, comparability to the case series, statistical efficiency, and general practicality makes them susceptible to selection bias of unpredictable direction and magnitude. Such biases arise when eligible controls are not representative of the population, or more strictly the person-time, that gave rise to the cases (Wacholder et al, 1992a; Wacholder et al, 1992b; Wacholder et al, 1992c).

Assume as in the same previous example that controls refuse to participate more often if they are smokers than if they are nonsmokers. We would then underestimate smoking in the control group and thereby overestimate both the difference between cases and controls and the excess risk. Hospital controls, neighborhood controls, and controls enrolled through searches of telephone lists have their own problems, and these have been extensively discussed (MacMahon and Trichopoulos, 1996).

In contrast to selection and information biases, issues of chance and confounding are equally relevant to cohort and case-control investigations (Hennekens and Buring, 1987; MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998).

ANALYSIS OF EPIDEMIOLOGIC STUDIES

Effect Measures

The underlying goal of epidemiology is to determine the magnitude of change in disease frequency caused by an exposure. How do we accomplish this? We could measure the cumulative incidence or incidence rate among those exposed to a factor. For example, we could observe that the incidence rate of breast cancer in a population of alcoholic women is 60/10,000 person-years. This information provides an estimate of the overall disease burden in this study base. However, we do not know how many cases would have arisen in the study base if all the women in this population had not been alcoholics. In epidemiology, the unexposed group stands in for the person-time experience of the exposed group had it not been exposed. Thus, we need to harvest information from both exposed and unexposed person-time.

There are several ways through which an association, or lack thereof, is assessed. Consider a population of women exposed to a high saturated fat diet and a group exposed to low saturated fat diets that are followed for 5 years to see if they develop breast cancer. The absolute effect of the high-fat diet would be the difference in the cumulative incidence between the two groups, or the difference in the incidence rates. Since the experience of the low saturated fat group should represent what would have happened to the high saturated fat group if they had not eaten the high saturated fat, and if the two groups are equivalent with respect to other breast cancer risk factors, the difference in risks or rates represents the excess risk or rate. These absolute-effect measures are called the risk difference and rate difference, respectively.

Although the absolute measures are easily interpreted, more common are effect measures that are taken as ratios and collectively known as the relative risk. This term includes the risk ratio, rate ratio, odds ratio, standardized mortality ratio, and standardized incidence ratio. The risk ratio is simply the cumulative incidence of disease among the exposed, divided by the cumulative incidence among the unexposed. The rate ratio is a ratio of the rates of disease among the exposed and unexposed. The odds ratio is the odds of disease among the exposed divided by the odds of disease among the unexposed. Lastly, the standardized mortality ratio or standardized incidence ratio is a ratio of the observed number of deaths or cases in a cohort, divided by the expected
number of deaths or cases in the general population, usually stratified by age and gender.

A relative risk value of 1 implies that the exposure under study does not affect the incidence of the disease under consideration. Values below and above 1 indicate a negative (inverse) and a positive association, respectively. For example, a relative risk of 0.5 implies that the disease occurs only half as frequently among exposed as among unexposed individuals; the studied factor appears to be protective. In contrast, if the relative risk is 1.5, then the occurrence (usually the incidence) is 50% higher among exposed than among unexposed individuals.

Studies based on follow-up of closed cohorts may be analyzed by using either cumulative incidence (risk) measures or by counting person-time and calculating incidence rate measures. Analyses based on cumulative incidence measures are only useful under certain conditions, such as no loss to follow-up, no competing risks, and unchanged exposure status throughout follow-up. In addition, study subjects should be followed for the same period of time. Whether or not these conditions are met, it is always valid to conduct analyses based on person-time, using incidence rate measures.

Interaction

The term interaction has been used to describe different biological and statistical concepts. Indeed, even in the epidemiologic literature, statements about interaction are often ambiguous and inadequately specified. From a biological point of view, component causes within the same sufficient cause may be thought of as interacting (Fig. 6–1). In other words, the exposures act synergistically to produce disease, since in the absence of one factor, disease will not occur by that mechanism. From an epidemiologic point of view, interaction is frequently characterized as effect-modification: That is, a factor A and factor B alone have a certain relationship with a disease, but together the factors have an effect different than that expected on the basis of the magnitude of their individual effects. The expectation of the joint effect of factors A and B can be assessed in either an additive or a multiplicative way.

We can use the example in Table 6–2 to illustrate how interaction is assessed. When a multiplicative scale is assumed, there is statistical interaction if the relative risk among those exposed to both factors A and B (that is, $RR_{AB}$) is different than the product of the two individual relative risks (that is, $RR_A \times RR_B$). When an additive scale is assumed, there is interaction if the $RR_{AB}$ is different than $RR_A + RR_B - 1$. In this example, the expected relative risk for someone with both exposures is 6.0 ($6.0 = [4.0 + 3.0 - 1]$) under the additive-effect assumption (Table 6–2A), whereas it is 12.0 (12.0 = 4.0 x 3.0) under the multiplicative-effect assumption (Table 6–2B).

Hence, interaction between two exposures is present when the relative risk is significantly different from what is expected according to a specified scale. Thus, for those with both exposures, we would have interaction on the additive scale if the relative risk is significantly different from 6.0 (Table 6–2A), and on the multiplicative scale if the relative risk is significantly different from 12.0 (Table 6–2B). If the relative risk following exposure to both factors compared to having neither is greater than the sum (minus the reference risk of 1, which should not be counted twice) or product of the individual risks, we call this interaction super additive or super multiplicative, respectively. If the relative risk is significantly lower, we refer to this as either subadditive or submultiplicative.

We can illustrate the concept of interaction using data from an epidemiological study of asbestos, smoking, and lung cancer risk. The source population for the data shown in Table 6–2C is a cohort of insulation workers from the United States and Canada (Hammond et al., 1979). The exposed person-time was the experience of over 12,000 male workers with at least 20 years of asbestos exposure. The comparison person-time came externally from the experience of more than 73,000 men of similar social class.

### Table 6–2A. Statistical interaction superadditive factors.

<table>
<thead>
<tr>
<th>Factor A</th>
<th></th>
<th>Factor B</th>
<th></th>
<th>RR_{AB}</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>1.0</td>
<td>-</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 6–2B. Statistical interaction and supermultiplicative factors.

<table>
<thead>
<tr>
<th>Factor A</th>
<th></th>
<th>Factor B</th>
<th></th>
<th>RR_{AB}</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>1.0</td>
<td>-</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 6–2C. Effects on lung cancer risk

<table>
<thead>
<tr>
<th>Smoking</th>
<th></th>
<th>RR_{smok}</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>1.0</td>
<td>10.9</td>
</tr>
<tr>
<td>+</td>
<td></td>
<td>10.9</td>
</tr>
</tbody>
</table>


Compared to men who were exposed to asbestos, but who were not smokers, was there an occupational hazard of lung cancer? The relative risk of lung cancer for those exposed to asbestos, not smokers, was 10.9 ($\text{RR}_{\text{smok}} = 10.9 + 5.2 - 1$). We observe interaction on
Table 6-2. Definitions of interaction. Relative risks of developing a certain disease among subjects exposed (+) or not exposed (−) to one or both factors denoted A and B. Subjects exposed to neither of these factors comprise the reference category and their relative risk is by definition 1.0.

**Table 6-2A. Statistical interaction on the additive scale with examples of subadditive and superadditive factors.**

<table>
<thead>
<tr>
<th>Factor A</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>−</td>
<td>1.0 (reference)</td>
<td>3.0</td>
</tr>
<tr>
<td>+</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>Subadditive</td>
<td>Expected under additive effects assumption</td>
</tr>
<tr>
<td></td>
<td>Superadditive</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6-2B. Statistical interaction on the multiplicative scale with examples of submultiplicative and supermultiplicative factors.**

<table>
<thead>
<tr>
<th>Factor A</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>−</td>
<td>1.0 (reference)</td>
<td>3.0</td>
</tr>
<tr>
<td>+</td>
<td>4.0</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>Submultiplicative</td>
<td>Expected under multiplicative effects assumption</td>
</tr>
<tr>
<td></td>
<td>Supermultiplicative</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6-2C. Effects on lung cancer risk of smoking, asbestos, and both factors.**

<table>
<thead>
<tr>
<th>Asbestos</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>−</td>
<td>1.0 (reference)</td>
<td>5.2</td>
</tr>
<tr>
<td>+</td>
<td>10.9</td>
<td>53.2</td>
</tr>
</tbody>
</table>


Compared to men who had neither exposure, the relative risk of those who were smokers, but who were not exposed to asbestos occupationally was 10.9; the relative risk of those exposed to asbestos, but who were not smokers, was 5.2. For those exposed to both asbestos and smoking, the relative risk of lung cancer was 53.2 compared to those with neither factor. In this example, there appears to be interaction on the additive scale, since the RR_smoker and asbestos = 53.2 is substantially higher than the expected relative risk of 15.1 under the additive model (RR_smoker + RR_asbestos − 1 = 10.9 + 5.2 − 1). We do not, however, observe interaction on the multiplicative scale, since the relative risk for both smoking and asbestos (53.2) does not represent a significant departure from what is expected under the multiplicative-effect assumption (56.7 = RR_smoker × RR_asbestos − asbestos = 10.9 × 5.2).

There are not any clear-cut guidelines on whether to assess interaction in the additive or multiplicative setting for the various disease outcomes examined in epidemiology, although both approaches are used (Brennan, 1999).

**Meta-analysis**

Random variation per se in epidemiologic studies is not an insurmountable problem.
Larger studies and eventually quantitative summary analyses are increasingly used. Such systematic statistical evaluations of results of several independent investigations can effectively address genuine chance-related concerns. Quantitative summary analyses have been termed meta-analyses and pooled analyses. There is no completely accepted distinction between the two terms, although meta-analysis is used more frequently when published results are combined. By contrast, in pooled analysis primary individual-level data from different studies may be made available to an investigator who undertakes the task of combining them. This facilitates the use of uniform exposure categories and statistical analyses across studies and may permit analyses that were not in the original publications. For instance, analyses of effect modification for which each initial study may have been too small to be informative.

Meta-analyses and pooled analyses have been widely and effectively used for randomized controlled trials and intervention studies, because in properly undertaken investigations of this nature confounding and bias are nonissues (Sacks et al, 1987). For observational epidemiologic studies, however, the role of meta-analysis is not universally accepted (Shapiro, 1994; Feinstein, 1995). Some investigators are concerned that no statistical summarization can effectively address problems generated by residual confounding, unidentifiable bias, or chance, perhaps following a multiple testing process.

Repeated demonstration of an association of similar direction and magnitude in several studies, undertaken by different investigators in different population groups, increases confidence in a genuine causal basis but cannot conclusively establish this. Nor do meta-analyses establish causality. These techniques essentially address the issue of chance and provide no guarantee that a particular bias, unrecognized confounding, or selective reporting have not operated in the constituent studies. It is at this stage that both biologic and epidemiologic considerations should be taken into account in interpreting the results of empirical studies.

Criteria for inferring causation from epidemiologic investigations have been proposed, over the years, by several authors, including MacMahon et al (1960), the US Surgeon General (US Department of Health, 1964), Sir Austin Bradford Hill (Hill, 1965), the IARC (1987), and others. In spite of differences in emphasis, a similar set of principles have been invoked by most authors. Sir Austin Bradford Hill (1965) advocated the nine widely used criteria listed in Table 6-3, to distinguish causal from noncausal associations.

The Hill criteria, although sensible and useful, do not separately address the inherently different issues that are posed by the results of a single study, the results of several studies, and the likelihood of causation in a certain individual case. The perceived likelihood of a particular disease moves for a continuous spectrum as adequate evidence accumulates. The evidence declared as sufficient by the International Agency for Research on Cancer (IARC) requires reevaluation of the evidence (Cole, 1987).

Table 6-3. The Hill criteria for inferring causality

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Strength</th>
<th>Consistency</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

not be confused with causation based on
ons alone.
a observational epide-
ed to address a specific
ng, the study is large,
ese of overt confound-
s, it is legitimate to
ferences. In contrast,
es problematic when a
s out to be statistically
ple, in a large but im-
ough that association
—but genuine—causal
also be the result of
unidentifiable
ps following a multi-
ration of an association
and magnitude in
rtaken by different in-
population groups,
in a genuine causal
istically establish this.
es establish causality.
tially address the in-
wide no guarantee that
recognized confound-
ing have not operated
ides. It is at this stage
and epidemiologic con-
taken into account in
als of empirical studies.
ng causation from ep-
itations have been pro-
s, by several authors, in
al et al (1960), the
 (US Department of
n Bradford Hill
.C (1987), and others.
emphasis, a similar
been invoked by most
Bradford Hill (1965)
idely used criteria list-
istinguish causal from
al, although sensible and
ately address the in-
es that are posed by
study, the results of the
likelihood of cau-
sation in a certain individual. In reality, the
perceived likelihood of a causal association
between a particular exposure and a specific
disease moves forward or backward in a
continuous spectrum as research results
accumulate. The evidence for causality is
declared as sufficient when a particular
threshold has been reached, but on occasion
requires reevaluation in the light of subse-
quent evidence (Cole, 1997).

The IARC Classification
The International Agency for Research on
Cancer (IARC) evaluates the risk of specific
agents to determine if they are carcinogenic
in humans. In order to come to a conclu-
sion, the IARC has implemented its own set
of criteria for evaluating the carcinogeni-
city of agents. After considering all the ev-
evidence, the IARC working group assigns the
agent to one of five categories, summarized
in Table 6-4. Group 1 indicates that there
is sufficient evidence to conclude that the
agent is carcinogenic to humans. A label of
group 2A means that there are insufficient
human data, but there is strong evidence
that the agent is carcinogenic in animal
models. Agents for which there is limited
evidence in humans and insufficient evi-
dence in experimental animals are assigned
to group 2B. Group 3 is used when there is
inadequate human and animal data to come
to a conclusion. Group 4 indicates that the
agent is most likely not a carcinogen in
humans based on adequate evidence sug-
gesting that it is not a carcinogen in both
animal models and human studies.

The Process of Causal Inference
Criteria for causality can be invoked, ex-
plicitly or implicitly, in evaluating the results
of a single epidemiologic study, although,

| Table 6-3. The Hill criteria for inferring causation |
|-----------------|-------------------------|
| **Criteria**    | **Definition**          |
| **Strength**    | A strong association is more likely to be causal. The measure of strength of an association is the relative risk and not statistical significance. |
| **Consistency** | An association is more likely to be causal when it is observed in different population groups. |
| **Specificity** | When an exposure is associated with a specific outcome only (for example, a cancer site or even better a particular histological type of this cancer), then it is more likely to be causal. There are exceptions, however, for example, smoking causing several forms of cancer. |
| **Temporality** | A cause should not only precede the outcome (disease), but also the timing of the exposure should be compatible with the latency period (in non-infectious diseases) or the incubation period (in infectious diseases). |
| **Gradient**    | This criterion refers to the presence of an exposure-response relationship. If the frequency or intensity of the outcome increases when an exposure is more intense or lasts longer, then it is more likely that the association is causal. |
| **Plausibility**| An association is more likely to be causal when it is biologically plausible. |
| **Coherence**   | A cause and effect interpretation of an association should not conflict with what is known about the natural history and biology of the disease, or its distribution in time and place. |
| **Experimental evidence** | If experimental evidence exists, then the association is more likely to be causal. Such evidence, however, is seldom available in human populations. |
| **Analog**      | The existence of an analogy (for example, if a drug causes birth defects, then another drug could also have the same effect) could strengthen the belief that an association is causal. |

Table 6-4. International Agency for Research on Cancer (IARC) classification of carcinogenicity of agents, mixtures or processes

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>The agent is carcinogenic to humans</td>
</tr>
<tr>
<td>Group 2A</td>
<td>The agent is probably carcinogenic to humans</td>
</tr>
<tr>
<td>Group 2B</td>
<td>The agent is possibly carcinogenic to humans</td>
</tr>
<tr>
<td>Group 3</td>
<td>The agent is not classifiable in terms of its carcinogenicity</td>
</tr>
<tr>
<td>Group 4</td>
<td>The agent is not carcinogenic to humans</td>
</tr>
</tbody>
</table>

Background

Causality is never be inferred on the basis of a single epidemiologic study, but the likelihood that an observed association is causal is strengthened when several of the following criteria are met: (1) minimal confounding; (2) minimal bias; (3) limited chance variation; (4) relatively strong association; (5) monotonic exposure-disease association, otherwise referred to as exposure-response or dose-response association; (6) internal consistency, exemplified by similarity of exposure-response patterns among various subgroups of study subjects; (7) compatibility of the temporal sequence of exposure and outcome with the known or presumed latency of the disease; and, lastly, (8) biologic plausibility, that is, a causal link between the exposure and the disease should be, at a minimum, biologically conceivable.

The general case (several studies, level II)

Establishment of the etiologic role of a particular exposure on the occurrence of a disease ideally requires strong epidemiologic evidence, an appropriate and reproducible animal model, and documentation at the molecular or cellular level of the morphological or functional pathogenetic process. Sometimes, an intended or unintended change, or natural experiment, greatly facilitates etiologic inference: This happens when, for example, an occupational group is exposed to high levels of compounds rarely encountered in other settings, a religious group avoids an exposure that is otherwise widespread, or a vaccine that creates herd immunity against a particular virus turns out to reduce the incidence of a certain form of cancer.

These conditions, however, are rarely collectively satisfied. Instead investigators have to be guided by the best available biomedical evidence in order to interpret correctly epidemiologic data from several studies. The following criteria need to be considered: (1) consistency, that is similarity (lack of heterogeneity) of results obtained by different investigators using different study designs in different populations; (2) overwhelming biomedical evidence for weak associations, whereas for strong associations reliance on powerful biomedical knowledge is less critical; (3) compatibility of exposure-response patterns across different studies exploring the exposure-disease association in different exposure ranges; (4) coherence, which requires results from analytic epidemiologic studies to be compatible with ecologic patterns and time trends, such as the increasing use of cigarettes by the population; (5) existence when one type of disease is strongly linked with one rather than several exposures associated with a certain exposure being associated with diseases; and (6) biological plausibility when a similar condition is shown to cause a similar disease in other settings.

Disease in a specific part of the body

Causality can be concluded between a particular exposure and a particular disease. This is not possible in a natural experiment, but the etiologic inference is strengthened when most of the following criteria are met:

1. The exposure under consideration is causal and specific to the population; (level II). (2) The relative risk is high. (3) The disease under consideration is specific and at least one other disease is similarly associated with the same exposure. (4) Biological plausibility is present.

For this conclusion following criteria must be met:

1. The exposure under consideration must be an etiologic factor of the disease under study. (level II). (2) The relative risk is high. (3) The disease under consideration is specific and at least one other disease is similarly associated with the same exposure. (4) Biological plausibility is present.
patterns and time trends, such as the increasing incidence of lung cancer over time, following the increasing use of tobacco products by the population; (5) specificity, which exists when one type of disease is consistently linked with one type of exposure rather than several exposures all being associated with a certain disease, or one type of exposure being associated with several diseases; and (6) biological analogy, which exists when a similar exposure has been shown to cause a similar disease in another species or a different form of the disease in humans. For example, viruses have been shown to cause leukemia in several animal species and at least one rare form of leukemia in humans.

None of these criteria can be considered as absolutely necessary for causal inference—a sine qua non. But the evidence for causality is strengthened when most of them are met.

**Disease in a specific person (level III)**

Causality can be conclusively established between a particular exposure as an entity and a particular disease as an entity. In contrast, it is not possible to establish such a link conclusively between an exposure and a particular disease of a given individual—for example, smoking in a patient with lung cancer. It is possible, however, to infer deductively that the specific individual’s illness was more likely than not caused by the specified exposure.

For this conclusion to be drawn, all the following criteria must be met (Cole, 1997): (1) The exposure under consideration, as an entity, must be an established cause of the disease under consideration, as an entity (level II). (2) The relevant exposure of the particular individual must have properties comparable (in terms of intensity, duration, associated latency, etc) to those that have been shown to cause the disease under consideration. (3) The disease of the specified person must be identical to, or within the symptomatological spectrum of, the disease that, as an entity, has been etiologically linked to the exposure. (4) The patient must not have been exposed to another established or likely cause of this disease. If the patient has been exposed to both the factor under consideration (for example, smoking) and to another causal factor (for example, asbestos), individual attribution becomes a function of several relative risks, all versus the completely unexposed: (a) relative risk of those who only had the exposure under consideration, (b) relative risk of those who had only been exposed to the other causal factor(s), and, (c) relative risk of those who have had a combination of these exposures. (5) The relative risk should be reasonably elevated (e.g., 2 or more).

The last criterion stems from the fact that the relative risk comprises a baseline component equal to 1, which characterizes the unexposed, plus another component that applies only to the exposed. When the relative risk is higher than 1 but less than 2 the individual who has been exposed and has developed the disease is more likely than not to have developed the disease for reasons not entirely due to the exposure. For instance, if the risk of a light-smoking 55-year-old man to suffer a first heart attack in the next five years is 6%, and that of a same-age non-smoking man is 4% (relative risk 1.5), then only 33% of the smoker’s risk (that is, 1/3 of the total 6%) can be attributed to his smoking. When the relative risk is higher than 2, a particular individual who has been exposed and has developed the disease under consideration is more likely than not to have developed the disease because of the exposure.

**CONCLUSION**

Manipulation of exposures in humans, many of which may be harmful, is frequently unethical, unfeasible, unethical, or both. Therefore, epidemiologists have to base their inferences on experiments that humans subject themselves to intentionally, naturally, or even unconsciously. The study of risk for lung cancer among smokers compared with nonsmokers is one classic example of a natural experiment.

Because human life is characterized by myriad complex, often interrelated, behaviors and exposures—ranging from genetic
traits and features of the intrauterine environment to growth rate; physical activity; sexual practices; use of tobacco, alcohol, and pharmaceutical compounds; dietary intake; exposure to infections, environmental pollutants, and occupational hazards; and so on—epidemiologic investigation is difficult and challenging. Given this complexity, it is not surprising that from time to time epidemiologic studies generate results that appear confusing, biologically absurd, or contradictory. However, it is reassuring that a wealth of new knowledge has been generated by epidemiologic studies over the last few decades. This knowledge now lays the scientific ground for primary prevention of many major cancers and other chronic diseases among humans globally.

A detailed study of epidemiologic methodology in any textbook (Hennekens and Buring, 1987; Miettinen, 1985; Walker, 1991; MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998; Rothman, 2002) can be fascinating and indeed necessary for those who want to pursue their own research. However, for the reader of this textbook, the general concepts introduced in this chapter should provide a sufficient basis. We have tried to convey that the sometimes esoteric theory of modern epidemiology can be condensed to a few central issues—namely (1) how to quantify and understand the impact of chance, (2) how to best harvest information on exposures and outcomes from a source population by using a cohort design, a case-control design, or variants thereof, (3) how to achieve valid results by minimizing the impact of confounding and bias, and, (4) how to address the central issue of causality in a structured way.

GLOSSARY

**Cause** A factor is a cause of a certain disease when alterations in the frequency or intensity of this factor—without concomitant alterations in other factors—are followed by changes in the frequency of occurrence of the disease, after the passage of a certain time period (latency, or induction period).

**Closed cohort** A closed cohort comprises a set of individuals who are followed for a defined period of time. After becoming a member of the cohort, an individual remains in the cohort until the end of the study, or development of the outcome.

**Competing risks** The risk of death from a certain disease competes with the risk of death from another disease by affecting time at risk. Competing risks generally bias risk ratios, but not rate ratios, since person-time allows for different follow-up time.

**Component cause** An exposure that acts in concert with other factors (component causes) to produce disease. None of these factors are sufficient in themselves to cause disease.

**Confidence interval** A statistical measure that provides range of possible values that include the true measure of association with a particular degree of certainty. For example, a 95% confidence interval provides a range of values that will include the true value 95% of the time.

**Confounding** A systematic error generated when another factor, that causes the disease under study or is otherwise related with it, is also related to the exposure under investigation, without being in the pathway that links exposure under investigation with the disease under study.

**Eco logic study** The study of exposure and the disease at the population level, rather than at the individual level.

**Epidemiology** The nonexperimental investigation of determinants of human disease.

**Experimental study** See randomized controlled trial.

**Information bias** A random, or nonrandom, misclassification of information on either the exposure, outcome, or confounding variables that leads to a biased estimation of the true effect.

**Loss to follow-up** The inability to follow beyond a certain point in time and thus ascertain the ultimate fate of individuals in a cohort study.

**Necessary cause** A factor or exposure that is essential in the etiology of the disease and without which the disease cannot occur.

**Open cohort** A cohort membership changes entering or exiting over time.

**Person-time** The sum of all study participants.

**p-value** A value that indicates the probability of observing, at least, or more extreme between a particular disease, if there were no association between the exposure and the disease.

**Recall bias** A miscue, common in cross-sectional studies, when subjects were asked if they had done the various activities that are associated with the disease.

**Relative risk** A descriptive comparison of the number of cases in exposed and unexposed groups.

**Selection bias** A systematic error generated from the process of the study or on account of participation bias occurs when the exposure and/or the disease are related to the selection process.
Risk of death from any cause or from a specific disease is generally bias relative to that of an individual remains follow-up time.

The human immunodeficiency virus is a necessary cause of acquired immunodeficiency syndrome, although other factors may be involved in order for the disease to occur.

Nonexperimental study See observational study.

Observational study A study in which the investigator cannot control the circumstances of the exposure.

Odds ratio A relative measure of association, which is calculated as the ratio of the odds of disease among the exposed divided by the odds of disease among the unexposed.

Open cohort A cohort of individuals whose membership changes over time, with people entering or exiting based on defining criteria.

Person-time The sum of all time spent by each study participant at risk for a disease.

p-value A value that indicates the likelihood of observing an association as extreme as, or more extreme than, the one found between a particular exposure and a certain disease, if there were in fact no association.

Randomized controlled trial An experimental study design in which the researcher randomly allocates subjects to groups that will be subjected or not to a particular exposure.

Recall bias A misclassification of an exposure, common in case-control studies, that occurs when subjects with the disease remember or report their exposures differently than those without disease.

Relative risk A term that collectively describes the various relative measures of association, that is the risk ratio, the rate ratio, the odds ratio, and the standardized incidence or mortality ratio.

Selection bias A systematic error that results from the process of selecting participants for the study or on account of factors that influence participation in the study. Selection bias occurs when the relationship between the exposure and the disease is different for those in the study than for those not in the study.

Study base The person-time of a group of individuals at risk for a disease from which an investigator aims to harvest information about disease occurrence.

Sufficient cause A minimal set of factors or exposures that inevitably produce the disease after a certain period of time.


