Epidemiology

By Daniel Wartenberg, FACSNET Scholar
Doug Ramsey and John Warner, FACSNET Editors;
Doris Ober, FACSNET Technical Editor
Foundation for American Communications

There are four most common types of epidemiological studies:
Cohort Study
Case Control Study
Occupational Epidemiological Study
Cross-Sectional Study

This chapter explains why and when epidemiologists prefer one type of study over another and describes strengths and weaknesses of each approach.

To begin an epidemiologic study, we decide what to study.

For this discussion, let's say we want to study prenatal exposure to electric and magnetic fields and the effect on a baby's birthweight. We look at the existing literature on birthweight to assess current knowledge and data. It is important to know if others have conducted similar studies in case they have uncovered specific design limitations or useful results, and this information is helpful in understanding the context of one's own study.

We believe that known risks include prematurity, poor prenatal care, low socioeconomic status, non-white ethnicity, large size of the mother, younger or older mothers, smoking, alcohol consumption and a host of other factors. Electric and magnetic field exposures are not known risk factors but have not been studied extensively. Therefore we wish to study them.

Cohort Study

The "What will happen to me?" study follows a group of healthy people with different levels of exposure and assesses what happens to their health over time. It is a desirable design because exposure precedes the health outcome — a condition necessary for causation — and is less subject to bias because exposure is evaluated before the health status is known. The cohort study is also expensive, time-consuming and the most logistically difficult of all the studies. It is most useful for relatively common diseases. To assess suitability, we find out the commonality of the disease we wish to study. Does it occur in 10 percent of all births, 1 percent of births, or 0.001 percent of births? For example, if low weight occurs in 10 percent or more of all births, then we might investigate a relatively small group of newborns, say 200 to 400, and characterize them with respect to their exposures during pregnancy. We would expect to see 20 to 40 low-weight babies in this group. We would want to know if, while pregnant, their mother's exposure to electric and magnetic fields was different from the other

180 to 360 births.

The cohort study approach is good for our hypothetical study because we can identify a number of pregnant women, characterize their exposure during almost their entire pregnancies and assess the babies' weights at birth. Thus we limit the possibility of investigator preferences, or "bias," affecting the selection of study subjects. We must make sure nearly everyone selected for our study participates, because those not participating may be different from those who do, causing another type of bias. The assumption of risk (such as exposure to electric and magnetic fields during pregnancy) necessarily precedes the outcome (birth), and that is a necessary condition for inferring a cause.

Finally, because we selected our study subjects on the basis of their exposure only, the cohort approach enables us to look at other pregnancy outcomes such as birth defects, spontaneous abortion or increased mortality, in addition to birthweight.

Case-Control Study

The "why me?" study investigates the prior exposure of individuals with a particular health condition and those without it to infer why certain subjects, the "cases," become ill and others, the "controls," do not. The main advantage of the case-control study is that it enables us to study rare health outcomes without having to follow thousands of people, and is therefore generally quicker, cheaper and easier to conduct than the cohort study.

One primary disadvantage of a case-control study is a greater potential for bias. Since the health status is known before the exposure is determined, the study doesn't allow for broader-based health assessments, because only one type of disease has been selected for study. If the condition we wish to study is rare — for instance, affecting less than 5 percent of the population — the cohort approach would not identify enough subjects from which to draw statistically reliable inferences, unless we looked at a very large number of subjects. For example, in our study of low birthweight, if between 500 and several thousand births would be needed to get 10 to 50 low-weight births, developing exposure information for all of those births would be cumbersome, expensive and time-consuming.

In the case-control design, we can review birth certificates of several thousand newborns and find a certain number that exhibited low birthweight — 50 to 100, for example — and a comparable number of normal birthweights, and compare them with respect to electric and magnetic field exposures. This approach has the advantage of identifying a sufficient number of cases of the rare outcome we wish to study out of a population of thousands of births. It then requires us to develop exposure and risk factor data only for the limited number of individuals in our study. Thus, it is quicker, easier and less expensive than the cohort design, which would require such information for all of the several thousand births. It is an approach commonly used for studies of cancers and other rare diseases. Also, because subjects were selected on the basis of outcome only, we can evaluate a variety of exposures, such as electric fields, magnetic fields, chemical exposures and so forth.

The case-control study has the disadvantage of selecting cases and controls after both the outcome and the assumption of risk have occurred. This makes substantial bias a possibility because we may inadvertently favor certain births for inclusion in our study, and because certain women who should have been eligible for our study were not (those pregnant mothers whose fetus spontaneously aborted, for instance). Once chosen on the basis of one outcome

(low birthweight), our subjects cannot be analyzed for certain other outcomes (spontaneous abortion), as they could in a cohort study.

Another consideration in choosing an epidemiological design is the commonness of the risk factor. Common exposures can be studied by either the cohort or case-control design. Rare exposures are best studied by the cohort method since groups are selected on the basis of their exposure status.

Consider our study at hand. Most of us are exposed at home to very low magnetic fields — under 10 milliGauss (mG). But some homes score as high as 40 mG, or higher, and some occupations measure exposures in the hundreds or thousands of mG. Let's define exposed houses as those having at least 10 mG. If, in our case-control study of low birthweight births, we were to compare the residential magnetic field exposure of the births, we likely would see few, if any, exposed houses and would not be able to draw any conclusions. On the other hand, if we conducted a cohort study, we could select houses with magnetic field exposures over 10 mG, and then compare the birthweights of babies in those houses with birthweights of babies in houses with magnetic fields less than 10 mG. Since low birthweight is far more common than high magnetic field exposure, the cohort design is more likely to produce a useful result.

Occupational Epidemiological Study

The occupational study can be designed using any standard epidemiologic design, simply selecting working people with particular jobs or exposures as subjects. The main advantage of this approach is that workers often have substantially higher exposures to certain risk factors than the typical population, which increases our chances of detecting an effect if one truly exists. The main disadvantages are that workers with various jobs differ substantially from one another in terms of risks, and that the working population is substantially different from the nonworking one (such the rich, elderly or disabled), making it difficult to generalize to populations with some nonworking people. We usually look to occupational settings to exploit situations of high exposure. The number of eligible subjects in these settings is smaller than in the general population, but that is more than balanced by the extreme levels of exposures often seen in the workplace, which increase our chances of seeing effects.

There are two caveats to occupational epidemiological studies. First, in the workplace, people are exposed to a variety of risk factors that may affect results. For example, many workers are exposed to a variety of chemicals (such as solvents) that are known or suspected carcinogens; to a variety of electric and magnetic fields at different intensities and frequencies; and to other factors such as stress and poor ventilation or air quality.

Second, the number of people exposed to the risk factor we're interested in may be much smaller in the workplace than in residences (for example, fewer workers service an electric power transmission line than live near the line), so in some cases it may be difficult to identify a sufficient number of exposed workers. And it may be difficult to identify a comparison population of workers not exposed to the risk factor (or exposed at a substantially lower level) who also have comparable characteristics with respect to other possible risks. For instance, we would not compare construction workers to business executives because of likely differences in lifestyle and occupational risk. We would expect to see differences in cancer rates between the two groups based on those factors alone, and it would be difficult to separate those risks from our primary interest in the study of exposures to electric and

magnetic fields. It is preferable to examine cancer rates among groups of similar workers, say among all telephone line repair workers, pole climbers, van drivers, dispatchers and supervisors. That is still not ideal, as many of these workers perform more than one task, but the approach is better than comparing telephone workers with business executives.

Cross-Sectional Study

The "Am I like my neighbors?" study compares groups in terms of their current health and exposure status and assesses their similarities. The main advantage is that the cross-sectional study is a particularly easy study to conduct, as we do not have to wait for the health outcome to occur or estimate what the level of exposure was likely to have been years ago. Its main disadvantage is that a cause can't be inferred, because only current health and exposure are being studied.

The cross-sectional study is the one in which we assess a group's health status and exposure status simultaneously. We might inquire about recent health problems (including breast cancer diagnosed in the past year) and assess the current electric and magnetic fields exposures in people's homes as part of the same survey. An important limitation of this approach is that it does not allow for changes over time, and thus cannot accommodate diseases that take time to develop. For example, someone exposed today to ionizing radiation may be diagnosed with leukemia (or another cancer) five, 10, or even 20 years from now, even though the leukemia may not be evident today, next week or next month. Therefore, much information may be lost by contemporaneous evaluation.

Implicit in the cross-sectional study is the assumption that the study population has been exposed for a long time and will continue to be exposed unless some intervention is effected. Although such studies can be used to identify possible associations and suggest worthwhile case-control or cohort studies for follow up, the cross-sectional study may not necessarily confirm causes.

Problems in Conducting Epidemiological Studies

By Daniel Wartenberg, FACSNET Scholar
Doug Ramsey and John Warner, FACSNET Editors; Doris Ober, FACSNET Technical Editor

Posted April 23, 1996; revised Jan. 24, 2000

Like any analytic methodology, epidemiology has its hazards. Following are discussions of the four common problems epidemiologists face in conducting their studies and interpreting results:

Selection Bias

Bias is a technical term for playing favorites in choosing study subjects or in assessing their exposure or disease status.

Once a study question has been formulated and study subjects identified, it is important that these subjects be recruited uniformly and that data about the health and exposure be collected

consistently. If certain subjects are not enrolled in the study, or if information is collected differently from different subjects, the resulting bias could invalidate the study. There are many different forms of bias.

Selection bias can occur when not everyone eligible to be in a study can be selected as a subject, and when those selected are different from those excluded, in a systematic way. If we compare leukemia rates in male telephone workers of a certain age to the overall rate of leukemia in U.S. workers of the same age, we will probably find that the telephone workers have the lower rate of disease. If we compare leukemia rates in these telephone workers, however, to occurrences of leukemia among electric utility workers of the same age, there probably won't be a difference. This is because workers in jobs that require physical exertion, such as phone and utility workers, are apt to be healthier than more sedentary workers such as office workers. In this example, both telephone and utility workers have similar physical requirements.

To avoid selection bias we must select comparison populations that are similar except for the specific factors under study, and that's often difficult to do.

Recall Bias

Sometimes the subjects' ability to recall and report past experiences may be affected by their preconceived ideas about a possible health hazard. For example, residents of a town, annoyed at the noxious smell of a local toxic waste site, demand action. Epidemiologists are asked to conduct a health status survey to compare the rate of health problems among those who live near the site and those who live farther away.

When we ask subjects how often in the past year they experienced headaches, coughs or colds, those who believe that the site is the cause of their ill health are likely to be able to document many such occurrences. These anti-waste site folks may even remember that every time they felt ill they had been particularly bothered by the smell of the site. The smell of the site, in fact, may have made them more aware than they otherwise would be of each and every minor illness. It gave them something to blame. Subjects who do not blame the site for their health status, may have experienced exactly the same number of minor illnesses, but since they did not associate their headaches with the site, they may not have paid the headache much attention or made mental notes of when those headaches occurred.

To compensate for recall difference, the investigator may ask study subjects if they believe the site may be responsible for their problems, and then compare the rates of disease among those who do believe it and those who don't to determine if bias exists. If not, we can analyze the data directly. If so, we must adjust for it. Bias also can occur in many other aspects of data collection, and failure to prevent, accommodate or adjust for bias can invalidate studies.

Misclassification

Misclassification is a technical term for mislabeling or mischaracterizing a study subject, and may occur with disease or exposure. For example, patients who die of cancer are often misclassified on death certificates. They may have one type of cancer, but if the cancer spreads, they may ultimately die of something else — another type of cancer, or pneumonia, or heart failure, for instance. Depending on the physician, either may be reported on the death certificate — but unless the original cancer is recorded, the subject has been misclassified and

the study results will be skewed.

Similarly, one can misclassify exposure. For example, in most studies of the health effects of cigarette smoking, exposure is determined by asking subjects if they smoke, and if so, how many cigarettes per day. Typically, people who recognize that smoking is undesirable underestimate the amount they smoke. Some even claim to be nonsmokers. When we classify these people as light smokers or nonsmokers, we unintentionally mislabel them.

If we are studying whether people who take care of their teeth get fewer cavities, we might ask people how often they brush their teeth. Those who brush regularly probably won't lie about it. Those who do not brush regularly may exaggerate how often they brush, since dentists tell us we are supposed to brush at least twice a day. Based on this incorrect data, we would mislabel some irregular brushers as regular brushers, and the results would show that a surprisingly small difference in brushing frequency makes a big difference in the number of cavities, a stronger effect than is really there.

Confounding

Confounding is the technical term for finding an association for the wrong reason. It is associated with both the risk factor and the disease being studied, but need not be a risk factor for the disease under study. The confounding variable can either inflate or deflate the true relative risk.

For example, in two studies conducted in Denver, Colo., investigators found that children with leukemia were more likely to live in areas with large electric power lines than children without cancer. Critics argued that the true risk might be traffic density and exposure to traffic-related air pollutants. In fact, areas that are more developed and therefore more populated have both higher traffic densities and larger electric lines. Both factors were related, in other words, and both were found to be associated with the disease.

To resolve this confusion, investigators looked at both factors simultaneously. In the combined analysis in which homes with similar traffic density were compared, the association between power line size and leukemia persisted. In the combined analysis, however, in which homes with similar power line size were compared, the association between traffic density and leukemia did not persist. Epidemiologists could call traffic density the confounder and power line size the possible risk factor for disease. Research is still inconclusive on whether living near power lines causes disease, or whether living in areas of high-traffic density causes disease. More data are needed to confirm any risk factor. Many possible factors are being evaluated and power line size may itself be a confounder.

The presence of confounding in epidemiological studies is both a common and important phenomenon. Many, many variables may be confounders in any given study. Some of their effects may be small, others may be large. Failure to account for the most important confounders may cause investigators to question the validity of the results obtained. Given the large number of such factors, it is never possible to account for all potential confounders.

Statistical Variation

Statistical variation is the technical term for chance fluctuations. Here's an example of statistical variation: we flip an evenly weighted coin 10 times. We don't expect always to get

five heads and five tails. Nor do we expect to get 10 heads and no tails. More likely, the ratios will be four and six, or maybe seven and three. But, if we flip the coin 10,000 times, we would expect to get nearly 5,000 heads and 5,000 tails.

Similarly, even if disease rates in two populations (say 1 million people each) are identical, they may not appear so in a study. If we study two identical populations with identical risk factors and exposures, and pick a sample of people for our analyses (for example, 1,000 from each), it likely will turn out that the disease rates measured will be similar to one another but not identical. As the number of people in our study increases (to 10,000 subjects each), we expect our study-based estimates of the true disease rates in the entire populations to be more accurate.

Used with the permission of FACSNET.org

Glossary

Centers for Disease Control and Prevention

A

AGE-ADJUSTED MORTALITY RATE. A mortality rate statistically modified to eliminate the effect of different age distributions in the different populations.

AGENT. A factor, such as a microorganism, chemical substance, or form of radiation, whose presence, excessive presence, or (in deficiency diseases) relative absence is essential for the occurrence of a disease.

AGE-SPECIFIC MORTALITY RATE. A mortality rate limited to a particular age group. The numerator is the number of deaths in that age group; the denominator is the number of persons in that age group in the population.

ANALYTIC EPIDEMIOLOGY. The aspect of epidemiology concerned with the search for health-related causes and effects. Uses comparison groups, which provide baseline data, to quantify the association between exposures and outcomes, and test hypotheses about causal relationships.

ANALYTIC STUDY. A comparative study intended to identify and quantify associations, test hypotheses, and identify causes. Two common types are cohort study and case-control study.

APPLIED EPIDEMIOLOGY. The application or practice of epidemiology to address public health issues.

ASSOCIATION. Statistical relationship between two or more events, characteristics, or other variables.

ATTACK RATE. A variant of an incident rate, applied to a narrowly defined population observed for a limited period of time, such as during an epidemic.

ATTRIBUTABLE PROPORTION. A measure of the public health impact of a causative factor; proportion of a disease in a group that is exposed to a particular factor which can be attributed to their exposure to that factor.

B

BAR CHART. A visual display of the size of the different categories of a variable. Each category or value of the variable is represented by a bar.

BIAS. Deviation of results or inferences from the truth, or processes leading to such systematic deviation. Any trend in the collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth.

BIOLOGIC TRANSMISSION. The indirect vector-borne transmission of an infectious agent in which the agent undergoes biologic changes within the vector before being transmitted to a new host.

BOX PLOT. A visual display that summarizes data using a ``box and whiskers'' format to show the minimum and maximum values (ends of the whiskers), interquartile range (length of the box), and median (line through the box).

C

CARRIER. A person or animal without apparent disease who harbors a specific infectious agent and is capable of transmitting the agent to others. The carrier state may occur in an individual with an infection that is inapparent throughout its course (known as asymptomatic carrier), or during the incubation period, convalescence, and postconvalescence of an individual with a clinically recognizable disease. The carrier state may be of short or long duration (transient carrier or chronic carrier).

CASE. In epidemiology, a countable instance in the population or study group of a particular disease, health disorder, or condition under investigation. Sometimes, an individual with the particular disease.

CASE-CONTROL STUDY. A type of observational analytic study. Enrollment into the study is based on presence (``case'') or absence (``control'') of disease. Characteristics such as previous exposure are then compared between cases and controls.

CASE DEFINITION. A set of standard criteria for deciding whether a person has a particular disease or health-related condition, by specifying clinical criteria and limitations on time, place, and person.

CASE-FATALITY RATE. The proportion of persons with a particular condition (cases) who die from that condition. The denominator is the number of incident cases; the numerator is the number of cause-specific deaths among those cases.

CAUSE OF DISEASE. A factor (characteristic, behavior, event, etc.) that directly influences the occurrence of disease. A reduction of the factor in the population should lead to a reduction in the occurrence of disease.

CAUSE-SPECIFIC MORTALITY RATE. The mortality rate from a specified cause for a population. The numerator is the number of deaths attributed to a specific cause during a specified time interval; the denominator is the size of the population at the midpoint of the time interval.

CENSUS. The enumeration of an entire population, usually with details being recorded on residence, age, sex, occupation, ethnic group, marital status, birth history, and relationship to head of household.

CHAIN OF INFECTION. A process that begins when an agent leaves its reservoir or host through a portal of exit, and is conveyed by some mode of transmission, then enters through an appropriate portal of entry to infect a susceptible host.

CLASS INTERVAL. A span of values of a continuous variable which are grouped into a single category for a frequency distribution of that variable.

CLUSTER. An aggregation of cases of a disease or other health-related condition, particularly cancer and birth defects, which are closely grouped in time and place. The number of cases may or may not exceed the expected number; frequently the expected number is not known.

COHORT. A well-defined group of people who have had a common experience or exposure, who are then followed up for the incidence of new diseases or events, as in a cohort or prospective study. A group of people born during a particular period or year is called a birth cohort.

COHORT STUDY. A type of observational analytic study. Enrollment into the study is based on exposure characteristics or membership in a group. Disease, death, or other health-related outcomes are then ascertained and compared.

COMMON SOURCE OUTBREAK. An outbreak that results from a group of persons being exposed to a common noxious influence, such as an infectious agent or toxin. If the group is exposed over a relatively brief period of time, so that all cases occur within one incubation period, then the common source outbreak is further classified as a point source outbreak. In some common source outbreaks, persons may be exposed over a period of days, weeks, or longer, with the exposure being either intermittent or continuous.

CONFIDENCE INTERVAL. A range of values for a variable of interest, e.g., a rate, constructed so that this range has a specified probability of including the true value of the variable. The specified probability is called the confidence level, and the end points of the confidence interval are called the confidence limits.

CONFIDENCE LIMIT. The minimum or maximum value of a confidence interval.

CONTACT. Exposure to a source of an infection, or a person so exposed.

CONTAGIOUS. Capable of being transmitted from one person to another by contact or close

proximity.

CONTINGENCY TABLE. A two-variable table with cross-tabulated data.

CONTROL. In a case-control study, comparison group of persons without disease.

CRUDE MORTALITY RATE. The mortality rate from all causes of death for a population.

CUMULATIVE FREQUENCY. In a frequency distribution, the number or proportion of cases or events with a particular value or in a particular class interval, plus the total number or proportion of cases or events with smaller values of the variable.

CUMULATIVE FREQUENCY CURVE. A plot of the cumulative frequency rather than the actual frequency for each class interval of a variable. This type of graph is useful for identifying medians, quartiles, and other percentiles.

D

DEATH-TO-CASE RATIO. The number of deaths attributed to a particular disease during a specified time period divided by the number of new cases of that disease identified during the same time period.

DEMOGRAPHIC INFORMATION. The ``person'' characteristics--age, sex, race, and occupation--of descriptive epidemiology used to characterize the populations at risk.

DENOMINATOR. The lower portion of a fraction used to calculate a rate or ratio. In a rate, the denominator is usually the population (or population experience, as in person-years, etc.) at risk.

DEPENDENT VARIABLE. In a statistical analysis, the outcome variable(s) or the variable(s) whose values are a function of other variable(s) (called independent variable(s) in the relationship under study).

DESCRIPTIVE EPIDEMIOLOGY. The aspect of epidemiology concerned with organizing and summarizing health-related data according to time, place, and person.

DETERMINANT. Any factor, whether event, characteristic, or other definable entity, that brings about change in a health condition, or in other defined characteristics.

DIRECT TRANSMISSION. The immediate transfer of an agent from a reservoir to a susceptible host by direct contact or droplet spread.

DISTRIBUTION. In epidemiology, the frequency and pattern of health-related characteristics and events in a population. In statistics, the observed or theoretical frequency of values of a variable.

DOT PLOT. A visual display of the actual data points of a noncontinuous variable.

DROPLET NUCLEI. The residue of dried droplets that may remain suspended in the air for long periods, may be blown over great distances, and are easily inhaled into the lungs and exhaled.

DROPLET SPREAD. The direct transmission of an infectious agent from a reservoir to a susceptible host by spray with relatively large, short-ranged aerosols produced by sneezing, coughing, or talking.

E

ENDEMIC DISEASE. The constant presence of a disease or infectious agent within a given geographic area or population group; may also refer to the usual prevalence of a given disease within such area or group.

ENVIRONMENTAL FACTOR. An extrinsic factor (geology, climate, insects, sanitation, health services, etc.) which affects the agent and the opportunity for exposure.

EPIDEMIC. The occurrence of more cases of disease than expected in a given area or among a specific group of people over a particular period of time.

EPIDEMIC CURVE. A histogram that shows the course of a disease outbreak or epidemic by plotting the number of cases by time of onset.

EPIDEMIC PERIOD. A time period when the number of cases of disease reported is greater than expected.

EPIDEMIOLOGIC TRIAD. The traditional model of infectious disease causation. Includes three components: an external agent, a susceptible host, and an environment that brings the host and agent together, so that disease occurs.

EPIDEMIOLOGY. The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems.

EVALUATION. A process that attempts to determine as systematically and objectively as possible the relevance, effectiveness, and impact of activities in the light of their objectives.

EXPERIMENTAL STUDY. A study in which the investigator specifies the exposure category for each individual (clinical trial) or community (community trial), then follows the individuals or community to detect the effects of the exposure.

EXPOSED (GROUP). A group whose members have been exposed to a supposed cause of disease or health state of interest, or possess a characteristic that is a determinant of the health outcome of interest.

F

FREQUENCY DISTRIBUTION. A complete summary of the frequencies of the values or categories of a variable; often displayed in a two column table: the left column lists the individual values or categories, the right column indicates the number of observations in each category.

FREQUENCY POLYGON. A graph of a frequency distribution with values of the variable on the x-axis and the number of observations on the y-axis; data points are plotted at the midpoints of the intervals and are connected with a straight line.

G

GRAPH. A way to show quantitative data visually, using a system of coordinates.

H

HEALTH. A state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.

HEALTH INDICATOR. A measure that reflects, or indicates, the state of health of persons in a defined population, e.g., the infant mortality rate.

HEALTH INFORMATION SYSTEM. A combination of health statistics from various sources, used to derive information about health status, health care, provision and use of services, and impact on health.

HIGH-RISK GROUP. A group in the community with an elevated risk of disease.

HISTOGRAM. A graphic representation of the frequency distribution of a continuous variable. Rectangles are drawn in such a way that their bases lie on a linear scale representing different intervals, and their heights are proportional to the frequencies of the values within each of the intervals.

HOST. A person or other living organism that can be infected by an infectious agent under natural conditions.

HOST FACTOR. An intrinsic factor (age, race, sex, behaviors, etc.) which influences an individual's exposure, susceptibility, or response to a causative agent.

HYPERENDEMIC DISEASE. A disease that is constantly present at a high incidence and/or prevalence rate.

HYPOTHESIS. A supposition, arrived at from observation or reflection, that leads to refutable predictions. Any conjecture cast in a form that will allow it to be tested and refuted.

HYPOTHESIS, NULL. The first step in testing for statistical significance in which it is assumed that the exposure is not related to disease.

HYPOTHESIS, ALTERNATIVE. The hypothesis, to be adopted if the null hypothesis proves implausible, in which exposure is associated with disease.

I

IMMUNITY, ACTIVE. Resistance developed in response to stimulus by an antigen (infecting agent or vaccine) and usually characterized by the presence of antibody produced by the host.

IMMUNITY, HERD. The resistance of a group to invasion and spread of an infectious agent, based on the resistance to infection of a high proportion of individual members of the group. The resistance is a product of the number susceptible and the probability that those who are susceptible will come into contact with an infected person.

IMMUNITY, PASSIVE. Immunity conferred by an antibody produced in another host and acquired naturally by an infant from its mother or artificially by administration of an antibody-containing preparation (antiserum or immune globulin).

INCIDENCE RATE. A measure of the frequency with which an event, such as a new case of illness, occurs in a population over a period of time. The denominator is the population at risk; the numerator is the number of new cases occurring during a given time period.

INCUBATION PERIOD. A period of subclinical or inapparent pathologic changes following exposure, ending with the onset of symptoms of infectious disease.

INDEPENDENT VARIABLE. An exposure, risk factor, or other characteristic being observed or measured that is hypothesized to influence an event or manifestation (the dependent variable).

INDIRECT TRANSMISSION. The transmission of an agent carried from a reservoir to a susceptible host by suspended air particles or by animate (vector) or inanimate (vehicle) intermediaries.

INDIVIDUAL DATA. Data that have not been put into a frequency distribution or rank ordered.

INFECTIVITY. The proportion of persons exposed to a causative agent who become infected by an infectious disease.

INFERENCE, STATISTICAL. In statistics, the development of generalizations from sample data, usually with calculated degrees of uncertainty.

INTERQUARTILE RANGE. The central portion of a distribution, calculated as the difference between the third quartile and the first quartile; this range includes about one-half of the observations in the set, leaving one-quarter of the observations on each side.

L

LATENCY PERIOD. A period of subclinical or inapparent pathologic changes following

exposure, ending with the onset of symptoms of chronic disease.

M

MEAN, ARITHMETIC. The measure of central location commonly called the average. It is calculated by adding together all the individual values in a group of measurements and dividing by the number of values in the group.

MEAN, GEOMETRIC. The mean or average of a set of data measured on a logarithmic scale.

MEASURE OF ASSOCIATION. A quantified relationship between exposure and disease; includes relative risk, rate ratio, odds ratio.

MEASURE OF CENTRAL LOCATION. A central value that best represents a distribution of data. Measures of central location include the mean, median, and mode. Also called the measure of central tendency.

MEASURE OF DISPERSION. A measure of the spread of a distribution out from its central value. Measures of dispersion used in epidemiology include the interquartile range, variance, and the standard deviation.

MEDIAN. The measure of central location which divides a set of data into two equal parts.

MEDICAL SURVEILLANCE. The monitoring of potentially exposed individuals to detect early symptoms of disease.

MIDRANGE. The halfway point or midpoint in a set of observations. For most types of data, it is calculated as the sum of the smallest observation and the largest observation, divided by two. For age data, one is added to the numerator. The midrange is usually calculated as an intermediate step in determining other measures.

MODE. A measure of central location, the most frequently occurring value in a set of observations.

MORBIDITY. Any departure, subjective or objective, from a state of physiological or psychological well-being.

MORTALITY RATE. A measure of the frequency of occurrence of death in a defined population during a specified interval of time.

MORTALITY RATE, INFANT. A ratio expressing the number of deaths among children under one year of age reported during a given time period divided by the number of births reported during the same time period. The infant mortality rate is usually expressed per 1,000 live births.

MORTALITY RATE, NEONATAL. A ratio expressing the number of deaths among children from birth up to but not including 28 days of age divided by the number of live births reported during the same time period. The neonatal mortality rate is usually expressed per 1,000 live births.

MORTALITY RATE, POSTNEONATAL. A ratio expressing the number of deaths among children from 28 days up to but not including 1 year of age during a given time period divided by the number of lives births reported during the same time period. The postneonatal mortality rate is usually expressed per 1,000 live births.

N

NATURAL HISTORY OF DISEASE. The temporal course of disease from onset (inception) to resolution.

NECESSARY CAUSE. A causal factor whose presence is required for the occurrence of the effect (of disease).

NOMINAL SCALE. Classification into unordered qualitative categories; e.g., race, religion, and country of birth as measurements of individual attributes are purely nominal scales, as there is no inherent order to their categories.

NORMAL CURVE. A bell-shaped curve that results when a normal distribution is graphed.

NORMAL DISTRIBUTION. The symmetrical clustering of values around a central location. The properties of a normal distribution include the following: (1) It is a continuous, symmetrical distribution; both tails extend to infinity; (2) the arithmetic mean, mode, and median are identical; and, (3) its shape is completely determined by the mean and standard deviation.

NUMERATOR. The upper portion of a fraction.

O

OBSERVATIONAL STUDY. Epidemiological study in situations where nature is allowed to take its course. Changes or differences in one characteristic are studied in relation to changes or differences in others, without the intervention of the investigator.

ODDS RATIO. A measure of association which quantifies the relationship between an exposure and health outcome from a comparative study; also known as the cross-product ratio.

ORDINAL SCALE. Classification into ordered qualitative categories; e.g., social class (I, II, III, etc.), where the values have a distinct order, but their categories are qualitative in that there is no natural (numerical) distance between their positive values.

OUTBREAK. Synonymous with epidemic. Sometimes the preferred word, as it may escape sensationalism associated with the word epidemic. Alternatively, a localized as opposed to generalized epidemic.

P

PANDEMIC. An epidemic occurring over a very wide area (several countries or continents) and usually affecting a large proportion of the population.

PATHOGENICITY. The proportion of persons infected, after exposure to a causative agent, who then develop clinical disease.

PERCENTILE. The set of numbers from 0 to 100 that divide a distribution into 100 parts of equal area, or divide a set of ranked data into 100 class intervals with each interval containing 1/100 of the observations. A particular percentile, say the 5th percentile, is a cut point with 5 percent of the observations below it and the remaining 95% of the observations above it.

PERIOD PREVALENCE. The amount a particular disease present in a population over a period of time.

PERSON-TIME RATE. A measure of the incidence rate of an event, e.g., a disease or death, in a population at risk over an observed period to time, that directly incorporates time into the denominator.

PIE CHART. A circular chart in which the size of each ``slice'' is proportional to the frequency of each category of a variable.

POINT PREVALENCE. The amount of a particular disease present in a population at a single point in time.

POPULATION. The total number of inhabitants of a given area or country. In sampling, the population may refer to the units from which the sample is drawn, not necessarily the total population of people.

PREDICTIVE VALUE POSITIVE. A measure of the predictive value of a reported case or epidemic; the proportion of cases reported by a surveillance system or classified by a case definition which are true cases.

PREVALENCE. The number or proportion of cases or events or conditions in a given population.

PREVALENCE RATE. The proportion of persons in a population who have a particular disease or attribute at a specified point in time or over a specified period of time.

PROPAGATED OUTBREAK. An outbreak that does not have a common source, but instead spreads from person to person.

PROPORTION. A type of ratio in which the numerator is included in the denominator. The ratio of a part to the whole, expressed as a ``decimal fraction'' (e.g., 0.2), as a fraction (1/5), or, loosely, as a percentage (20%).

PROPORTIONATE MORTALITY. The proportion of deaths in a specified population over a period of time attributable to different causes. Each cause is expressed as a percentage of all deaths, and the sum of the causes must add to 100%. These proportions are not mortality

rates, since the denominator is all deaths, not the population in which the deaths occurred.

PUBLIC HEALTH SURVEILLANCE. The systematic collection, analysis, interpretation, and dissemination of health data on an ongoing basis, to gain knowledge of the pattern of disease occurrence and potential in a community, in order to control and prevent disease in the community.

R

RACE-SPECIFIC MORTALITY RATE. A mortality rate limited to a specified racial group. Both numerator and denominator are limited to the specified group.

RANDOM SAMPLE. A sample derived by selecting individuals such that each individual has the same probability of selection.

RANGE. In statistics, the difference between the largest and smallest values in a distribution. In common use, the span of values from smallest to largest.

RATE. An expression of the frequency with which an event occurs in a defined population.

RATE RATIO. A comparison of two groups in terms of incidence rates, person-time rates, or mortality rates.

RATIO. The value obtained by dividing one quantity by another.

RELATIVE RISK. A comparison of the risk of some health-related event such as disease or death in two groups.

REPRESENTATIVE SAMPLE. A sample whose characteristics correspond to those of the original population or reference population.

RESERVOIR. The habitat in which an infectious agent normally lives, grows and multiplies; reservoirs include human reservoirs, animals reservoirs, and environmental reservoirs.

RISK. The probability that an event will occur, e.g. that an individual will become ill or die within a stated period of time or age.

RISK FACTOR. An aspect of personal behavior or lifestyle, an environmental exposure, or an inborn or inherited characteristic that is associated with an increased occurrence of disease or other health-related event or condition.

RISK RATIO. A comparison of the risk of some health-related event such as disease or death in two groups.

S

SAMPLE. A selected subset of a population. A sample may be random or non-random and it may be representative or non-representative.

SCATTER DIAGRAM. A graph in which each dot represents paired values for two continuous variables, with the x-axis representing one variable and the y-axis representing the other; used to display the relationship between the two variables; also called a scattergram.

SEASONALITY. Change in physiological status or in disease occurrence that conforms to a regular seasonal pattern.

SECONDARY ATTACK RATE. A measure of the frequency of new cases of a disease among the contacts of known cases.

SECULAR TREND. Changes over a long period of time, generally years or decades.

SENSITIVITY. The ability of a system to detect epidemics and other changes in disease occurrence. The proportion of persons with disease who are correctly identified by a screening test or case definition as having disease.

SENTINEL SURVEILLANCE. A surveillance system in which a pre-arranged sample of reporting sources agrees to report all cases of one or more notifiable conditions.

SEX-SPECIFIC MORTALITY RATE. A mortality rate among either males or females.

SKEWED. A distribution that is asymmetrical.

SPECIFICITY. The proportion of persons without disease who are correctly identified by a screening test or case definition as not having disease.

SPORADIC. A disease that occurs infrequently and irregularly.

SPOT MAP. A map that indicates the location of each case of a rare disease or outbreak by a place that is potentially relevant to the health event being investigated, such as where each case lived or worked.

STANDARD DEVIATION. The most widely used measure of dispersion of a frequency distribution, equal to the positive square root of the variance.

STANDARD ERROR (OF THE MEAN). The standard deviation of a theoretical distribution of sample means about the true population mean.

SUFFICIENT CAUSE. A causal factor or collection of factors whose presence is always followed by the occurrence of the effect (of disease).

SURVEILLANCE. see PUBLIC HEALTH SURVEILLANCE

SURVIVAL CURVE. A curve that starts at 100% of the study population and shows the percentage of the population still surviving at successive times for as long as information is available. May be applied not only to survival as such, but also to the persistence of freedom

from a disease, or complication or some other endpoint.

T

TABLE. A set of data arranged in rows and columns.

TABLE SHELL. A table that is complete except for the data.

TRANSMISSION OF INFECTION. Any mode or mechanism by which an infectious agent is spread through the environment or to another person.

TREND. A long-term movement or change in frequency, usually upwards or downwards.

U

UNIVERSAL PRECAUTIONS. Recommendations issued by CDC to minimize the risk of transmission of bloodborne pathogens, particularly HIV and HBV, by health care and public safety workers. Barrier precautions are to be used to prevent exposure to blood and certain body fluids of all patients.

V

VALIDITY. The degree to which a measurement actually measures or detects what it is supposed to measure.

VARIABLE. Any characteristic or attribute that can be measured.

VARIANCE. A measure of the dispersion shown by a set of observations, defined by the sum of the squares of deviations from the mean, divided by the number of degrees of freedom in the set of observations.

VECTOR. An animate intermediary in the indirect transmission of an agent that carries the agent from a reservoir to a susceptible host.

VEHICLE. An inanimate intermediary in the indirect transmission of an agent that carries the agent from a reservoir to a susceptible host.

VIRULENCE. The proportion of persons with clinical disease, who after becoming infected, become severely ill or die.

VITAL STATISTICS. Systematically tabulated information about births, marriages, divorces, and deaths, based on registration of these vital events.

Y

YEARS OF POTENTIAL LIFE LOST. A measure of the impact of premature mortality on a population, calculated as the sum of the differences between some predetermined minimum or desired life span and the age of death for individuals who died earlier than that predetermined age.

Z

ZOONOSES. An infectious disease that is transmissible under normal conditions from animals to humans.