

Epidemiology

Mark Schiffman, M.D.

Applications of Epidemiology to Pathology Studies 1301
 Prevalence, Incidence, and Mortality Rates of Disease 1302
 Geographic Differences and Time Trends in Disease Occurrence 1303
 Validating New (or Old) Histopathologic Diagnostic Distinctions 1304
 Judging Intra- or Interpathologist Agreement 1304
 Epidemiologic Studies of Disease Etiology 1304
 Follow-Up Studies of Patients with the Same Pathologic Diagnosis 1305

Randomized Clinical Trials 1306
 Screening for Gynecologic Malignancies 1307
Basic Statistical Concepts 1307
 Variability as a Fundamental Principle of Pathology 1307
 Error Versus Bias 1308
 Descriptive Data 1308
 The Basic Contingency Table 1309
 Measures of Risk (Absolute, Relative, and Attributable Risks) 1310
 Causal Intermediates and Surrogate Endpoints 1313
 Measures of Interpathologist Agreement 1313
 Screening Terms 1314

The Receiver-Operating Characteristic (ROC) Curve 1315

Problem Areas 1315
 Dividing a Spectrum of Disease into Categories 1315
 The Need for Masking 1316
 Standardization of the Scientific Art of Pathology 1316
 Specimen Adequacy Versus the Bias of Convenience Samples 1316
 Deciding How Large a Study to Do: Statistical Significance Versus Practicality 1317
 Incorporating Research into Pathology Practice 1317

Most pathologists are part-time epidemiologists as well. The two medical disciplines are more closely allied than many realize. Epidemiologists study the distribution and determinants of diseases in human populations. In current medical practices, diseases are often defined by histopathologic diagnoses or by clinical pathologic test values. Thus, whenever a pathologist shifts intellectually from the level of the individual slide or specimen to thinking about a group of diagnoses, an informal epidemiologic question is being raised. For example, "How common is this diagnosis?" is a question of prevalence or incidence. "Why am I seeing so many cases of this type of tumor?" is a question of time trends. "How would my colleague interpret these slides compared with me?" is a question of interpathologist agreement. And, "What causes this disease I am seeing every week?" is a question of etiology that can be addressed by pathologists working as epidemiologists, or with them.

This chapter is meant to introduce the major epidemiologic concepts of greatest use to patholo-

gists who are considering a research project, or who wish to think more formally at the population level about their case material or diagnostic criteria. The review is certainly not exhaustive; rather, it is meant to be quite informal and readable and to encourage the pathologist to pursue epidemiologic projects and collaborations. The first section, accordingly, is organized around types of possible epidemiologic studies that a pathologist might wish to pursue. The next section outlines nonmathematically a few basics of statistical thinking that pathologists need to know if they wish to do more formal epidemiologic research. The third section discusses a few problem issues that usually emerge when epidemiologists and pathologists work together.

Applications of Epidemiology to Pathology Studies

This section illustrates the types of epidemiologic projects that a pathologist may undertake, either

informally or formally. The examples are drawn mainly from the author's experience conducting etiologic and screening studies of gynecologic neoplasia, especially cervical neoplasia.

Throughout this section, epidemiologic terms will be introduced and simply defined. There is a useful dictionary of epidemiology for readers interested in learning more terminology.⁷ For a more complete understanding of basic epidemiologic concepts, the reader is referred to one of several introductory texts.^{2,5,10}

Prevalence, Incidence, and Mortality Rates of Disease

One of the first questions that an expert or novice epidemiologist is likely to ask about a disease under study is "How common is it?" The pathologist at the microscope is interested in how common various conditions are, as one element of differential diagnosis (witness the maxims, "Rare diseases occur rarely." and "If you hear hoofbeats, think of a horse not a zebra.")

For the pathologist considering a research study, the frequency of disease occurrence is crucial for two reasons. On the practical level, very rare conditions are difficult to study epidemiologically because the statistical principles underlying epidemiology require moderately large numbers to deal with chance, which is the unavoidable and defining characteristic of observational studies in humans.

More importantly, the amount of disease in a population is the starting point for epidemiologic thought, leading to all the major epidemiologic comparisons, such as "How much disease occurs in population A compared to population B, and what does the difference tell us? Why is the amount of disease changing over time? What risk factors are associated with groups having the most disease?"

Because measuring the occurrence of disease is so important to epidemiologists, they find it important, like skiers discussing snow, to define terms carefully using a resultant epidemiologic jargon (in the good sense of the word). A few key terms related to the frequency of disease occurrence are essential and worth memorizing by anyone interested in epidemiology.

The *prevalence* of a disease is the number of occurrences of the disease in a given population at a given time, for example, "Twenty percent of the patients seen in this clinic have at least reactive changes on their Papanicolaou smears." Often, prevalence is discussed with reference to a single point in time, as in a screening program, yielding a *point prevalence*:

"Two percent of the screening smears last month showed changes suggestive of CIN."

The *incidence* of disease is the number of new cases that develop in a given time period. Accordingly, *incident disease* refers to new disease whereas *prevalent disease* refers to all the cases in the population, whether new or chronic. The connection between prevalence and incidence is the *duration* of the condition (Prevalence = Incidence × Duration). Therefore, the prevalence of rabies in a given week is close to the incidence because duration is short, whereas the prevalence of a long-duration disease, such as rheumatoid arthritis much exceeds the incidence for any time period. A more subtle and relevant example of how incidence and prevalence relate via duration is the following. In studies of young women, we noted about 10 years ago that the point prevalence of human papillomavirus (HPV) infection was about the same as the yearly incidence, suggesting a duration of infection of approximately 1 year. In follow-up studies, we have now confirmed that HPV infections do last about a year.

Incidence is most often defined as a yearly rate, as in "36,000 incidence cases of uterine corpus cancer were diagnosed in the United States in 2000." However, *lifetime cumulative incidence* is also an intuitively useful term, meaning the estimated risk of occurrence of a disease over a woman's life: "About 1% of women in the United States will develop cervical cancer in their lifetime." For chronic diseases such as endometriosis, genital herpes, or specific gynecologic cancers, incidence is usually thought of as a one-time phenomenon; that is, second primaries rarely occur. (In contrast to second primaries, recurrences of the same disease imply that it is prevalent, not incidence.) For acute, self-limited, or curable conditions such as gonorrhea, incidence must be defined over a narrow range of time appropriate to the duration of the illness.

Rates of death from a disease are measured as the *mortality rate*. The connection between incidence and mortality is, of course, survival, measured often by the *case:fatality ratio*.

In summary, the epidemiologist is interested in the prevalence, incidence, and mortality rates of a disease as the fundamental basis of further study. These terms can be applied to any study population, whether that population is a single gynecologic practice or hospital, a city, a country, or the world.

National incidence and mortality data are most often cited when discussing the scope of a medical problem. Where can national data be obtained? In the United States, pathologists are probably aware that mortality rates from all causes are compiled

and available from a variety of sources, most notably and simply from the National Center for Health Statistics (6525 Belcrest Road, Room 1064, Hyattsville, MD 20782; 301-458-4636).

Mortality rates are usually the most reliable gauge of disease occurrence for highly fatal diseases when comparing different populations worldwide or time periods. Of course, there are obvious uncertainties and errors in ascribing causes of death, but mortality rates are reasonably well recorded and useful for many cancers. Mortality is not useful as a measurement of disease occurrence in regions where the diagnostic workups are lacking or the case: fatality ratio has been altered sharply by improved treatment.

In contrast, if a condition is not often fatal, mortality rates may not be useful at all for disease surveillance. It is often more difficult to obtain reliable incidence data, and the researcher must rely on data from voluntary registries, published surveys, or occasionally government-mandated registries. For cancer, fortunately, the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program compiles incidence rates for a (nonrandom but stable) 10% sample of the U.S. population. The most accessible source of SEER cancer incidence and survival data (as well as national cancer mortality data) is *CA—A Cancer Journal for Clinicians*, published annually by the American Cancer Society and mailed free on request.⁴ More detailed cancer data can be obtained from other American Cancer Society publications, such as *Cancer Facts and Figures*, or from the SEER program itself (National Cancer Institute, Bethesda, MD 20892, or <http://seer.cancer.gov>). The International Agency for Research on Cancer compiles international incidence and mortality rates derived from cancer registries of varying quality.^{9,18}

Geographic Differences and Time Trends in Disease Occurrence

Pathologists may wish to go beyond descriptions of disease occurrence at a place and time to compare rates between geographic areas or over time. The usual hope is that the comparisons may yield clues to etiology and pathogenesis. A cautious approach is critical because of the omnipresent effects of chance on observational data. How can one tell if the amount of disease in one place or time is truly different from the amount found earlier or elsewhere? Disease rates fluctuate over time and place. Many geographic differences and temporal trends do not persist over time, appearing random (to the limit of our understanding!).

Hence, there is a need for statistics as one of the disciplines underlying epidemiology. Distinguishing chance differences from true differences requires statistical thinking and an appreciation of the types of differences that arise by chance. This point is important, because overinterpretation of chance differences is one of the most common errors that novice epidemiologists make when comparing disease rates from one place or time to another. For example, many cancer "outbreaks" where several neighbors get similar tumors turn out to be quite explainable as chance clusterings of events, expected for common malignancies such as breast cancer. A good bit of advice might be to treat health statistics like the monthly economic news: it takes a long-term trend or a persistent difference to trust that something important is happening.

When comparing one place to another, or analyzing time trends, the cardinal rule is to make sure that the comparison is valid. A checklist of common-sense questions should be asked:

1. Are the rates being compared truly comparable (incidence, prevalence, mortality)? In particular, are the sources of data comparable (for example, a mandatory registry cannot be compared to a voluntary reporting system because of differences in the completeness of reporting).
2. Are the diagnostic criteria the same in both comparison groups? This particular problem has plagued the interpretation of time trend data regarding minor cervical cytologic abnormalities because increased recognition by pathologists of subtle koilocytotic changes cannot be easily distinguished from increased incidence of koilocytotic atypia.
3. Are the two populations comparable in age and other factors affecting risk of disease? No one would think of comparing the prevalence of cervical epithelial neoplasia (CIN) in a gynecologic referral practice to the prevalence in a screening clinic because, of course, the prevalence would be higher in the referral clinic. Some researchers, however, make the analogous mistake of comparing populations that differ with regard to age, socioeconomic status, or other more subtle characteristics related to the risk of disease (called *confounding variables* in epidemiologic jargon). Most importantly, almost all diseases vary in incidence and prevalence by age, and thus almost all comparisons should take age into account. The section on error and bias (following) mentions simple meth-

ods of adjustment for age and other confounding variables. The statistical bases of making geographic and temporal comparisons are covered in the sections on descriptive data and measures of risk.

Validating New (or Old) Histopathologic Diagnostic Distinctions

The creation and refinement of pathologic classifications can be aided by epidemiologic corroboration. For example, the Bethesda System of cervical cytology combines koilocytotic atypia and CIN 1 as low-grade squamous intraepithelial lesion (LSIL). This combination was supported by epidemiologic data. The two diagnoses, which are not reliably distinguishable on morphologic grounds, share the same epidemiologic profiles of younger average age and varied human papillomavirus (HPV) types, as compared to the older average age and restricted HPV types found in higher-grade lesions. As another example, a recent pathologic study of squamous vulvar cancer, which proposed new pathologic subtypes, was strengthened by a separate epidemiologic analysis showing that the new subtypes had different epidemiologic characteristics.⁶ Pathologists and epidemiologists can work iteratively to refine disease classifications, asking each other "Do categories X and Y look the same or different from your point of view?"

Judging Intra- or Interpathologist Agreement

Pathology agreement studies have been motivated by the needs of both disciplines. Pathologists are obviously concerned with the reliability of the diagnoses they make. Epidemiologists are concerned with uniform case definition in their studies. When comparisons of intra- and interpathologist agreement are performed, the epidemiologist can serve the role of scientific organizer, ensuring independence of the reviews by *masking* the reviewers (also called *blinding*) to each other's diagnoses until after the data are complete. It is the widespread opinion of epidemiologists that unmasked comparisons, in which reviewers have access to each other's diagnoses, have limited scientific value. Like all human beings, pathologists tend to agree much more in public than in private, and masking provides a guarantee that a comparison rather than a consensus is being achieved. In the area of cervical pathology, the diagnosis of CIN by either cytology or histology has proven much more variable among experts when masked comparisons were performed than initially

expected. Surprisingly, the extensive histologic reviews of specimens from loop electrosurgical excision procedures (LEEP) exhibit almost as much interpathologist variability as cytology.¹⁷ Many morphologic judgments that pathologists make regarding cancer precursor lesions are clearly difficult, regardless of tissue type and quantity.

Epidemiologic Studies of Disease Etiology

Epidemiologists attempt to find the determinants of disease by statistically correlating the presence or absence of possible *exposures* (often called *risk factors*) with the presence or absence of disease. Epidemiologic studies attempting to relate exposures and disease are called *analytic studies*, as distinguished from *descriptive studies* that yield rates of disease without directly addressing etiology.

A description of the many types of analytic studies is beyond the scope of this chapter. At the simplest level, *prospective* or *cohort studies* start with the measurement of an exposure in a group of study participants who are followed over time. The investigators then compare incidence rates or *absolute risk* of disease in the exposed versus the unexposed groups. The ratio of the incidence rate in exposed subjects divided by the incidence rate in the unexposed is called the *incidence rate ratio*. The reader might correctly expect that there are as many types of rate ratios as there are types of rates (e.g., *prevalence rate ratio*, *lifetime cumulative risk ratio*). Many epidemiologists casually refer to the entire group as the *relative risk* of exposed versus nonexposed subjects and use the abbreviation *RR* as a general shorthand.

Prospective studies are the most appealing type of analytic study because they most directly determine how commonly disease occurs in exposed versus unexposed individuals. The relative risk, directly measured, is an intuitively clear answer to the question: "If a woman has this characteristic (the exposure), how much more likely is she to develop the disease, compared to a similar unexposed woman?" The absolute risk translates as "How likely to get disease is an exposed woman?" (see later: "Measures of Risk"). The problem with prospective studies of cancer is that they are expensive, usually take years to organize and complete, and must be very large to generate enough cases of cancer for reliable estimates of risk, even for common tumors including *in situ* neoplasia.

Other analytic study designs try, in general, to estimate the relative risk estimates that might be obtained in the ideal prospective study, while saving time and money. Analytic studies that start by col-

lecting a series of *cases* (women diagnosed with a given disease) and appropriate *controls* (women without that disease who are measured for comparison) are called *case-control* studies. The exposures of interest are ascertained for both groups and the relative risk (RR) of disease among the exposed versus the unexposed is estimated by calculating the ratio of the odds of exposure in cases versus controls (for more explanation, see the statistical section on measures of risk).

The estimation of the prospective relative risk by the case-control *odds ratio* (OR) is one of the most important statistical concepts in epidemiology, and one of the most subtle. For this statistical approximation to be valid, incident case and controls must be chosen to be strictly comparable. The control group must represent the group of women at risk of developing disease at the time the incident case was diagnosed, otherwise the estimation of the relative risk can be grossly mistaken because of *bias* (a non-random or systematic error in estimation of a statistic, to be distinguished from *random error*).

In practice, it is very difficult to define and recruit an unbiased sample of the general population of women that gave rise to the cases appearing in one hospital or clinic. Thus, all kinds of compromises of convenience and practicality must be made, and it becomes difficult to avoid bias in choosing controls. For example, smoking causes or worsens so many kinds of illness that it is very difficult to use hospitalized controls to estimate the relative risk of a disease associated with smoking (such as many cancers). The exposure to smoking in the hospitalized controls is elevated compared to the general at-risk population; thus, the odds ratio obtained in a naively conducted hospital-based study tends to provide too low an estimate of the relative risk.

Because case-control studies are so commonly used as an analytic design, choosing proper controls is one of the two most important aspects of epidemiology. The other is assuring proper measurements of exposure and disease. The mark of a good epidemiologist is a dedicated attention to control selection and measurement error, whereas many novices tend to focus more on the cases and data analysis while relying on a *convenience sample* of whichever controls are most easily available.

Besides prospective and case-control studies, another common analytic study design is the *cross-sectional* study, in which exposure and disease status are ascertained concurrently for a study population. An example would be a screening study of HPV infection and abnormal cervical cytology, in which all women attending a clinic are tested for viral DNA at the same time the cytologic smear is

taken. The analysis of a cross-sectional study is somewhat similar to that of a case-control study, but the researcher must be careful because the cases are a combination of incident and prevalent disease. The odds ratio that is computed in a cross-sectional study is a good estimation of the prospective relative risk only if certain conditions are met, including an assumption that the disease under study is rare (an assumption not met for cervical cytologic abnormalities in many clinics).

The pathologist collaborator should play a key role in all analytic studies of diseases whose definitions rely on nonroutine pathologic expertise. Misclassification of disease status can be very damaging to a study because the result of misclassification on correlative statistics, like the relative risk and odds ratio, is, generally, to reduce the apparent strength of the association between disease and exposure. If the disease is defined poorly enough, no epidemiologic risk factors may be found even if they exist.¹³ Moreover, it is often very difficult to measure the risk factors (exposures) without substantial error, whether laboratory testing or interviews are being used. The combination of multiple errors in measuring both exposure and disease can literally make a study worthless. For example, early studies correlating HPV DNA detection and CIN revealed only a moderate association, in that less than 50% of cases were found to be HPV positive. Moreover, HPV infection was not apparently associated with sexual activity, an established strong risk factor for CIN. These weak associations were a result of misclassification. Subsequent studies with better HPV tests and expert review of pathology revealed that virtually all cases of CIN contain HPV DNA and that HPV is the sexually transmitted agent explaining the association of sexual activity and risk of CIN.

As the result of the strong, damaging effects of misclassification on epidemiologic studies, epidemiologists pay careful attention to the pathologic classifications that define their study cases and controls and often establish formal collaborations with reviewing pathologists as part of epidemiologic studies.

Follow-Up Studies of Patients with the Same Pathologic Diagnosis

Clinicians, pathologists, and epidemiologists are all interested in learning what happens to patients diagnosed with a given disease. For a possibly fatal disease, survival rates are critical, whereas for other chronic diseases progression rates are often estimated. It is often of interest to divide the patients into groups, to determine whether subtypes of disease follow different courses, or whether different

treatments influence outcome. The *randomized clinical trial* (see following) is a specialized version of such a follow-up study, in which subjects are randomly assigned to various treatment groups to maximize the comparability of the groups. The hope is that the randomization will minimize differences in both known and unknown confounding variables that could bias the comparison.

Follow-up studies almost invariably involve the concept of *time to an event*. In other words, it is important *when* incidence, progression, or death occur, not just *if* they occur. Let us discuss this issue in the context of studies of disease outcome (as opposed to disease incidence). Clearly, all participants in any follow-up study or clinical trial eventually die; the question is when (and why). A good treatment prolongs time to death whereas a bad type of disease shortens it. Because of the critical notion of "time to event" in epidemiologic follow-up studies, such studies depend heavily on actuarial methods, such as survival curves and life-table analyses, when comparing exposed to unexposed patients or treated to untreated patients. The central statistical concept in such studies is a kind of rate called a *hazard*, which refers to the risk of an outcome occurring in a unit of follow-up time. A hazard is computed as the number of *events* (e.g., death, cure, or progression) divided by the amount of *person-time* of follow-up. Person-time is computed individually for each participant as the observation time between her entry into the study and her exit. The total person-time for a study group is the sum of all the individual observation periods. For example, 10,000 women followed for a year or 100 women followed for 10 years both yield 1,000 person-years of follow-up time. Twenty deaths arising during that follow-up would yield an estimated hazard of 20 deaths per 1,000 person-years in both situations.

A hazard is a special kind of rate because it is conceived of as the rate of an outcome (disease incidence, progression, mortality, or whatever) at a single moment in time, as the mathematical "limit" of the rate as time "goes to zero." Accordingly, the hazard of disease can change from moment to moment as conditions change. An HPV-infected woman lights up a cigarette and her hazard for progression to invasive cervical cancer probably increases. She quits smoking the next day and her hazard decreases.

Moreover, the computation of the denominator of hazards, person-time of follow-up, requires some training and thought. For each successive time interval during follow-up, the denominator of women at risk for an event changes. For example, women are lost to follow-up as they drop out of the study, or they die for other reasons, or they experience the

event itself (because one can only progress for the first time or die once). Thus, computing the proper amount of person-time during which the events occurred requires some knowledge of *censoring*, which is the proper deletion of irrelevant follow-up time during which the subject was not truly at risk of the outcome.

It is useful to compute the hazard of conditions like death that happen once and do not reverse. Life-table methods are more confusing when a condition can come and go. For example, say we want to study "HPV infection" without defining specific types. However, the term "HPV infection" is like "a cold." Multiple types can present a confusing picture of clearance and "recurrence," with ambiguous meaning. Proper counting of events and censoring of person-time are very difficult in this context, making simpler analyses more appealing, such as the computation of cumulative incidence rate ratios (ever infected versus not infected over the course of study).

Usually, researchers are not content to describe the simple survival or progression curve of a disease after diagnosis. They wish to determine which factors affect the hazard, that is, what the relative or *proportional hazard* of death, etc., might be for women in different groups defined by pathologic differences or treatment types. The proportional hazard is almost identical to the incident rate ratio already discussed, but the denominator is person-time of follow-up, not just time. Proportional hazard analyses are too complex to be described here, and pathologists performing follow-up studies might consider consulting an epidemiologist or biostatistician early in the design phase of such projects. Data collection must be organized carefully to permit a correct determination of person-time.

Randomized Clinical Trials

Randomized clinical trials are conceptually simple prospective analyses, with eligible women divided into treatment arms. Randomization serves to balance known and unknown biases in the arms. There is a placebo or standard treatment arm, compared to one or more new treatment arms. These trials are very appealing as a court of judgment regarding best medical practice when "equipose" exists, that is, we do not know which practice is best. Such trials are highly influential. However, they are surprisingly difficult and should not be undertaken as quickly as observational studies. The maxim "Do no harm" is applicable because the participants' fate is influenced by the randomization. Clinician judgment as to the individual's optimal treatment must explicitly

be set aside. The stakes are always high when a trial is under way. We want a result quickly and definitively, but the two objectives are in conflict. The time-honored rules used to ensure fairness can seem bureaucratic and rigid. "Clinical trialists" are statisticians, epidemiologists, and others who specialize in randomization, data monitoring, intermediate analyses ("data peeks" before the scheduled end of follow-up), and stopping rules (in case the intermediate analyses reveal an especially good or bad outcome). Pathologists contribute expert review of entry diagnoses and outcomes. Often, the burden of pathology review is large, reproducibility is paramount, and the review process is highly controlled by central administration. Of all collaborations with epidemiologists, pathologists should be most wary of clinical trials because of the inevitable attendant requirements. Still, the rewards are vital.

Screening for Gynecologic Malignancies

Screening is inherently epidemiologic; thus, the pathologist involved in screening programs (e.g., cervical cytologic screening or CA-125 testing) needs to understand the interrelated concepts of sensitivity, specificity, and predictive value. The basics are outlined later in a statistical section on screening.

A common mistake in evaluating the results of a screening trial is to ignore the clinical setting. The *sensitivity* of a screening technique (percentage of diseased women who test positive) and its *specificity* (percentage of disease-free women who test negative) theoretically do not change when the test is taken out of a high-risk hospital clinic to be applied to the general population. But most clinicians are more interested in the *positive predictive value* and *negative predictive value*, two statistics that are highly dependent on the clinical setting. The positive predictive value is the percentage of women testing positive who truly have disease. The negative predictive value is the reassurance that disease does not exist given a negative test result.

Here is an important practical point: Given the same sensitivity and specificity (i.e., the same assay accuracy), positive predictive value decreases sharply as the prevalence of the disease decreases. True positives can be swamped by false positives once the test is applied to many normal women. Therefore, the same screening test that looks promising because of high sensitivity in a high-risk clinic will often perform poorly in the general population, producing so many false positives compared to the disease yield that the costs outweigh the benefits. As a general rule, specificity is perhaps surprisingly important as a requirement for a screening test. A screening test such as a

tumor marker must be highly specific (negative in virtually all nondiseased women, certainly more than 90%) to be cost-effective for general population screening.

Basic Statistical Concepts

Hopefully, the preceding discussion has firmly established the relevance of epidemiology to gynecologic pathology research and even daily practice. Epidemiologic work requires an understanding of biostatistics. This section presents the bare basics of what the author believes pathologists collaborating in epidemiologic research might wish to know about biostatistical methods. Introductory biostatistics texts are available and easy to read for the pathologists wishing to work independently or for those who want computational formulae for chi-square or other commonly used tests.

Variability as a Fundamental Principle of Pathology

Virtually all measurements that one could make about a human population are variable. Height, weight, fine points of anatomy, metabolic patterns, serum levels of hormones, and nutrients are all commonly recognized to be variable. The same variability is seen by pathologists at the tissue and cellular levels and by research pathologists at the molecular biologic level (e.g., varying tissue levels of DNA adducts given equivalent carcinogenic exposures, genetic polymorphisms in human genes, and varying molecular responses to infection with viral DNA). Even the intricate, multistep molecular pathways to cancer demonstrate substantial variability between individuals who develop the same type of malignancy.

Variability in pathology is mainly described by *categorical* or *discrete* data and statistics, as compared with *continuous* data and statistics (the province of the mean, median, and standard deviation). Similar (but not identical) histologic and cytologic appearances are categorized and named. More attention is paid to the borderlines and overlaps of the categories, rather than subtler differences within the categories (unless splitting into finer categories is being considered). Categorical data analysis relies on *contingency tables*, which are discussed in a following section. Contingency tables such as the common 2×2 table are frequently counts of categorical data; for example, how many (not what percent of) CIN 2 lesions demonstrated aneuploidy or not, compared with how many CIN 3 lesions demonstrated aneuploidy or not.

The variability in categorical data such as pathology categories shows up in diagnostic error, that is, the misassignment of a patient to the wrong category. In general, error cannot be avoided. To the epidemiologist, categorization of variable biologic continua virtually dictates that there will be error. If two categories blend into each other with regard to a characteristic (even one as complex and general as microscopic appearance), they cannot be perfectly separated based on that characteristic. Thus, pathologists search for additional characteristics to discriminate difficult-to-distinguish indeterminate cases, such as immunocytochemistry, but these ancillary measurements also have error and overlap. There is a field of statistics called *discriminant analysis* in which the goal is to determine how many characteristics must be measured to maximize correct assignment to overlapping categories. This complicated set of statistical methods underlies the development of computer-assisted cytology screening.

Error Versus Bias

Error is inevitable, but epidemiologists hope that it is mainly random, not systematically pushing the data in one way or the other. *Random error* reduces the *reliability* of repeated measurements, affecting their *precision*, and reduces the perceived strength of correlations, but the average measured value still becomes increasingly true or *accurate* as the study size increases. Systematic error, called *bias*, impacts directly on the accuracy of the measurement; no matter how large a study based on biased measurements is, the answer will be wrong. Thus, epidemiologists struggle to reduce random measurement error, but they have an even stronger dislike of biased measurements. If the exact direction and magnitude of a fixed bias were known, the data could be adjusted (like a scale that always reads three pounds too heavy), but adjustments for bias are not usually possible.

Epidemiologists combat error and bias in a few standard ways. To quantify and reduce random error, reliability is measured by repeating data collection, whether that involves reasking a question, rerunning an assay, or submitting a pathology slide for rereview.

For continuous variables, statistics of reliability begin with the *variance*. It is the sum of the squared deviations of measurements from their arithmetic average or *mean*, divided by the number of data points minus one. The *standard deviation* is the square root of the variance, and is commonly used to indicate the "spread" of a group of numbers. The standard deviation of a measurement can be com-

puted for individual members of the overall study population or for repeated samplings of a study statistic such as the mean (in which case it is called the *standard error* of the mean). Standard errors are important in making *confidence intervals* around the mean, when we compare different populations to see whether they are statistically significantly different regarding the characteristic under study.

When epidemiologists assess laboratory assays, they often consider the *coefficient of variation (CV)*, which is the ratio of the standard deviation to the mean. Low CVs (under 10% is excellent) indicate high assay reproducibility, although they do not ensure accuracy. Remember that a reproducible assay can still be biased.

For categorical variables such as pathology interpretations, statistics of reliability include the simple percentage of agreement and more complicated statistics mentioned below in the section on measures of interpathologist agreement. Epidemiologists would like to compare the pathology diagnoses to a reference standard of truth, but such reference standards virtually never exist. Certainly, there is no source of absolute truth in pathology, only advancing degrees of expertise correlated with decreasing amounts of diagnostic error. Therefore, to reduce bias in pathology, researchers are limited to the comparison of different experts. To the extent that truly independent experts agree (without consideration of each other's opinion), the possibility that either one is biased is reduced. To reduce the possibility of bias, epidemiologists try to ensure that all study measurements are made independent of each other so that knowledge of one variable cannot bias a decision about another. The difficulties of masking are discussed in a later section.

Descriptive Data

The terms used most often to describe and summarize descriptive data, such as prevalence and incidence, were defined earlier in the section on geographic differences and time trends and are not repeated. A few additional statistical concepts critical to the interpretation of descriptive data should be mentioned.

First, there is an important choice of scale in the plotting of descriptive data. The scale of the *y* or vertical axis greatly affects the appearance of the data and must always be noted when examining plotted data. A log scale flattens curves and reduces the apparent strength of trends and differences whereas an arithmetic scale does the opposite. On a log scale, an increasing, straightline trend implies an exponential, not linear rate of increase.

A common error in inference when interpreting descriptive data is the *ecologic fallacy*, the attribution of causality to an association seen only in descriptive data. For example, the international risk of colon cancer (mortality rates for each country plotted on a graph) correlates with the average dietary intake of those countries for fat, meat, and sugar and with the average amount of sunlight (the major determinant of vitamin D levels). To assume automatically that all four variables are true risk factors for colon cancer at the level of the individual would be an example of the ecologic fallacy, confusing descriptive data for analytic (individual level) data.

In the interpretation of time trend data, the possibility of a *cohort effect* must be kept in mind. A cohort effect, familiar by analogy to anyone who studies the sociology of baby-boomers, is the variation in disease occurrence that occurs in a population over time, as successive birth cohorts (persons of the same age) experience the unique environment that typifies their life course. For example, based on cross-sectional prevalence data compiled in Portland, Oregon, in 1991, the prevalence rates of koilocytotic atypia of the cervix decreases sharply with increasing age from a peak at about 20–25 years. This age trend might represent a biologic phenomenon, the result of immunity, with many women becoming infected with HPV at the time of initiation of sexual intercourse, then becoming increasingly immune and having fewer new sexual partners as they age. Or, the age trend could also reflect a cohort effect, with changing sexual practices and increasing prevalence of HPV infection over the past decades placing younger women today at higher risk for koilocytotic changes compared with their older sisters and mothers.

To distinguish cohort effects from simple age trends requires a *cohort analysis*, a type of descriptive graphing in which the age-specific prevalence rates are graphed separately for each birth cohort. These analyses are usually difficult enough in interpretation to merit a statistical consultation.

The Basic Contingency Table

The pathology slide of epidemiology is the contingency table, the basic form of which is the 2×2 table (Table 27.1). Most important epidemiologic findings, relating an exposure to risk of a disease, have been derived and can be expressed in this simple form. Extension of the table to more rows or columns does not change the concepts, only the statistical complexity.

The most common statistics computed from a contingency table are simple proportions or percentages (proportions of 100%), which can then be compared: "Ninety percent of the group with disease were smokers [$(a/a + c) = 0.90$] compared with 20% of the nondiseased [$(b/b + d) = 0.20$]. These proportions could be compared statistically using the well-known *t-test* or another test of the difference between independent proportions. More often, the *chi-square statistic* is computed, which gives equivalent interpretations but has a slightly different intent.

The chi-square test is meant to determine whether the disease categories and the exposure categories are associated or independent; that is, does being exposed affect the probability of having disease? Chi-square values are derived by comparing the expected counts of *a*, *b*, *c*, and *d*, to the values that would be expected if disease and exposure were totally independent. For example, the expected value of *a* is the cross product of $(a + b) \times (a + c)$ divided by *n*. The divergence of observed from expected values for all the *cells* of the table (*a*, *b*, *c*, *d*) are summed to derive the chi-square statistic. The larger the statistic, summarizing how much observed counts differ from expected, the more likely disease and exposure are associated by more than chance.

The chi-square statistic obtained is compared to the tabled values of the *chi-square distribution* to yield a *p-value*, the probability of observing such a chi-square value if disease and exposure are not related. In other words, this is the probability of concluding that an association exists in error. To falsely

Table 27.1. The basic contingency table

	<i>Disease</i>	<i>No disease</i>	<i>Total</i>
Exposed or test positive	a	b	a + b
Unexposed or test negative	c	d	c + d
Total	a + c	b + d	a + b + c + d = n

accept an association, when the *null assumption* would be correct, is considered a type 1 error. This name was chosen perhaps because it is generally considered a more important scientific error than failing to detect a true association (a type 2 error). If the p -value is less than an appropriate cutpoint, such as 0.05 or 0.01, then convention dictates that chance is unlikely to explain the degree of association seen in the table and the association is considered *statistically significant*.

Thanks to many published cautions, most clinicians and researchers know that a strict dependence on p -values is incorrect because the magnitude of the p -value is dependent on the size of the study. Smaller studies require stronger associations to achieve the same level of statistical significance; thus a p -value of 0.06 in a small study by no means rules out a true exposure-disease association whereas a highly statistically significant difference from a huge study may be so small as to be clinically irrelevant.

Contingency tables larger than 2×2 should be analyzed in a methodical and hierarchical fashion, not restricting the analysis to the most "significant looking" internal comparisons. First, the evidence for association in the full table should be assessed and, if there is none, then the analysis should stop. A common mistake some novices make is to look at a large contingency table, choose the most interesting difference seen, then test the significance of that extracted comparison. Given a large enough contingency table, some subtables will yield statistically significant results by chance alone. Permitting a pre-screening of the data before applying a statistical test to the most divergent data points is wrong. If one wishes to define the likely source of the association when the overall contingency table indicates statistical significance, the proper approach is to analyze smaller subtables in a complete and hierarchical manner. A formal description of the proper approach to contingency table analyses can be found in standard biostatistics texts.

When the number of study subjects is very small, such that the expected count in any cell is less than about five, then chi-square analyses are unreliable and should be replaced by a test called *Fisher's exact test*. Of course, if the study is too small, no result will be statistically significant.

One other key point about contingency tables is that the two measurements (disease status and exposure, for example) must be assumed to be independent as one embarks on statistical testing. Although a significant chi-square statistic indicates that the measurements are not independent, the initial or *null hypothesis* of independence is what the

test is designed to reject. Thus, standard chi-square analyses should not be performed to test tables where the measurements are explicitly correlated, as in interpathologist agreement studies (see later) or comparisons of the efficacy of two cell collection techniques used in the same group of patients. For these *paired-sample comparisons*, the *McNemar's test* is easy to use. The test ignores the points of agreement of the two measurements and tests the statistical significance of the amount of divergence.

It would also be wrong to include more than one measurement per subject in a standard contingency table. Measurements from a given person tend to be "auto-correlated," that is, more alike than random measurements. A difficult and evolving field of epidemiology explicitly considers multiple measurements from subjects. For example, in a prospective cohort study, it may be very interesting to study the patterns of mildly abnormal cytologic interpretations over time that indicate a risk of subsequent severe neoplasia. The level of study remains the woman, not the slide, and a simple contingency table cannot be used as it would lump together all the interpretations naively.

Measures of Risk (Absolute, Relative, and Attributable Risks)

The chi-square provides limited information regarding the strength of an association (yes/no). Therefore, epidemiologists often prefer instead to compute the more informative statistic, the relative risk (or odds ratio estimate of the relative risk). These key terms were defined in the section on epidemiologic studies of disease etiology. In this section, the relation of the terms to the contingency table are explained, with a brief discussion of ancillary topics such as statistical adjustment of confounding variables, interaction, and confidence intervals.

Suppose a prospective study started by defining an exposed group and an unexposed group of women, then followed the two groups for disease occurrence. The absolute risk of disease following exposure can be represented as an incidence rate $a/(a + b)$. The time period for this incidence rate is implicitly the duration of follow-up. The absolute risk of disease in the unexposed group, analogously, would be the incidence rate $c/(c + d)$. The ratio of these absolute risks would be the relative risk (specifically, the incidence rate ratio) in exposed versus nonexposed women, $a/(a + b)$ divided by $c/(c + d)$. A relative risk of approximately 1.0 implies the exposure is not related to risk of the disease. A relative risk greater than 1.0 implies an increased risk. For example, a relative risk of 2.0 means that the

risk of disease in exposed women is twice that of unexposed women. In contrast, a relative risk between 0.0 and 1.0 indicates a protective association (a relative risk of 0.5 implies a halving of risk associated with the exposure). Prospective studies permit the computational directness and intuitive quality of the relative risk calculation, and the ability to decompose the relative risk into the absolute risks among the exposed and unexposed groups.

In contrast, absolute risks cannot usually be calculated in case-control studies, because the true numbers of exposed women ($a + b$) and unexposed women ($c + d$) are not known. In fact, in 2×2 tables from case-control studies the values $a + b$ and $c + d$ are meaningless and should never be computed. The numbers of cases ($a + c$) and controls ($b + d$) are chosen first, and not in proportion to the true ratio of cases to controls in the population. Cases are almost always sampled in excess; in fact, oversampling cases to overcome the limitation of rarity is the major reason to perform a case-control analysis.

As mentioned earlier, although case-control data do not permit direct calculation of the relative risk, the odds ratio provides a valid estimate of it if the following assumptions are met. The cases and controls must represent an unbiased sample of all women with and without disease in the population. The disease in question must be rare if prevalent cases are studied. If the cases are all incident, the rare disease assumption is not as important, unless the disease is so common that a nonnegligible percentage of the population is developing it at any given time.

To understand these points more intuitively, again consider a prospective study. The odds of disease in exposed women is a/b , very close to the risk of disease $a/(a + b)$ if a , the occurrence of disease among the exposed, is very infrequent. Similarly, the odds of disease in nonexposed women is c/d , close to the risk of the disease if uncommon in the nonexposed women, $c/(c + d)$. With a little algebra, it is easy to see that the relative odds or odds ratio for a rare disease (a/b divided by c/d , often computed as the cross-product ad/bc) is quite close to the relative risk.

The important point is that the cross-product ad/bc can be computed from a case-control study without knowing the total number of exposed and unexposed women. So long as the odds a/c and b/d are unbiased with regard to the entire population, then a/c divided by b/d equals ad/bc equals the prospective odds ratio of a/b divided by c/d . The key is to select an unbiased sample of cases and controls. Because epidemiologists usually try to recruit

all cases occurring in a population, bias among cases is not usually an issue unless participation rates are poor. The place where bias is a major concern is among the controls. Epidemiologists spend most of their intellectual energy attempting to ensure that the ratio b/d in controls (also thought of as the percentage of controls exposed to the risk factor) is unbiased compared to the same ratio in the whole population that gave rise to the cases. Without the elimination of bias, the odds ratio does not estimate the relative risk, and the case-control design will yield a false result.

Confounding is the type of bias that concerns epidemiologists the most, particularly when they are conducting case-control studies or nonrandomized prospective studies. *Confounding variables* are factors that influence both the risk of disease and the likelihood of exposure to a risk factor under study. The relationship between exposures, confounding variables, and disease outcome is illustrated in Fig. 27.1.

When assessing whether an exposure, such as genital herpes infection causes cervical cancer, the researcher must consider and adjust for the confounding influence of HPV infection, the sexually transmitted agent that is the central cause of cervical cancer. Women who have more sexual partners are more likely to be both herpes type 2 and HPV infected (i.e., the confounding variable HPV is linked to the likelihood of the study exposure, herpes). The apparent influence of herpes type 2 on risk of cervical cancer is reduced by statistically adjusting for HPV infection status. In summary of this important point, epidemiologic analyses must adjust statistically for the influence of confounding factors to generate unbiased risk estimates. Note that confounding factors are true risk factors for disease, despite the name that suggests confusion; it is the exposure-disease association that is under question.

Adjustment for confounding is commonly undertaken by one of three methods: *exclusion*, *stratification*, or *regression modeling*. Exclusion is exem-

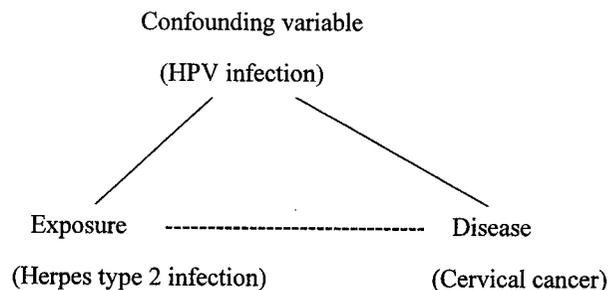


Fig. 27.1. Confounding

plified in the foregoing example by restricting the analysis to women known to be infected with HPV. Using stratification, rather than excluding any subjects, the association of sexual behavior with cervical cancer could be examined separately in each of the two strata (HPV-/HPV+), providing two unconfounded estimates akin to those derived by exclusion. The risk estimates could then be pooled to obtain a global estimate for the risk of genital herpes adjusted for HPV. This kind of stratified analysis is commonly performed using a group of procedures called a *Mantel-Haenszel analysis* in recognition of its developers.

A more conceptually difficult approach that is widely used is *logistic regression analysis*, a multivariable regression technique available in the major statistical software packages such as SAS, STATA, and BMDP. Logistic regression is especially well suited to calculation of the odds ratio as an estimate of the relative risk in case-control studies. This technique permits the simultaneous estimation of the relative risks for multiple risk factors, adjusted for each other's confounding influences. A discussion of this technique, and its uses and misuses, is beyond the scope of this chapter. The commercially available statistical packages offer multivariable regression packages in a seductively simple format that might inspire some novice epidemiologists to perform complicated analyses. However, to master the art of multivariable regression analysis takes statistical training and apprenticeship. Moreover, the results cannot be "checked" easily. It is wise to both avoid and distrust complicated analyses, especially because the bulk of what can be learned from most data sets can be expressed using simple tables and intuitively approachable statistics. In short, all modeling should be checked against simple tables for commonsense agreement.

Adjustment for confounding is often not perfectly achieved, particularly when the confounding variable cannot be measured well or when variables under study are highly correlated. In fact, it is sometimes virtually impossible using statistical methods to adjust for the confounding influences of correlated variables. For example, the most conceptually difficult areas of chronic disease epidemiology relate to time. In all data analyses involving time, the correlated effects of age at first exposure, duration of exposure, and latency (time since first exposure) are among the most difficult to figure out.

Sometimes the risk of an exposure varies by the level of another exposure. For example, the risk of esophageal cancer associated with smoking is much higher among alcohol drinkers than among non-

drinkers. This effect modification is often called *interaction*. Extreme positive effect modification is sometimes called synergy, but that term is inexact and probably is worth avoiding. Effect modification is different from confounding, in that no global adjustment to arrive at a single correct risk estimate for the exposure is possible. The risks truly vary by levels of the effect modifier. The proper approach is to present the risk estimates for the exposure separately for each level.

It is common to place *confidence intervals* around relative risk estimates to indicate the likely range of the true risk that we are trying to estimate. Confidence intervals take into account only random error, not bias, and are conceptually somewhat similar to *p*-values though more informative. Thus, a 95% confidence interval and a *p*-value of 0.05 are both commonly chosen as standard and have analogous interpretations. For example, if the relative risk of an exposure for a disease is 1.8 with a 95% confidence interval of 1.1 to 3.0, this implies that given random error, the true relative risk has a 95% chance of falling within that range. If the confidence interval for a relative risk excludes 1.0, the result is conventionally considered statistically significant. A relative risk with confidence intervals including 1.0 indicates no statistically significant association between exposure and disease. As with *p*-values, confidence intervals should be used as a guide but not followed slavishly in interpreting data.

Most analytic epidemiologic research centers on estimation of relative risks. Another very useful concept, especially for public health applications of epidemiologic results, is the *attributable risk*, also known as the *attributable proportion* or *etiologic fraction*. These terms subsume several computational forms and subtle differences in meaning, but the general meaning is clear: how much of the disease (from 0 to 100%) is due to the exposure and would theoretically disappear if the exposure were eliminated. One useful computational formula for the attributable risk, using the notation in Table 27.1, is $\text{Attributable risk} = [(a/a + c) \times (1 - 1/RR)] \times 100\%$. In words, the fraction of disease attributable to the risk factor is equal to the percentage of cases of the disease who are exposed, adjusted for the strength of the estimated relative risk. Although the formula may appear a bit complicated, it is very easy to use. The adjustment part of the formula $(1 - 1/RR)$, goes to 0 as the relative risk goes to 1.0 and goes to 1 as the relative risk goes to infinity. Thus, even if all cases are exposed, the attributable risk will be 0% if all controls are also exposed because the RR is 1.0 and the adjustment term is 0.

Causal Intermediates and Surrogate Endpoints

Increasingly, many pathology and epidemiology studies of gynecologic neoplasia do not include invasive tumors. There is a keen interest in the validation of biomarkers and intermediate/surrogate endpoints for screening, diagnosis, and etiologic research. Cancers arise as multistep processes. An exposure can become a biologically effective internal dose, resultant genetic alteration can lead to a subtle lesion, and the precursor can progress to cancer. Each step might be reversible and influenced by the genetic susceptibility of the individual. The earliest *intermediate endpoints* are often common and reversible, such as HPV infection. Later steps are less common and more fixed, such as progression of a simple HPV infection to CIN 3. Although molecular biologists may view oncogenesis as a series of "molecular hits," epidemiologists may discuss "conditional probabilities." For example, conditional on a woman being infected with an oncogenic type of HPV, what is the probability of progression to CIN 3? Conditional on having CIN 3, what is the probability of invasion? If the CIN 3 lasts for 5 years without regression, how does that affect the probability of invasion?, and so on. Oncogenesis occurs mechanistically but, lacking all the details, epidemiology presumes that events will happen by useful measurements of "chance."

The importance of a biomarker or intermediate endpoint can be evaluated using the relative risk of cancer when positive compared to when negative. A high relative risk implies importance. A *surrogate endpoint* is a more stringent term. Many studies examining the associations between biomarkers are not clinically relevant because no association with attributable risk of disease is directly made. If a biomarker is a valid surrogate endpoint, then reducing its occurrence should proportionately reduce the occurrence of the cancer itself. CIN 3 is a good surrogate endpoint for invasive cervical cancer.

The statistical evaluation of possible intermediate endpoints is linked to the analysis of confounding algebraically, but there are important differences of interpretation. When a biomarker or preinvasive lesion is proposed as an intermediate endpoint for a cancer, it should share the general risk factor profile of that cancer. In fact, its consideration by statistical adjustment should "explain" the association of known epidemiologic risk factors for that cancer. If not, the validity of the intermediate endpoint as a surrogate for the cancer is in question. As an example, the risk of ovarian cysts

detected by transvaginal ultrasound is not reduced by multiparity and oral contraceptive use. These are two very powerful protective factors in the etiology of ovarian cancer, casting doubt on the etiologic relevance of most of the cysts found by ultrasound.³ On the other hand, HPV infection almost completely explains the strong association of sexual behavior and risk of cervical cancer, as befits a central causal intermediate.¹¹

Measures of Interpathologist Agreement

Simply put, there is no universally accepted statistical measure of interrater agreement. The problem is adjustment for the influence of chance agreement, which varies with the numbers of categories and the composition of the study population. All currently available statistical methods have limitations and, therefore, it is best when possible to present the actual data to the reader, in addition to any percentage or statistic.

Consider a study of interpathologist agreement for the categories of the Bethesda System of cervical cytology. A group of 100 smears was given to pairs of pathologists, who were asked to rate them as normal or benign reactive changes, atypical squamous cells of undetermined significance (ASCUS), LSIL, or HSIL (high-grade squamous intraepithelial lesion).^{14,15} The trouble with simply calculating percent agreement is not only that some agreement is expected by chance. The results are strongly dependent on how the smears are chosen. If mainly normal smears were submitted, the percentage of agreement would be high. If a wide range of changes were equally represented, then agreement would undoubtedly decrease. In general, the most information is obtained by choosing a wide range of smears, oversampling the rarer grades to achieve a balanced study group.

The most widely used, more sophisticated statistic of agreement of use to pathologists is the *kappa statistic*. The kappa statistic computes the proportion of agreement in excess of the expected chance agreement. Kappa values can range from 1.0 (perfect agreement) to less than 0.0 (zero implies only chance agreement). The statistic has some limitations.⁸ Only tables of identical size can be compared, and the statistic is slightly dependent on the prevalence of disease. Also, the interpretation of kappa values is not absolutely clear cut, in that researchers disagree as to what defines good agreement. In general, values greater than 0.75 represent excellent agreement beyond chance, 0.40 to 0.75 is fair to good, and less than 0.40 indicates poor agreement

beyond chance.¹ An asymmetry chi-square, analogous to a multcategory McNemar's test, is often calculated with the kappa statistic. The purpose is to test whether one rater is yielding systematically more severe interpretations than the other, or whether disagreements are randomly distributed.

Many pathologists are willing to admit how difficult it is to distinguish grades of intraepithelial neoplasia as assessed by cytology. Objective molecular measurements such as HPV DNA testing are useful in clarification of equivocal cytology.¹⁴ However, as mentioned earlier, much less is said about the irreproducibility of histologic diagnoses or intraepithelial neoplasia. Interpathologist agreement based on cytology or histology tends to be moderate, not excellent,¹⁶ which is a sobering thought given the importance of the interpretations in guiding patient management. To the knowledgeable epidemiologist, misclassification of pathology is inevitable and not a matter of fault in most instances.

Screening Terms

Screening is a special area of epidemiology distinct from descriptive or analytic studies. It is rare to find a useful screening test. Finding a strong risk factor for a disease does not imply that we should screen for that risk factor, because the factor is often too common in the general population to permit its use as a trigger for clinical action.

Screening terms have very exact meanings, which may vary from other common uses of the same terms. In Table 27.1, the women in cell "a" have *true-positive* screening tests, in that they have the disease and tested positive. The women in cell "c" have *false-negative* results, because they have the disease but tested negative. The *sensitivity* of a test, also called the true-positive rate, is the percentage of diseased women who test positive [$a/(a + c)$ in Fig. 27.1]. The screening sensitivity must be clearly distinguished from the analytic sensitivity of a laboratory assay, which has a different meaning. Typically, the more analytic sensitivity the better. However, increasing screening sensitivity can lead to decreasing specificity, as indicated in the following section on ROC curves.

The *true-negative* results are in cell "d"; the *false-positive* results are in cell "b". The *specificity*, also called the true negative rate, is the percentage of women without the disease who test negative [$d/(b + d)$]. The concept of specificity is more important in screening than most realize. Because the overwhelming majority of women in a population do not have the disease under study, as the specificity per-

centage falls even slightly, the absolute numbers of false-positive screening tests rise dramatically in comparison to the number of true positives.

Therefore,² decreased specificity leads to low *positive predictive value*, the percentage of women with a positive test who truly have the disease [$a/(a + b)$]. Positive predictive value is, for many diseases, the major screening statistic of interest. Clinicians ask: If a woman tests positive, what is the likelihood that she will have disease confirmed on referral to the next clinical step (e.g., colposcopically directed biopsy, laparoscopy, or more major surgery). Low positive predictive value leads to overreferral and overtreatment.

For grave diseases, where overtreatment of normal women is less of a concern than not missing any cases, the *negative predictive value* is a very important concept of reassurance. The negative predictive value is the percentage of women who test negative who are truly disease free [$d/(c + d)$]. A clinician may ask, accordingly, "If the test is negative, what is the percentage assurance that the disease is not present and that I can safely stop the diagnostic workup?" The sensitivity of the test is usually the key determinant of negative predictive value.

When screening is mentioned, there is always an implicit notion of a *reference standard* or *gold standard* of disease. The performance of screening tests is described statistically in relation to this reference standard, and if it is flawed, then the screening statistics will be flawed. For example, colposcopically directed biopsy with pathologic diagnosis is often taken as the reference standard of cervical intraepithelial neoplasia (CIN), but the colposcopic biopsy may be misdirected or the histopathologic diagnosis may be in error. Thus, the true performance of screening tests such as cytology, cervicography, or HPV testing may be misinterpreted when compared with the results of colposcopically directed biopsies.¹⁷

Screening tests may detect prevalent disease or predict the future diagnosis of disease, and the two time frames may be confused. If some type of HPV test could truly predict incipient cervical neoplasia, even when biopsies were still negative, it would be misleading to compare the HPV screening result only to prevalent (same-day) disease defined by biopsies.

Another mistake is the following: Researchers who wish to compare the sensitivity of two screening tests double-test a research population, referring for a definitive diagnostic procedure those women who are positive for either screening test. If they then compute and report the "sensitivity" of each test, an error of circular reasoning has been made.

Because both screening tests could have missed disease (double false-negatives), the true sensitivity of either test cannot be known without referring all women in the study population for the definitive workup. Sometimes, in large studies, it suffices to refer a random sample of the women who screen negative on both tests, as a way of correcting (or of verifying, to think optimistically) the estimates of sensitivity.

The point of this discussion is that, when screening terms such as sensitivity or specificity are mentioned, then the reference standard must be explicitly stated and, if necessary, questioned.

The Receiver-Operating Characteristic (ROC) Curve

Some of the current controversy regarding the proper clinical management of inconclusive cervical cytologic smears centers on the competing needs for good negative predictive value (assurance that we are not missing any high-grade disease) and good positive predictive value (desire to not overtreat). This problem highlights an inescapable feature of screening (or more fundamentally of trying to categorize overlapping distributions): increased sensitivity virtually always leads to decreased specificity and, as a corollary, increasingly reassuring negative predictive value can only be obtained at the price of decreased positive predictive value.

There is a formal method for choosing the proper screening *cutpoint* (e.g., the viral load threshold of a DNA-based assay meriting colposcopic referral to detect CIN 2-3 or cancer) to achieve an optimal compromise between sensitivity and specificity. The technique is called the *receiver-operating characteristic (ROC)*, because the approach was developed to test how well an electronic receiver could distinguish signals from electrical noise. The concepts are useful and well explained in a few key articles that are recommended to anyone wishing to evaluate a screening test.^{12,19}

In brief, most test measurements range from zero to some high value. It is conceivable to set a series of cutpoints that define a positive screening test result demanding further attention. Lower cutpoints may detect more cases of disease but refer more women. In a ROC curve, sensitivity for detection of the target disease is plotted against 100% specificity. The expression 100% minus specificity is very close to percent referred. A very good screening test will have very high sensitivity and specificity. In other words, it will detect women with disease but refer few extra women. The quality of

screening or diagnostic tests is easy to compare using ROC curves.

Problem Areas

The major goal of including an introduction to epidemiology in a textbook on gynecologic pathology was to encourage pathologists to do epidemiologic studies and to work with epidemiologists. Accordingly, it may be worth alerting the pathologist to recurrent problem areas that exist at the juncture of the two disciplines. This section quite informally catalogs a few practical problems that appear to arise most commonly.

Dividing a Spectrum of Disease into Categories

Unfortunately, some epidemiologists may seek out pathologists to perform a service function of "making sure the cases are right," without understanding much about pathology (just as pathologists might seek out statisticians to do a rote data analysis or to figure "How many cases are needed for statistical significance?"). Providing rote pathology review may prove a difficult collaboration, because epidemiologists are prompted by their statistical methods to seek overly simplistic and discrete categorization of disease outcomes. Because the statistical methods for considering a spectrum of disease are difficult to perform and understand, epidemiologists tend to simplify disease measurements into a few (ideally two) reliably distinguished categories, such as "invasive cervical cancer" versus "normal." But, as the example of cervical neoplasia demonstrates, diseases may exist as a spectrum of changes that are impossible to divide perfectly into a few categories.

When an epidemiologist asks a pathologist to state whether a slide shows disease (i.e., defines a case) or not (i.e., rejects the case), an uncertain or heavily qualified diagnosis is difficult to force into the study dichotomy. Often, the epidemiologist must subsequently exclude the uncertain diagnoses from the analyses. It is possible to perform a "malicious analysis" in which the uncertain cases are added to the analysis as cases, then reanalyzed as controls, to see whether the uncertainty in pathologic definition affects the comparisons being made. However, too large a proportion of uncertain diagnoses can make an analytic study unreliable.

The collaborating epidemiologist must be willing to understand diagnostic error as a fact of nature

and not a failing of pathologists. The pathologist must be willing to sacrifice absolute truth to simplify the statistical data to the point of understanding. The limitations of epidemiology should be recognized. As a great physician-epidemiologist once said: "Epidemiology is a butcher shop; don't try to use a scalpel." In other words, epidemiology can only study strong risk associations, because even strong associations are made to appear weak by unavoidable measurement errors and biases. Truly weak associations will probably be missed by all but the largest and luckiest studies. With this in mind, the routine use of pathologic qualifiers such as "consistent with" and "cannot exclude" should be abandoned for epidemiologic studies, with the recognition that diagnostic errors will exist (the extent of which should be measured by reliability studies and reported).

The Need for Masking

Epidemiologists tend to mask all data collection as an automatic part of good research technique to avoid the influence of possible subtle biases that could distort risk estimates. Thus, they do not routinely tell interviewers the disease status of the subjects to minimize bias in questioning, they do not tell laboratory collaborators the identity of specimens until the results are obtained, and they ask pathologists to make their diagnoses with a minimum of information regarding the patients. Pathologists working together (panel reviews) tend to agree more readily than if the independent opinions are compared. The social tendency to promote consensus may be the cause. Epidemiologists are seeking a completely independent decision from pathologists, without influence from previous diagnoses or clinical tests, which often are being studied as risk factors for the current condition. All common statistical tests assume that the study measurements are completely independent of each other; thus, using any piece of data to influence a decision on another piece of data is wrong.

Pathologists, however, realize that diagnoses are best made in the context of complete information regarding the patient, and that asking for a microscopic diagnosis out of context, as one would demand a lab result from a machine, risks error. Some pathologists incorrectly view the request for masking as a sign of distrust of their intellectual integrity or ability to make an independent decision. The request is actually a sign of epidemiologists' belief that everyone is biased about every decision unless masked. As a revealing example, an epidemiologists' wine tasting group in Maryland covers all labels

from the bottles before tasting and unmask the results only after the "data" (opinions) are in. Fortunately, it is usually easy for good collaborators to achieve a balance between automatic demands for complete masking and the kind of complete disclosure of study information that could lead to serious biases.

Standardization of the Scientific Art of Pathology

A more thorny problem arises when epidemiologists challenge the accuracy and reliability of pathologic diagnoses, either as part of a formal pathology agreement study or as part of a larger epidemiologic project. This challenge takes the form of calculation and publication of rates of (dis)agreement between experts or between the expert and himself/herself on different days. The epidemiologist is trained to believe that all biological phenomena are variable and that all measurements of biologic phenomena are prone to random error. The pathologist has the weighty daily task of being the final arbiter of disease definition, a responsibility that does not mesh well with error.

The epidemiologist author has learned something about the world of gynecologic pathology only because of the intellectual humility of expert gynecologic pathologists (responsible for several of the chapters in this text) whose curiosity outweighed their urges to preserve their national reputations for infallibility. Most of the comparisons performed have related to the cytopathology and histopathology of cervical intraepithelial neoplasia and benign "look-alikes." Agreement rates between expert pathologists have been only fair at best but have led to a greatly increased understanding of the diagnoses.

A pathologist may feel irritated at the demands for reliability studies from new epidemiologist colleagues. If so, it might help to ask the eager-beaver epidemiologist when they had last compared their design or analytic performance in a masked comparison with other epidemiologists. Because such painful comparative exercises are almost never perpetrated by epidemiologists on themselves, mutual humility and curiosity should reign.

Specimen Adequacy Versus the Bias of Convenience Samples

Epidemiologists seeking to minimize bias are loath to permit exclusions from a complete series. They suspect that the excluded members of the set will differ from those included in a systematic (biased), rather than random, way. Thus, epidemiologists

working with pathologists wish to start their analyses by considering the entire collection of pathologic specimens available, winnowing out as needed to usable specimens but always with an eye to possible biases of exclusion that could affect the general applicability of the results. Epidemiologists distrust *convenience samples*, groups of specimens that happen to be available for testing or for review. Pathologists may view the task of defining and retrieving all relevant specimens from their center to be unnecessary. It may be difficult to decide in advance when a convenience sample is sufficient and when a more definitive collection is required. In general, convenience samples are useful for preliminary methodologic work, such as checking if genomic DNA can be amplified from the paraffin blocks available, but such studies cannot be used to reach definitive, generalizable conclusions.

Deciding How Large a Study to Do: Statistical Significance Versus Practicality

Bigger is better for the epidemiologist. It is not much more difficult to do a statistical analysis of 1000 patients than 100; in fact, it is methodologically easier because the numbers are clearly sufficient. However, the pathologist collaborator may view it differently. The question of study size is almost always negotiable, in that bigger studies permit the detection of smaller differences, but the critical difference that needs to be detected is usually open to discussion.

There are minimum numbers of subjects that permit epidemiologic analyses. It is impossible to generate a statistically significant result with fewer than 5 subjects, regardless of how strong an association is. Thirty subjects is another breakpoint. Thirty subjects is a common minimum number in that common statistics such as means start to "behave" more reliably when there are about 30 or more data points. About 200 cases and 200 controls are needed to find reliably a relative risk of about 2.0 (a doubling of risk), given typical prevalences of common exposures. Case-control studies of more than 1000 subjects are relatively rare. Cohort studies, however, often require thousands or even tens of thousands of subjects to generate enough disease endpoints for analysis. Clinical trials range from small (20 subjects) to large (thousands of subjects) based on the size of the difference being sought. In general, small studies miss weak associations, do not permit adequate adjustment for confounding, and generate less reliable estimates of risk. Still, many landmark studies of new topics have been small.

The key to defining the proper size of the study is to agree on the hypothesis and the range of expected results. Sample size calculations are very assumption dependent and usually demand information not available until the study is completed. Most epidemiologists choose a reasonable number based on cost and time available, then compute the *statistical power* of such a study to detect associations of various strengths. It is standard to require the study to have an 80% or greater chance of finding (as statistically significant) the key disease-exposure association under study, assuming the association truly exists. Epidemiologists therefore commonly accept a 20% chance of making a type 2 error (failing to "observe" a true association) whereas they restrict themselves to approximately a 5% chance of making of a type 1 error (falsely declaring a null association to be significant). As scientific skeptics, epidemiologists stack the deck against themselves to avoid being rash. When they are making multiple comparisons, they often reduce the required level of significance below 1% to even tougher standards of evidence.

For the pathologist, boredom and time commitment can be real problems in big epidemiologic collaborations. Pathology quality assurance group members can easily spend 10 hours a week on review work of fairly monotonous, unchallenging cases. Of course, the friendly epidemiologic collaborator will be monitoring to avoid any drift in diagnostic interpretations over time. There will be cases sent back with relabeling to assess intrapathologist reproducibility. The situation requires dedication, trust, and scientific interest. In truth, to answer big questions often takes big studies by a cooperative team.

Incorporating Research into Pathology Practice

The value of well-characterized pathology collections is increasing. The field of molecular diagnostics is being powered at the speed of molecular biology. The clinical relevance of new findings and potential assays, however, can only be evaluated at the restraining speed of clinical studies and epidemiology. A few new themes are emerging.

In this volume, there are discussions of genomics, RNA microarrays, and proteomics. However, venerable old histology collections are usually not useful for archival studies of DNA and especially RNA because of the destructive nature of acidic fixatives. Even neutral buffered formalin is not nucleic acid "friendly." Pathologists who wish to conduct a lot of molecular work come under pressure to per-

form frozen sections or to use ethanol or other fixatives favoring the molecular analysis as well as the morphology. As a very pragmatic point, will pathologists keep tissues past the regulatory requirements to promote science, at the expense and risk that accompany archived materials?

As another issue of practicality and trust, institutions need to work together more than ever before to further new leads into the origins of relatively rare tumors and new subdivisions of neoplasia. Issues of relabeling, confidentiality, and ambiguities of informed consent can stop conceptually appealing multiinstitutional collaborations.

For those readers who have actually displayed exceptional interest by completing this chapter, a good question might be "Where do we interdisciplinary types go from here?" Journals of common general interest to pathologists and epidemiologists are rare. Funding committees often are not composed to evaluate our jointly conceived projects. Interdisciplinary meetings for pathologists and epidemiologists are difficult to imagine and virtually nonexistent. For now, we meet in response to specific research questions, in ad hoc meetings and collaborations.

For the future, however, consider this. I am in a research group that previously contained only epidemiologists and clinicians. Then we added molecular biologists. Starting this year, we will have our first full-fledged pathologist epidemiologist. Laser capture microdissection and microarrays are beginning to replace questionnaires and abstracts as the "meat and potatoes" of cancer epidemiology.

References

1. Fleiss JL (1981) *Statistical methods for rates and proportions*, 2nd Ed. Wiley, New York
2. Gordis L (2000) *Epidemiology*, 2nd Ed, Saunders, Philadelphia
3. Hartge P, Hayes R, Reding D, et al (2000) Complex ovarian cysts in postmenopausal women are not associated with ovarian cancer risk factors. Preliminary data from the PLCO cancer screening trial. *Am J Obstet Gynecol* 183: 232-237
4. Greenlee RT, Murray T, Bolden S, Wingo PA (2000) Cancer statistics, 2000. *CA A Cancer J Clin* 50:7-33
5. Knapp RG, Miller MC (1992) *Clinical epidemiology and biostatistics*. National medical series for independent study. Williams & Wilkins, Baltimore
6. Kurman RJ, Toki T, Schiffman MH (1993) Basaloid and warty carcinomas of the vulva. *Am J Surg Pathol* 17: 33-145
7. Last JM, Abramson JH (1995) *A dictionary of epidemiology*, 3rd Ed. Oxford University Press, New York
8. Maclure M, Willett W (1987) Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 126: 161-169
9. Pisani P, Parkin DM, Bray F, Ferlay J (1999) Estimates of the worldwide mortality from 25 cancers in 1990. *Int J Cancer* 83:18-29 (see erratum *Int J Cancer* (1999) 83:870-873)
10. Sackett DL, Haynes RB, Tugwell P, Guyatt GH (1991) *Clinical epidemiology: a basic science for clinical medicine*. Lippincott, Philadelphia
11. Schiffman MH, Bauer HM, Hoover RN, et al (1993) Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia. *J Natl Cancer Inst* 85:958-964
12. Schiffman M, Herrero R, Hildesheim A, et al (2000) HPV DNA testing in cervical cancer screening: results from women in a high-risk province of Costa Rica. *JAMA* 283:87-93
13. Schiffman MH, Schatzkin A (1994) Test reliability is critically important to molecular epidemiology: an example from studies of human papillomavirus infection and cervical neoplasia. *Cancer Res* 54:1944s-1947s
14. Sherman ME, Schiffman MH, Lorincz AT, et al (1994) Towards objective quality assurance in cervical cytopathology: correlation of cytopathologic diagnoses with detection of high-risk HPV types. *Am J Clin Pathol* 102:182-187
15. Smith A, Sherman ME, Scott DR, et al (2000) Review of the Bethesda System Atlas does not improve reproducibility or accuracy in the classification of atypical squamous cells of undetermined significance. *Cancer Cytopathol* 90:201-206
16. Stoler MH, Schiffman M, ALTS Group (2001) Interobserver reproducibility of cervical cytologic and histologic diagnoses: realistic estimates from the ASCUS-LSIL triage study (ALTS). *JAMA* 285:1500-1505
17. Wacholder S, Armstrong B, Hartge P (1993) Validation studies using an alloyed gold standard. *Am J Epidemiol* 137:1251-1258
18. Whelan SL, Parkin DM, Masuyer E (1991) *Patterns of cancer in five continents*, IARC scientific publication no. 102. Oxford University Press, New York
19. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots. *Clin Chem* 39:561-577