3 MEASUREMENT

We live in a cancerophobic society. For several decades the man on the street has been bombarded with the carcinogen of the week to the point of numbing exhaustion. This epidemic reached ludicrous limits when it was announced, in all seriousness, that mother's milk "caused" cancer because it contained trace amounts of PCBs and other awful chemicals, and that children should be breast-fed for a maximum of 6 months.

In part, the present dilemma can be laid at the feet of zealous legislators and news-hungry media folks; in part, the problem exists simply because our technical expertise has far outstripped our legislative apparatus. Laws about cancer in the environment were passed several decades ago, when the prevailing attitude was that any amount of a carcinogen in the soil, air, or water was too much. Since that time, technical improvements in analytic instrumentation have allowed us to detect trace amounts of chemicals that are orders of magnitude smaller than the amounts detectable when the laws were passed (literally equivalent to a martini made with a drop of vermouth in a swimming pool of gin). However, the laws remain on the books and any attempt to repeal them at this stage would promote a rapid demise to any political career.

In part, too, the issue is epidemiologic. Epidemiologists, oncologists, and toxicologists tend to view the issue of causation as a binary variable — either something causes cancer or it doesn't. Admittedly, some attempt is made to quantify the risk by extrapolation from animal data to humans. Nevertheless, it would certainly assist the field, and perhaps our quality of life, if we would pause to ask just how much cancer a particular agent might cause. Of course, this question demands some means of quantifying the degree of risk to life and limb from a particular agent.

This chapter deals explicitly with this issue, discussing a variety of **measures of association** used by epidemiologists. The problems to which these measures can be applied are far ranging, from the estimation of the risk to health from an environmental agent, to the benefit of treatment, to the agreement between a diagnostic test and a "gold standard," and to issues of observer agreement.

ISSUES IN CHOOSING A MEASURE

The issue of measurement is critical to much of science. Lord Kelvin, a distinguished physicist of the 1800s, once said

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science whatever the matter may be.

Epidemiology is not immune to these admonitions. The issue of measurement in many sciences is, by and large, a technical issue of instrumentation, and of developing the right bit of apparatus to measure some phenomenon with the appropriate degree of precision. In epidemiology the issues are a bit more conceptual, and much thought must be directed to the appropriate selection of which variable to measure in the first place. Often the choice of variable represents a deliberate compromise; for example, in looking at the effects of an educational strategy for practicing physicians one could decide to measure the increase in knowledge of the participants, a variable that is likely sensitive to the educational strategy and can be easily tested with methods like multiple choice questions. Unfortunately, this choice begs the issue of whether the increased knowledge will be translated into a change in physician behavior with patients. In turn, we should worry whether the doctor's admonitions will change patient behavior, whether this behavior change will actually result in improved health, and whether the improvement in health will result in increased longevity or decreased morbidity. It is evident that the further we get from the intervention, the more socially relevant the outcomes are, but the less likely they are to be sensitive to the intervention.

THE DIMENSIONS OF MEASUREMENT

Epidemiologists have categorized the wide number of potential choices in the measurement of the effects of illness into the six Ds — death, disease, disability, discomfort, dissatisfaction, and debt. A little creativity can easily result in some additions to the list: psychiatrists would like to look at dysphoria and depression, and sociologists might examine disenfranchisement or dysfunction.

Some of these variables, like death and debt, are relatively easy to measure, and hence are frequently used in studies in epidemiology. Others, like dissatisfaction and disability, are notoriously difficult to measure, and have been the making of many a career in epidemiology. We will avoid, for the most part, the technical issues surrounding the measurement of these variables; the important point is that the Ds serve as a reminder that

measurement of dependent variables or outcomes need not be confined to the traditional measures like death and disease.

The choice of an outcome variable is almost inevitably a compromise based on the interplay among the several following factors:

1. **Precision of Measurement.** Measures that are subject to a large degree of random variation or individual interpretation are less useful than measures that are more precise. The judgment of precision cannot be made on an a priori basis; careful studies have shown appallingly high error rates in many areas of clinical medicine, such as radiology, that conventional wisdom would suggest are highly objective. Methods to assess precision are reviewed later in this section.
2. **Logistical Factors.** Measures are often chosen simply because they are inexpensive. Cost is certainly one criterion, as are other logistical factors like the likelihood of obtaining compliance, or the ease of entering the data.
3. **Ethical Issues.** Some measurements are unsuitable for ethical reasons. No ethics committee would permit coronary angiography to be performed on all patients in a trial, regardless of cost, simply because of the risks associated with the procedure (unless the test was a part of the patients' regular care).
4. **Importance.** Often the most important variables, in terms of their burden on the affected individual, are the most impractical to use in studies. One good example is death. It has considerable importance to the individuals involved. However, although it is precise and easy to measure, death is often rejected as an outcome variable in studies because it occurs too infrequently (thank goodness), and thus the follow-up period required would be too long. As a result, investigators often substitute other variables that are less important, but more available for measurement. As one example, hypoglycemic agents were adopted because they demonstrated the appropriate effect on blood sugar, which is much easier to measure than diabetes (although not as relevant). Much later the widespread use of the drugs was discontinued because long-term studies showed that the lower blood sugar level had no impact on longevity or complications from the disease.
5. **Sensitivity.** For a variable to be useful, there must be some reasonable chance that it is related to, or likely to change with, the independent variable under study. As an example, researchers often select a laboratory test result as a measure of effect of a risk factor or therapeutic intervention. For instance, several studies have looked at the effect of formaldehyde on lower respiratory tract disease using measures of pulmonary function as the dependent variable. The choice is reasonable in some respects: pulmonary function can be measured with a high degree of precision and relatively cheaply. The data can be elicited from patients far more easily than by using such alternatives as symptom diaries, which may cause severe problems with compliance. The difficulty is that the effects of formaldehyde may not be detectable with this measure because they are likely to occur shortly after exposure and dissipate rapidly, and so they have vanished by the time patients arrive at the clinical setting for testing. Also, relatively large changes in pulmonary function, of the order of 20 percent, are required to show any effect on patients' function. For a similar reason, the use of death as an endpoint, however important, is unlikely to be sensitive to any subtle changes resulting from low-level exposure. Of course, if formaldehyde is suspected as a potential human carcinogen, the use of death as a measure, specifically respiratory cancer death, is uniquely appropriate.

The important implication of these considerations is that issues of measurement are central to much research in epidemiology. The choice of an appropriate measure is a complex exercise in compromise. Just as investigators should be aware of the issues involved in this choice, critical readers of the literature should examine closely the variables used in a reported investigation to determine whether they are appropriate for the research goals.

TYPES OF VARIABLES

When considering issues of measurement, it is useful to make a distinction among different types of variables. Although there are various ways to describe the different variables, the important distinction is between those variables that are **categorical**, such as dead/alive, diseased/normal, or Protestant/Catholic/Jewish/other, and those that are **continuous**, like diastolic blood pressure, hemoglobin level, height, and many subjective states, such as pain, disability, or mood. Categorical variables can only take on certain discrete values. By contrast, continuous variables can, in theory, assume an infinite number of values.

Within these broad classes there is often a further subdivision. Categorical variables are classified into *nominal* variables, which are named categories like dead/alive, male/female, or white/Oriental/African, and *ordinal* (ordered) categories like Stage I/Stage II/Stage III cancer or much improved/improved/same/worse/much worse. The distinction between the two is that there is no order implied for nominal variables — whites are no higher or lower than Orientals or Africans. In contrast, there is a clear order implied in ordinal variables (e.g., staging in cancer).

Continuous variables are also divided into two classes. With *interval* variables the distance between points has some quantitative meaning, so that the difference between a blood pressure of 95 mm Hg and 105 mm Hg is the same as the difference between 110 mm Hg and 120 mm Hg. For *ratio* variables, the ratio of two quantities has meaning (e.g., the ratio of two temperatures expressed in degrees Kelvin). These latter two concepts are understood better by considering violations of the rule. A rating scale going from "much below average" to "much above average" is not an interval variable, because the distance between "much below average" and "below average" has no real meaning — it certainly would not be easy to demonstrate, for example, that it is the same as the difference between "average" and "slightly above average." In a similar vein, the ratio of two temperatures expressed in absolute or Kelvin degrees has some meaning, but degrees on the Celsius scale are not ratio variables — 20° C is not twice as hot as 10° C.

The distinction between categorical and continuous variables is important, since it influences nearly every way we think about them, as will become evident in the remainder of this section. However, the difference between nominal and ordinal variables is only important in the application of some slightly esoteric statistical tests that work for ordered categories but not for nominal categories. Similarly, there is virtually no importance to the distinction between interval and ratio variables, so the less said the better.

MEASUREMENT WITH CATEGORICAL VARIABLES

We began this section on measurement with the suggestion that much of the confusion surrounding the carcinogenic risk of many environmental hazards is a result of inadequate attention paid to the quantification of risk. In this section we will develop a number of ways to approach the issue of risk assessment. There are two parts to the question: (1) deciding on the appropriate way to **measure** the health effect, and (2) deciding on some way to express the **association** between the supposed cause and the outcome.

For the moment let us define the issue a little more precisely. Without getting into the specifics of risks from radiation, PCBs, dioxin, ethylene dibromide, Agent Orange, video display terminals, or hydro lines, it would seem apparent that we are being bombarded with all sorts of chemical, electromagnetic, nuclear, and particulate delights that never assaulted our ancestors. That being the case, one possible result of the overall impact of all these insults to the organism would be an increase in the overall rate of cancer over the past century or so. If these pollutants are indeed devastating our health, this should be reflected in a gradual increase in cancer rates as time passes.

As we shall see, this *seems* simple enough, but it isn't. First, should we count all cases of cancer, or all deaths from cancer? After all, to the extent that our therapies are getting better we might actually be curing some folks, which would make the death rate drop even though there may be just as many or even more cases around. On the other hand, we're also getting better at detecting cancer with methods like Pap smears and mammography, which weren't available a few years or decades ago. The effect of this might be to inflate the apparent number of cases in recent years, although it would have less impact on deaths, since by the time someone dies from it, cancer is fairly obvious.

For convenience and convention we call the counting of cases the measurement of **frequency**, and the counting of deaths the measurement of **impact**. We will explore the issue of the overall effect of the environment using both these measures, by examining the risk of cancer in the 1930s and the 1980s to see if we can detect the effect of a (questionably) deteriorating environment.

MEASURES OF FREQUENCY

Measures of frequency focus on the **occurrence** of disease as opposed to the sequelae of disease (in particular, death). There are a number of ways one can approach the counting of disease. The choice is based on the unpleasant reality that it takes some time to do a study, and while the clock is ticking, new folks are unfortunately developing a disease at the same time that some lucky souls are being cured of it (at least for some diseases) and others are dying of it. All this coming-and-going in and out of the study wreaks havoc with any attempt to count who actually has the disease. To

overcome this state of affairs, epidemiologists have worked out a few standard ways of counting bodies, warm or otherwise.

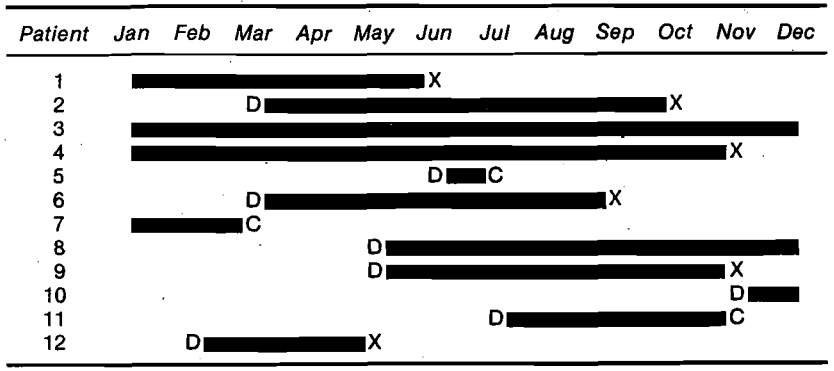
To return to our original problem, let's suppose we wish to count all cases of cancer (of all types) in Canada in 1987. Having agreed on the criteria for diagnosis and carefully set up a sampling frame that is perfectly representative of the population of Canada, or, alternatively, having developed a reporting mechanism for all cases in Canada, we now start counting on January 1 and stop on December 31. All the counts come pouring in, and the systems analysts and statisticians are rubbing their hands in glee at all the years of prospective employment ahead. Now the embarrassing questions emerge.

To illustrate the difficulty, let's examine what happened at the cancer reporting center in Plumcoulee, Manitoba. There are a total of 200 people in this farming community, a fact that we'll need to know later. The reports to the center are shown in Table 3-1.

It is obvious from the table that we can get wildly different estimates of the amount of cancer in Plumcoulee depending on how we choose to do the counting. If we just look at the number of cases around at any point in time, we find four in January, three in December, and eight in July. If we count the total number of folks who were reported this year, the answer is 12. If we count the number of new cases in 1987, it's eight. Finally, there were six deaths from cancer in that year.

There are, however, some standard ways to report the data, as we'll discuss in this section.

TABLE 3-1 Reports of Cancer, Plumcoulee, Manitoba



D = first diagnosis; ▬ = with disease; C = cured; X = died

INCIDENCE

Incidence is defined as follows:

$$\text{Incidence} = \frac{\text{Number of new cases in a fixed time period}}{\text{Number of people at risk}}$$

Usually the period of study is chosen to be 1 year, in which case we speak of the **annual incidence**. In Plumcoulee (see Table 3-1) there were eight new cases of cancer in 1987. If we had decided to focus on the 3-month incidence, there were three new cases from January to March.

The denominator, or number of people at risk, is not quite 200 people because patients 1,3,4, and 7 already had cancer and thus could not be counted as "at risk"; this reduces the denominator to 196. Thus the annual incidence is:

$$\text{Annual Incidence} = \frac{8}{196} = 0.0408 \text{ cases per year}$$

Usually the incidence of disease is much lower than in this example, and the correction for preexisting cases is unnecessary. Further, to make things more readable, incidence is often cited as cases per 1,000 (in this example, 40.8 cases per 1,000 per year), or even as cases per million per year for very rare disorders.

PREVALENCE

If we are planning screening programs, disease incidence is of immediate interest. However, if we are concerned with the provision of services for people with the disease, such as palliative care, our immediate concern is "How many people actually have the disease at any point in time?" This quantity is called the **prevalence**, which is defined as follows:

$$\text{Prevalence} = \frac{\text{Number of people with the disease}}{\text{Number of people at risk}}$$

In contrast to incidence, prevalence is determined at a single point in time. Still considering the data from Table 3-1, perhaps the most rational point in time to choose is the middle of 1987, or July 1. Looking at the table, we find that patients 2, 3, 4, 6, 8, and 9 had cancer at this time. Patient 5 was cured sometime in July, and should be counted on July 1, whereas patient 11 was diagnosed in July and would likely not enter the count. This leaves a total of seven cases in the numerator.

Again the denominator is not quite 200. By July 1, patients 1 and 12 were deceased, so the denominator is only 198. Finally, the prevalence is:

$$\text{Prevalence} = \frac{7}{198} = 0.0354 = 35.4 \text{ per } 1{,}000$$

PERIOD PREVALENCE

A close analogy to the incidence is the **period prevalence**, which is based on the number of people with the disease over a defined period of time (usually 1 year). The formal definition is:

$$\text{Prevalence} = \frac{\text{Number of people with the disease over the time period}}{\text{Number of people at risk over the time period}}$$

The calculation of annual prevalence in our example from Table 3-1 is straightforward. There are 12 people identified as having cancer in that year, and 200 at risk, so the period prevalence is simply $12/200 = 0.06 = 60$ per 1,000. If we were to calculate the quarterly prevalence for the first quarter of the year, we would include only patients 1, 2, 3, 4, 6, 7, and 12; the period prevalence for 3 months is therefore $7/200 = 0.035 = 35$ per 1,000.

RELATION BETWEEN PREVALENCE AND INCIDENCE

The previous definitions were slightly different in dimensions. Incidence is based on a fixed time period, and is quoted per month or year. However, prevalence is calculated at a single point in time. It happens that the two quantities have an interesting relationship, which involves the average duration of disease:

$$\text{Prevalence} = \text{Incidence} \times \text{Duration}$$

It's not easy to demonstrate the relationship mathematically, but it is easy to show that it is reasonable. Think of a chronic, but relatively nonlethal disease like rheumatoid arthritis (RA). Once an individual acquires the disease, he/she carries it until death, so the duration is calculated by subtracting the average age at onset from the expected life span. Thus, each new case of RA is added to the pool of prevalent cases, and although relatively few cases may be added each year, there are a large number of prevalent cases around. So the prevalence of RA is much greater than annual incidence.

By contrast, the ordinary cold has a duration of a few days at most, and kids can often get more than one per year. In this situation the annual incidence might approach, or even exceed, 1,000 per 1,000. Yet unless there's an epidemic around, relatively few people have a cold at any time, so the prevalence of colds is not nearly as high as the incidence — perhaps 50 per 1,000. Because the duration is very short, the prevalence is much lower than the annual incidence.

The relationship may seem to be of only arcane interest. However, it is often easier to obtain published data on disease prevalence than on incidence; yet if you want to do an intervention or prevention study, it is usually of greater interest to know how many new cases you are likely to get. Through the use of this formula and a reasoned guess at the duration of the disease, you can arrive at a plausible estimate of the number of new cases.

CASE FATALITY RATE

While we're examining the fate of the Plumcoulee patients from Table 3-1, we might as well introduce a term that links disease *frequency*, or the likelihood of developing the disease, to disease *impact*, which is the likelihood of dying from the disease.

First of all we note that a total of six persons from Plumcoulee died of cancer during the study. It is natural to express this quantity in a similar manner to our measures of disease occurrence to form a quantity called the **mortality rate**; which is defined as follows:

$$\text{Mortality Rate} = \frac{\text{Number of deaths from disease in a time period}}{\text{Number of people at risk}}$$

Studying the data from Plumcoulee, we see that six people (patients 1, 2, 4, 6, 9, and 12) died of cancer in 1987. There were 200 people at risk, so the **annual mortality rate** was $6/200 = 0.03$, or 30 per 1,000.

As we shall see in the discussion on measures of impact, this approach is a fairly crude basis for comparison. However, there is another relationship evident from the display. When relating frequency to impact we might wish to study the likelihood that a disease may be fatal. This quantity is called the **case fatality rate**, and is defined as follows:

$$\text{Case Fatality Rate} = \frac{\text{Number of deaths from disease in a time period}}{\text{Number of people with the disease}}$$

In the present example there were 12 people with cancer in Plumcoulee in 1987, and six deaths; the case fatality rate is therefore $6/12 = 50$ percent per year.

MEASURES OF IMPACT

We began this discussion with the idea that one broad way to determine whether all the industrial pollutants have affected human health was to examine the rates of cancer over several decades, to see if any increasing trend was evident. We briefly discussed the advantages and disadvantages of looking at disease frequency (cases of cancer) and disease impact (deaths from cancer).

The impact of disease need not focus entirely on death. For a chronic disease like arthritis, disease impact would more appropriately be calculated using measures of activities of daily living, function, or quality of life. However, for the example we have been pursuing we will focus on mortality. The measurement of mortality has one major advantage over the measurement of frequency, namely that relatively complete archival sources are available and have been for several decades (or for several centuries in Great Britain). Instead of setting up a reporting system such as was proposed for Plumcoulee and allowing it to run for a few decades while we

epidemiologists cool our collective heels, we can conduct a retrospective impact study. In this discussion we use actual data, based on Canadian statistics for 1933 and 1973, to examine our research hypothesis that the increased level of chemical, radiologic, and particulate pollution in Canada in the intervening 40 years has led to an increase in the observed rate of death from cancer.

MORTALITY RATE

To test this hypothesis, let's tum to our desk copy of Canadian statistics. We look up the appropriate sections and compile the data (Table 3-2). To make the comparison easier, it makes sense to work out the number of deaths per 1,000 population. This is called the annual mortality rate, which is defined as follows:

$$\text{Annual Mortality Rate} = \frac{\text{Number of deaths in a year}}{\text{Total population}}$$

For 1933 the annual mortality rate is 11,056 per 10,500,000, or 1.05 per 1,000. For 1973 the annual mortality rate is 44,877 per 21,400,000, or 2.10 per 1,000.

From these data it would appear that the rate of cancer has nearly doubled in 40 years. We may conclude that perhaps there is evidence of a significant health effect of pollutants. Nevertheless there are a number of steps we can take to refine the comparison.

PROPORTIONAL MORTALITY RATE

We were in the fortunate position when we calculated the mortality rate to have a good estimate of the denominator, or the population at risk. Federal census takers in the Western world go to great pains and expense to determine how many people there are in the country in given years (perhaps so they can ensure complete tax returns to pay for the census). However, in many situations where research is conducted on subpopulations (e.g., workers exposed to welding fumes or residents near a landfill site) it would be very difficult or impossible to determine on the basis of existing records how many people were in the denominator in a given year.

On the other hand, it is much easier to determine the cause of death of all the people in a population who died, because death certificates are a legal necessity. We can reason that if pollutants are causing more cancer in 1973

TABLE 3-2 Canadian Cancer Statistics

	1933	*1973*
[illegible]		

than they did in 1933, proportionately more deaths should be caused by cancer than by other causes in 1973 than in 1933. This approach is called the **proportional mortality rate** or **PMR**. It requires no knowledge of the people at risk, only mortality data. The PMR is defined as follows:

$$\text{PMR} = \frac{\text{Number of deaths from a particular cause}}{\text{Total number of deaths}}$$

It turns out that in 1933 there were a total of 122,850 deaths recorded in Canada. In 1973, 236,200 deaths were recorded. The resultant PMRs are shown in Table 3-3.

It appears that the same trend to higher mortality rates in 1973 is present in these data. Of course one alternative explanation is that proportionately more people were dying of cancer in 1973 simply because fewer people were dying from everything else. This makes some sense because tuberculosis, diphtheria, and other serious infectious diseases were present in 1933 but absent in 1973. Certainly there is some evidence that this may be occurring; males born in 1933 had a life expectancy of 41.1 years, whereas men born in 1973 had a life expectancy of 68.2 years.

This example also nicely illustrates the strengths and weaknesses of the PMR method. Its strength is that it can be applied in situations in which only minimal data are available; its weakness is that a high PMR is always open to two interpretations: (1) more deaths from the cause of interest or (2) fewer deaths from everything else.

AGE-SPECIFIC MORTALITY

In general, cancer is a disease of old age. Although a few young persons die of cancer, in most circumstances there is a period of a few decades between exposure to some cancer-causing agent and the onset of the disease. This must be kept in mind when contrasting 1933 with 1973; not only might more people have died from other causes in 1933, as we mentioned previously, but also more people might have died young from other causes and not lived long enough to develop cancer.

To determine if this reasoning results in an alternative explanation for the higher observed cancer mortality in 1973, we could look only at the death rate from cancer in older people (e.g., older than 75 years of age). We could

TABLE 3-3 Proportional Mortality Rates for Cancer

	1933	*1973*
[illegible]		

then calculate the cancer mortality rate in this age segment. The result is called the **age-specific mortality rate**, which is defined as follows:

$$\text{Age-Specific MR} = \frac{\text{Number of deaths in a particular age range}}{\text{Total number of deaths in a particular age range}}$$

Let's work this example through. In 1933 there were 5,126 Canadians older than 75; in this group there were 110 cancer deaths. Therefore the age-specific mortality rate is 110/5,126 = 21.5 per 1,000. Similar data from 1973 indicate that there were 915 cancer deaths among the 35,295 Canadians over age 75, which results in an age-specific mortality rate of 915/35,295 = 25.9 per 1,000.

These rates are indeed a little closer than the overall mortality rates we looked at earlier, thereby suggesting that a partial explanation for the differences is simply that people were dying of other causes in 1933 and were not living long enough to develop cancer. However, it is unfortunate that in order to make this comparison it was necessary to ignore most of the data.

STANDARDIZED MORTALITY RATE

The discussion on age-specific mortality rate suggested that if we restricted our view to those individuals who survived long enough to be at risk of developing cancer, there was a smaller difference in cancer rates between 1933 and 1973 than was evident when we simply looked at overall mortality. The difference between the two sets of data reflects (1) the influence of age on mortality rates from a specific disease, and (2) differences in the age distributions between the Canadian population in 1933 and 1973.

Most diseases show a strong relationship with age. Risk from chronic diseases like heart disease and cancer increases with age, whereas infectious diseases are more common in the young. Even pedestrian mortality shows a strong bimodal distribution with age, and strikes the very young, who lack awareness of the dangers of traffic, and the very old, who can no longer see and hear danger as well as before (or run as fast!).

Because of the strong influence of age on disease mortality rates, any comparison between two different populations is considerably strengthened by correcting for the differences in age distribution. This approach is called the **standardized mortality rate** or **SMR**, and builds on the age-specific mortality rate. Having broken down the deaths in the population of interest by age and created age-specific mortality rates, we then use them with the distribution of age in a reference or standard population to create an overall projected mortality rate. There are four basic steps in the process:

1. Calculate the age-specific mortality rate for each age range in the population of interest.

2. Multiply this rate by the number of people in the age range in the standard population. This then determines the number of individuals in the standard population who would die from the disease.
3. Add up the total number of projected deaths across all age levels of the standard population.
4. Finally, convert this to a mortality rate by dividing by the total numbers in the standard population.

For example, to compare the cancer mortality in 1933 and 1973, we will project them both onto a reference population distribution (in this case the population distribution of Canada in 1970, but any year could have been chosen). The method is illustrated in Table 3-4.

After this lengthy process we then can determine that the standardized mortality rate for cancer deaths for 1933 is 2,510 per 1 million, or 2.51 per 1,000. Similar calculations can be performed for cancer deaths in 1973 and from all other causes in both 1933 and 1973, always using the 1970 population as the standard. These calculations are shown in Table 3-5.

TABLE 3-4 Calculations for Standardized Mortality Rate

1	*2*	*3*	*4*	*5*	*6*
Age Range	*1933 Pop.*	*Cancer Deaths*	*Age-Specific Mortality (Col. 3 ÷ Col. 2)*	*1970 Pop.*	*Standard Deaths (Col. 4 × Col. 5)*

TABLE 3-5 Standardized Mortality Rates per 1,000

	1933	*1973*

Some of our suspicions are therefore correct. People indeed died at a much faster rate from other causes in 1933 than in 1973—15.12 per 1,000 versus 8.91 per 1,000, respectively. There nevertheless appears to be an excess cancer risk persisting in 1973 of about 20 percent (3.10 versus 2.51 per 1,000). However, this is considerably less than the doubled risk originally calculated using the unstandardized mortality rates.

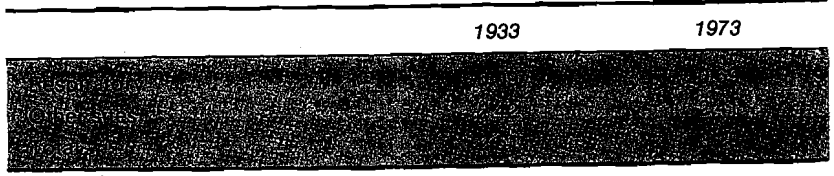
SUMMARY

The standardized mortality rate is about the best estimate of the mortality arising from a particular cause, and is virtually a prerequisite for any comparison across different populations. Proportional mortality rates are a weak alternative, useful only in situations in which there are no denominator data available.

It should be kept in mind that the application of SMRs corrects for the confounding effect of age, and possibly of sex differences, but that's all. To conclude that *any* observed difference results from a particular cause requires the elimination of all other possible causes. The point is nicely illustrated by a final run at the 1933-1973 comparison.

The difficulty arises from the use of a historical control, as described in *Research Methodology*. To conclude that the observed difference between 1933 and 1973 is caused by industrial pollution requires that we eliminate from suspicion all the other differences between 1933 and 1973. One difference in particular is staring us in the face—cigarette smoking. Smoking per capita has increased steadily from the turn of the century until recent times, and cigarettes are a known and strong causal factor in lung cancer. These facts suggest that we may further understand the cause of the increase of cancer deaths from 1933 to 1973 by separating respiratory cancer from cancers of all other sites (since the latter are only weakly related to smoking). If we do this, and calculate SMRs for respiratory cancer and other sites, all the differences between 1973 and 1933 can be accounted for by a sevenfold difference in respiratory cancer rates (Table 3-6). This of course doesn't prove that smoking, rather than pollution, is the cause of the increase. However, it does suggest that there is no general impact of air, water, and foodborne chemicals on human health reflected in cancer rates.

TABLE 3-6 Respiratory Cancers vs Cancers from All Other Sites*



* Standardized mortality rates per 1,000

MEASURES OF ASSOCIATION WITH CATEGORICAL VARIABLES

We began this discussion with the assertion that much of our fear about cancer and the environment was a result of inadequate quantification of the additional risk. To this point we have dwelt on measurement issues and sought means to measure the health effects in an unbiased manner. We now wish to explore methods to measure the *strength of association* between two variables.

We have already used some rough-and-ready measures of association. We found in the last section that there was a sevenfold higher risk of respiratory cancer in 1973 than in 1933. We could restate the data in two other ways: (1) the risk of respiratory cancer increased from 0.09 per 1,000 to 0.69 per 1,000, or (2) there was a risk of cancer of 0.60 per 1,000 attributed to the different circumstances in 1973 and 1933. In the next few examples we will formalize these concepts.

Let's begin with a new example that is related to therapeutic benefit. The issue is the relationship between cholesterol and heart disease. For a long time a strong association between serum cholesterol and heart disease has been known; however, the implications of this finding were not clear. Did a high level of cholesterol "cause" heart disease, or was it simply a marker of a certain genetic predisposition? The key issue has been whether it could be demonstrated that lowering cholesterol levels by diet or drugs would reduce the rate of heart disease.

This was finally demonstrated in 1985 by the Coronary Primary Prevention Trial (CPPT), a randomized trial that was conducted at a number of clinics in North America. The researchers began by screening nearly half a million men to find a group of 3,900 who had very high serum cholesterol levels (above 256 mg per deciliter) but as yet no evidence of disease. The men also had to comply with a fierce regimen. The drug, called cholestyramine, was foul-smelling, foul-tasting, and gut-wrenching, and had to be taken in water six times per day. The researchers eventually found their bunch of docile souls who would go along with the treatment. They were randomized into two groups (the placebo was concocted to taste just as bad) and followed for 7 to 10 years. After the dust settled there were 30 cardiac deaths in the drug group and 38 in the control group, figures that were statistically significant. There was no overall difference in death rates, but this won't concern us. The ways in which these data might be displayed are discussed next.

RELATIVE RISK

The data from the cholestyramine study appear in Table 3-7. The **relative risk**, as the name implies, is a measure of the likelihood of occurrence of the target event (death or disease) in those exposed and not exposed to the agent of interest. It is defined as follows:

$$\text{Relative Risk} = \frac{\text{Mortality rate (or incidence) in exposed group}}{\text{Mortality rate (or incidence) in unexposed group}}$$

TABLE 3-7 Data from Cholestyramine Study

	Cardiac Deaths	Alive	Total
Cholestyramine	30	1,870	1,900
Placebo	38	1,868	1,906

Mortality rates in the two groups are 30 per 1,900 and 38 per 1,906. Therefore, the relative risk from cholestyramine is 30/1,900 ÷ 38/1,906 = 0.792. To put it another way, the risk of cardiac death in the treated group was 1.00 − 0.792, or 21 percent lower than in the placebo group, a risk reduction of 21 percent.

The data can be presented in another way. We could turn the question around and ask what the relative risk of cardiac death resulting from the absence of a drug is. This relative risk is the inverse of the previous calculation: 38/1,906 ÷ 30/1,900 = 1.26.

ETIOLOGIC FRACTION

Closely related to the notion of risk reduction is a concept called the **etiologic fraction (EF)**. When considering a risk factor for a disease, in this case high cholesterol levels, we are interested in what fraction of the cases of cardiac death has high cholesterol levels as its etiology. Since there were 38 deaths in this cohort when high cholesterol levels were present, and 30 deaths when this risk factor was absent, we could define the proportion of cardiac deaths, or the EF, as follows:

$$EF = \frac{\text{Mortality in exposed group} - \text{Mortality in unexposed group}}{\text{Mortality in exposed group}}$$

For the CPPT trial (see Table 3-7) the etiologic fraction is (38 − 30) ÷ 38 = 21 percent. This is the same number as, although a different concept than, the risk reduction we calculated earlier.

ATTRIBUTABLE RISK

The relative risk gives some indication of the increased risk (in the case of a risk factor) or benefit (in the case of a therapy) in relative terms. However, we would often like to examine the actual increase or reduction in incidence or mortality attributed to the cause. This is called the **attributable risk (AR)**, and is defined as follows:

AR = Mortality rate (or incidence) in exposed group − Mortality rate (or incidence) in unexposed group

In the cholestyramine example (see Table 3-7) the attributable risk of cardiac death (attributable to the absence of the benefit derived from cholestyramine) is 38/1,906 − 30/1,900 = 4.1 per 1,000.

The example nicely illustrates the important differences between the two concepts of relative risk and attributable risk. The CPPT trial began with a highly selected cohort of people with very high cholesterol levels, followed them for a long time (7 to 10 years), and indeed demonstrated a statistically significant risk reduction of 21 percent. However, this amounted to a reduction in risk of cardiac death of only four per 1,000, compared with a total rate of death in both groups of about 70 per 1,000.

RELATIVE ODDS

The concepts of association we have discussed so far work well for most situations in which we wish to examine the effect of a particular risk factor on the subsequent occurrence of disease. However, there is one study design, the case-control study (see *Research Methodology*), in which things don't quite fit. Case-control studies are used in situations in which the likelihood of developing disease is low, or there is a long latency before the onset of disease. Typically, both these conditions apply to the investigation of risk factors in cancer. In these circumstances we assemble a group of people with the disease (cases) and an appropriate set of people without disease (controls), usually of the same size, and we examine the exposure of the two groups to the risk factor of interest.

As one example, continuing our cancer theme, Table 3-8 was derived from one of the original studies linking lung cancer to smoking.

The fact that the rate of lung cancer overall is so high is a sure clue that we are dealing with a case-control study, since if these data were based on a cohort study that assembled persons who did and didn't smoke, we would arrive at the alarming conclusion that the overall rate of lung cancer was about 34 percent. However, if we continue along the lines we had done previously, we could calculate a risk of cancer in the exposed group of 659/684, or 96 percent, and in the unexposed group of 984/1,332, or 74 percent. The relative risk of lung cancer is then, using the previous methods, 0.96 ÷ 0.74 = 1.30. Although the final result seems plausible, the intermediate steps are insane because of the nature of the design. In fact,

TABLE 3-8 Lung Cancer and Smoking

	Cases	Controls	Total
Smoker	659 (A)	984 (B)	1,643
Nonsmoker	25 (C)	348 (D)	373
Total	684	1,332	2,016

lung cancer is much rarer than we have made it out to be; the controls without cancer are sampled from a much larger population of healthy folks than are the cases.

Although we cannot calculate from these data an actual risk of getting lung cancer, we can frame things in a different way. We begin with the cases and play a gambling game, asking the odds that this person was exposed to the suspected carcinogen. When a gambler says that the odds of a candidate's being elected are 1:4, he is saying that the probability of his being elected is one-quarter that of his not being elected, and since these probabilities add to one, a little mental arithmetic shows that the probability that this candidate will be elected is 20 percent. Similarly, the odds that an individual with lung cancer was exposed to tobacco are A/C = 659/25 = 26.4; and the odds that an individual in the control group was exposed is B/D = 984/348 = 2.83. The **relative odds** of lung cancer from tobacco exposure are then:

$$\text{Relative Odds} = \frac{\text{Odds of exposure for cases}}{\text{Odds of exposure for controls}} = \frac{A/C}{B/D}$$
$$= 26.4/2.83$$
$$= 9.33$$

DIAGNOSTIC TESTS

The twentieth century has seen dramatic changes in disease patterns in the Western world. Since the advent of effective antibiotics, vaccines, and, perhaps most important, adequate nutrition and sanitation, most people in industrialized countries can look forward to a full life. Our present preoccupation is with chronic, lifestyle-related diseases for which there are unlikely to be any "magic bullets" in the foreseeable future.

One result of these changes is that epidemiologists have moved away from their historical roots in the study of epidemics to such diverse activities as the study of occupational risks or trials of therapeutic agents, in order to maintain employment. (One result of this shift in employment patterns is that books such as this are now required to tell health professionals what epidemiologists do.)

However, thanks to a new infectious disease, AIDS, that has all the devastating characteristics of the traditional scourges of mankind like cholera and the black plague, epidemiologists find themselves the center of attention at cocktail parties. We need not devote any space in this section to describing the natural history, prevalence, modes of transmission, or risk factors of AIDS—these are taught to elementary school students. However, we will use this disease as an instructive example of a measurement problem, the application of diagnostic tests.

There are now two high-risk populations for AIDS—homosexuals because of sexual contact and street drug users because of the sharing of contaminated needles. Before the advent of adequate screening tests, there was a third high-risk segment—people requiring blood transfusions for any reason. In particular, a significant number of hemophiliacs acquired AIDS as a result of their exposure to large numbers of transfusions. However, since 1985 all blood products are routinely screened for AIDS using the enzyme linked immunosorbent assay (ELISA) test.

As diagnostic tests go, ELISA is a very good one indeed. This is fortunate, because the consequences of the test are severe. If an individual has AIDS antibodies, there is at least a 30 percent chance of developing the disease, and AIDS has nearly a 100 percent mortality. The consequences of a false positive are also severe. If we tell someone he/she has antibodies when this isn't the case, we are causing massive anxiety and lifestyle changes. Conversely, if we miss blood products containing antibodies, the chance of infecting someone is high.

Let us examine the performance of this test in two populations: (1) in a homosexual population, in which the prevalence of AIDS antibodies is about 50 percent, and (2) in routine screening of blood donations, in which the prevalence of antibodies is about 0.2 percent.

TRUE POSITIVE, FALSE POSITIVE, TRUE NEGATIVE, AND FALSE NEGATIVE RATES

Let us imagine that the ELISA test is being used as a diagnostic test for a high-risk population, e.g., homosexuals in New York City. Actual figures for this group indicate that the prevalence of the AIDS antibody is about 50 percent.

To examine the test performance, we could screen a group of individuals and compare the test result with their true status. Truth isn't easy to come by, but in this case there is a more expensive, but virtually perfect test called the Western blot test. We could take samples from the group and perform both tests on the samples. If we were to screen 1,000 individuals with the test and compare the test result to the "gold standard," the results would be similar to those found in Table 3-9.

TABLE 3-9 Results of ELISA vs Western Blot Test in Screening of 1,000 Homosexuals from New York City

	Gold Standard (Western Blot)		
	Antibodies	No Antibodies	Total
ELISA Positive	498 (A)	4 (B)	502
ELISA Negative	10 (C)	488 (D)	498
Total	508	492	1,000

The characteristics of tests are usually described in terms of the letters (A, B, C, D) in the four cells of the table. One way of describing the test's performance is as follows:

$$\textbf{True Positive Rate} = \frac{\text{People with positive test and disease}}{\text{All people with disease}}$$
$$= A/(A + C) = 498/508$$
$$= 98.03 \text{ percent}$$

$$\textbf{False Negative Rate} = \frac{\text{People with negative test and disease}}{\text{All people with disease}}$$
$$= C/(A + C) = 10/508$$
$$= 1.97 \text{ percent}$$

$$\textbf{True Negative Rate} = \frac{\text{People with negative test and no disease}}{\text{All people without disease}}$$
$$= D/(B + D) = 488/492$$
$$= 99.19 \text{ percent}$$

$$\textbf{False Positive Rate} = \frac{\text{People with positive test and no disease}}{\text{All people without disease}}$$
$$= B/(B + D) = 4/492$$
$$= 0.81 \text{ percent}$$

SENSITIVITY AND SPECIFICITY

Another way of describing the test's characteristics has its origins in the biochemistry laboratory. We speak of **sensitivity**—how sensitive the test is at detecting disease—and **specificity**—how good the test is at rejecting samples that are not diseased. Let's use the data from Table 3-9.

The test sensitivity is a measure of the test's ability to detect people with the disease, and is measured as follows:

$$\text{Sensitivity} = \frac{\text{Number with disease who have a positive test}}{\text{Number with disease}}$$
$$= A/(A + C) = 498/508$$
$$= 98.03 \text{ percent}$$

Conversely, the test specificity measures the ability of the test to correctly identify people who do not have the disease, and is measured as follows:

$$\text{Specificity} = \frac{\text{Number without disease who have a negative test}}{\text{Number with disease}}$$
$$= D/(B + D) = 488/492$$
$$= 99.19 \text{ percent}$$

As you can see, sensitivity is the same as true positive rate, and specificity is the same as true negative rate.

POSITIVE AND NEGATIVE PREDICTIVE VALUE

The descriptions thus far give some picture of the characteristics of the test. However, the denominator for both sensitivity and specificity assumes some knowledge of the true state of affairs, since it is based on people who do or don't have the disease. Clinicians rarely have the luxury of a "gold standard"; if they did, they wouldn't be doing the test. Putting it another way, assume you are about to advise someone who has just received a positive ELISA. Do you tell the individual that he/she has AIDS antibodies? What is the chance that someone with a positive ELISA does not have antibodies? These probabilities are embodied in the concepts of **positive predictive value** and **negative predictive value**, in which the denominators are

based on people with positive and negative tests. Again using the data from Table 3-9, these values are measured as follows:

$$\text{Positive Predictive Value} = \frac{\text{People with positive test and disease}}{\text{All people with positive test}}$$
$$= A/(A + B)$$
$$= 498/502$$
$$= 99.20 \text{ percent}$$

$$\text{Negative Predictive Value} = \frac{\text{People with negative test and no disease}}{\text{All people with negative test}}$$
$$= D/(C + D)$$
$$= 488/498$$
$$= 97.99 \text{ percent}$$

RELATIONSHIP BETWEEN PREVALENCE AND PREDICTIVE VALUE

The data we have presented so far give a fairly encouraging picture of the ELISA test. If someone has a positive test, we can be 99.2 percent certain that person really has AIDS antibodies. However, the calculations were based on a situation where the prevalence of antibodies was high (about 50 percent). In different circumstances the picture may not be as rosy. For example, experience in screening blood donations has shown that the prevalence of AIDS antibodies is actually closer to 0.2 percent. As pointed out over 3 decades ago, this change in prevalence may drastically affect the usefulness of the test.

Working out a new contingency table (as in Table 3-9) we now have a prevalence of 0.2 percent; two people out of the 1,000 will have antibodies, and 998 will not. Because the prevalence is so low, imagine screening 1,000,000 units of blood, of which about 2,000 will have antibodies (whether we use 1,000 samples or 1,000,000 does not affect the results at all, it just eliminates decimal points during the calculations). Since the test has a sensitivity of 98.0 percent, $0.98 \times 2{,}000 = 1{,}960$ persons will test positive with ELISA (cell A) and 40 will test negative (cell C). Now there will be $1{,}000{,}000 - 2{,}000$, or 998,000 normal units of blood. We know from our previous data that the specificity of the test is 99.2 percent; there will be a total of $0.992 \times 998{,}000 = 990{,}016$ normal units that test negative (cell D). Conversely, there will be $998{,}000 - 990{,}016 = 7{,}984$ normal units of blood that have positive ELISA tests (cell C). The new data appear in Table 3-10.

TABLE 3-10 Prevalence of AIDS per 1,000,000 Units of Blood

		Gold Standard (Western Blot)		
		Antibodies	No Antibodies	Total
ELISA	Positive	1,960 (A)	7,984 (B)	9,944
	Negative	40 (C)	990,016 (D)	990,056
	Total	2,000	998,000	1,000,000

If we now recalculate the predictive values, they look like this:

$$\text{Positive Predictive Value} = 1{,}960/9{,}944$$
$$= 19.7 \text{ percent}$$

$$\text{Negative Predictive Value} = 990{,}016/990{,}056$$
$$= 99.99 \text{ percent}$$

The picture is now very different than in the first situation. If a person has a negative test, there is virtual certainty that he/she truly is AIDS-negative. However, a positive test is nearly uninterpretable because more than 80 percent of the positive test results come from people who don't have antibodies!

In actual practice any blood that tests positive is sent for a repeat ELISA and a Western blot test. If ELISA remains positive and the Western blot is negative, the blood is discarded but the donor is not told. If they are both positive, the donor is informed and contacts traced.

Thus in general, the prevalence of disease has a profound effect on the usefulness of a test. If the prevalence is low, the positive predictive value of the test is low and the negative predictive value high. Conversely, if the prevalence of disease is very high, the negative predictive value is low but the positive predictive value is high.

BAYES' THEOREM

In the previous discussion we calculated the probability that a person with a positive ELISA had AIDS antibodies, given known data about the prevalence of antibodies and the characteristics of the test. However, we had to take a roundabout route by calculating a new contingency table (see Table 3-10) and then working out the appropriate values. There is an algebraic shortcut, called **Bayes' theorem**, that permits this calculation directly. To do the calculation, we will also introduce some new symbols that frequently appear in the epidemiologic literature:

$$\begin{aligned}
P(D) &= \text{Probability of disease before the test} \\
&= \text{Prevalence} = 0.2 \text{ percent} \\
P(T+ \mid D) &= \text{Probability of positive test given the disease} \\
&= \text{Sensitivity} = 98.0 \text{ percent} \\
P(T+ \mid \bar{D}) &= \text{Probability of positive test given no disease} \\
&= (1 - \text{Specificity}) = 0.8 \text{ percent} \\
P(T- \mid D) &= \text{Probability of negative test given the disease} \\
&= (1 - \text{Sensitivity}) = 2.0 \text{ percent} \\
P(T- \mid \bar{D}) &= \text{Probability of negative test given no disease} \\
&= \text{Specificity} = 99.2 \text{ percent}
\end{aligned}$$

According to Bayes' theorem, the probability of disease given a positive test, $P(D \mid T\bar{v})$ (i.e., the positive predictive value), is as follows:

$$P(D \mid T+) = \frac{P(D) \times P(T+ \mid D)}{P(D) \times P(T+ \mid D) + [1.0 - P(D)] \times P(T+ \mid \bar{D})}$$

$$= \frac{0.2 \times 98}{(0.2 \times 98) + (0.8 \times 99.8)} = \frac{19.6}{99.44} = 19.7 \text{ percent}$$

A similar calculation could be done to get the negative predictive value.

Bayes' theorem can also be used in an iterative fashion. If we had a situation involving a series of laboratory tests, we could now calculate the posttest probability for the second, third, and subsequent tests. In each case we would use the calculated posttest probability from the previous test as the pretest probability for the calculation of the next test.

RECEIVER OPERATING CHARACTERISTIC CURVES

One measure that is frequently employed for evaluating the effectiveness of diagnostic systems is the **receiver operating characteristic (ROC)** curve. Particularly popular in radiology, it has roots in electrical engineering and psychophysics.

Imagine a laboratory test that has continuous values, such as cardiac enzymes, and consider the problem of attempting to find an appropriate cut-point where any value above the point is considered a positive (i.e., indicative of myocardial infarction), and any point below is considered negative or normal. If we set the point too high, we will miss a number of mild myocardial infarctions, but will avoid false positives. Conversely, a point set too low will catch all the myocardial infarctions at the cost of filling cardiac care unit beds with normal (non-myocardial infarction) patients. This situation is illustrated in Figure 3-1.

As we move the cut-point from right to left, we will initially pick up true positives and few false positives. However, as we pass the center of the myocardial infarction distribution, the rate of pickup of the false positives will increase, and the true positives decrease, to the point that nearly all the increase is false positives. Plotting the true positive rate on the Y-axis and the false positive rate on the X-axis, we generate the ROC curve, as in Figure 3-2.

The ROC curve has some interesting features. First, we note that a perfect test would pick up only true positives at first, then after the true positive rate is 100 percent, only false positives; this describes a curve going vertically along the Y-axis and then horizontally along the top. Conversely, a useless test picks up both true and false positives at the same rate, and traces out a line at 45 degrees. The extent to which the ROC curve "crowds the corner" is a measure of the value of the test. This is measured by the area between the curve and the 45 degree line. Second, the best cutoff to minimize overall errors occurs when the tangent to the line is at 45 degrees; displaying the data this way therefore permits a rational selection of cutoff. The advantage of the ROC approach is that it permits a clear separation between the intrinsic value of the test, as captured in the area under the curve, and the errors associated with an inappropriate choice of cutoff.

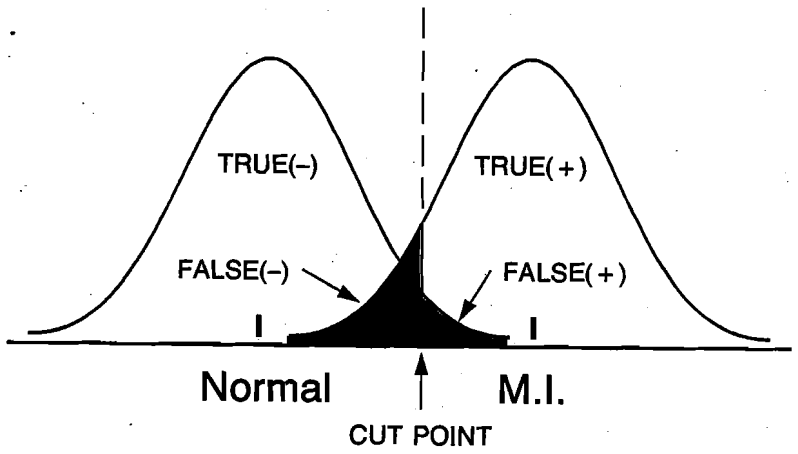


Figure 3–1 Determining the cut-point of a test for myocardial infarction.

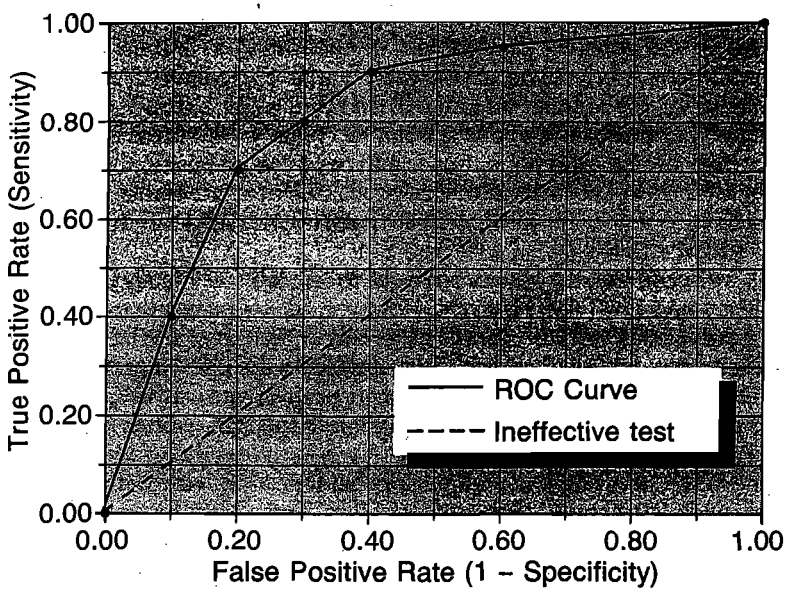


Figure 3-2 ROC curve.

ACCURACY

As yet we have not considered any measure of the overall accuracy of the test. One approach that is very straightforward is to simply sum up the numbers on the diagonal of the table, cells A and D, and place them over the total of all cells. Let's use the data from our two AIDS examples (Tables 3-9 and 3-10).

The overall accuracy of the test, based on data from Table 3-9, is (498 + 488)/1,000 = 98.6 percent. For the lower prevalence situation in Table 3-10, the accuracy is (1,960 + 990,016)/1,000,000 = 99.198 percent. Even though the test is much less useful in the low prevalence case, the accuracy has improved, since the huge numbers of true negatives have predominated in the calculation of accuracy. Because of the possibility of misleading results from this approach, most assessments of accuracy are performed by correcting for chance agreement using a statistic called Cohen's Kappa.

CHANCE CORRECTION USING COHEN'S KAPPA

As we have just seen, the likelihood of agreement between a test result and a "gold standard" is affected by the prevalence of disease. In the extreme case we could consider the application of a clinical sign, right-handedness, to a classical "disease" of Victorian times—self-pollution, or masturbation. Right-handed people are in the majority with about 90 per-

TABLE 3-11 Prediction of Depression from Test Results

Test Results	Depression: Present	Depression: Absent	
Positive	18 (A)	7 (B)	25
Negative	12 (C)	63 (D)	75
	30	70	100

cent of the population. If we are in a population where everyone does "it," the test will be right 90 percent of the time, without conveying any information whatsoever.

To avoid this trap it is desirable to correct for chance agreement. Taking a little less extreme example, consider the data in Table 3-11, which predict depression as diagnosed by expert interview using DSM-III criteria, from a self-completed questionnaire.

The accuracy, as determined before, is (A + D)/N = (18 + 63)/100 = 81 percent. What agreement would we expect by chance? Chance means that there is, in fact, no association between the two variables. Consider first the A cell. We know that on the average 30 percent of all people in the sample have depression, or 30 people. If there is no association between the two variables, we would expect that the same proportion of people with and without depression would have a positive test, simply equal to the overall proportion of positive tests, or 25 percent. So by chance, 25 percent of the 30 depressed people, or 7.5 people, would be in cell A. Similarly, there should be 75 percent of the nondepressed people, or 52.5 people, in cell D. The agreement expected by chance is (7.5 + 52.5)/100 = 60 percent. We actually observed 81 percent. It's not necessary to figure out the numbers in cells B and C because we don't use them in the calculation. The chance corrected agreement, called **Kappa**, is defined as:

$$\text{Kappa} = \frac{\text{Observed agreement} - \text{Agreement by chance}}{1.0 - \text{Agreement by chance}} = \frac{0.81 - 0.60}{1.0 - 0.60}$$

$$= 0.21/0.40 = 0.525$$

As a result, the agreement corrected for chance has been reduced from 81 percent to 53 percent.

MEASUREMENT WITH CONTINUOUS VARIABLES

Historically, epidemiology was concerned with the distribution in time and place of disease epidemics; in more recent times clinical epidemiology has focused on the testing of therapies directed to prolonging life by reducing the incidence of such catastrophic events as heart attacks and strokes. In these situations the unit of analysis is the case of disease or death, and measurement issues focus on the verification of presence or absence of disease.

However, physicians and epidemiologists are increasingly coming to recognize that, for many diseases, there is little to be gained in *quantity* of life from foreseeable advances in biomedicine and there is much more potential for gain in *quality* of life. Innovations such as palliative care and geriatric medicine are explicitly not directed to the cure of disease or extension of life; rather, they are an attempt to improve the quality of life.

From the perspective of epidemiology, research in this area presents new measurement challenges. The measurement of quality of life is a new science; different methods proliferate, and seldom yield the same results. There is possibly more error of measurement than might be expected in categorical measures like diagnosis. Conventional approaches to evaluation of measures, such as comparison with a "gold standard," are inapplicable, because no such criterion currently exists, and no clinical equivalent of the autopsy or biopsy will ever be available. Epidemiologists must acquire new skills, borrowed from such disciplines as psychology, education, and economics, in order to understand and contribute to the development of these measures.

With rare exceptions, these outcomes are based on continuous measurement, originating in rating scales or checklists completed by observers or patients. Approaches to the measurement of association with these measures involve unfamiliar concepts like reliability and construct validity. Usually analysis is conducted using parametric statistics, which assume an interval level of measurement, and normal (bell-shaped) distributions. This section briefly reviews some of these concepts. We are not trying to be comprehensive; instead, we will recommend additional readings for readers who wish to venture further.

MEASURES OF ASSOCIATION

To examine the issues of measurement with continuous variables, we will use an example from rheumatology. The issues here are prototypical of the issues we raised in the beginning of this discussion. The diseases of rheumatology — rheumatic arthritis, osteoarthritis, ankylosing spondylitis, and lupus — are rarely fatal, but often are severely incapacitating because they inflict pain, deformity, and dysfunction on their victims. To examine the efficacy of their therapies, rheumatologists have developed a large number

of measures of disease severity. Some emerge from the laboratory, such as erythrocyte sedimentation rate and rheumatoid factor, but appear to have little relationship with clinical measures of function. Some appear to be "objective" clinical descriptions of disease process, such as counts of involved joints or erosion counts (from observations of bone erosions on hand roentgenograms) and walk times. On closer scrutiny, however, these objective descriptions appear to have a great deal of variation among observers and relatively little relationship with measures of the patient's function. Finally, some measures are based on the patients' own assessment of their function and health, and run the gamut from a simple 10 cm line (called a visual analog scale, presumably to obscure its simplicity) on which the patient puts a mark to indicate his perceived health, to indices of function containing tens or hundreds of questions.

To make sense of this potpourri, it is essential to review empirical evidence that the measures are doing what was intended by their makers. When these questions are examined, the evidence falls into two broad classes. The researcher assessing **reliability** asks whether the measures are giving the same answer over different situations (e.g., different observers or the same observer on two occasions separated by a short time interval). The researcher studying **validity** asks whether the measure is assessing what is intended. Does the index of function related by the patient really assess function, or is the score related to the patient's mood, social status, or whatever?

Because the measures are continuous, we cannot simply place the data into a 2×2 table as we used before. (We could do this, but the shoehorn act comes at an awful cost of loss of information, e.g., any height above 5'6" [168 cm] is classified as tall.) Instead we must measure the degree to which an individual who is high on one measure or occasion is high on a second measure or occasion, and the converse. The methods to develop these measures are explored further in the next discussions.

PEARSON CORRELATION

By far the most common measure of association for continuous variables is the **Pearson product-moment correlation**. It was invented in the early 1900s by one of the founders of modern statistics. The correlation is based on the idea of fitting the data by a straight line, as illustrated in Figure 3-3.

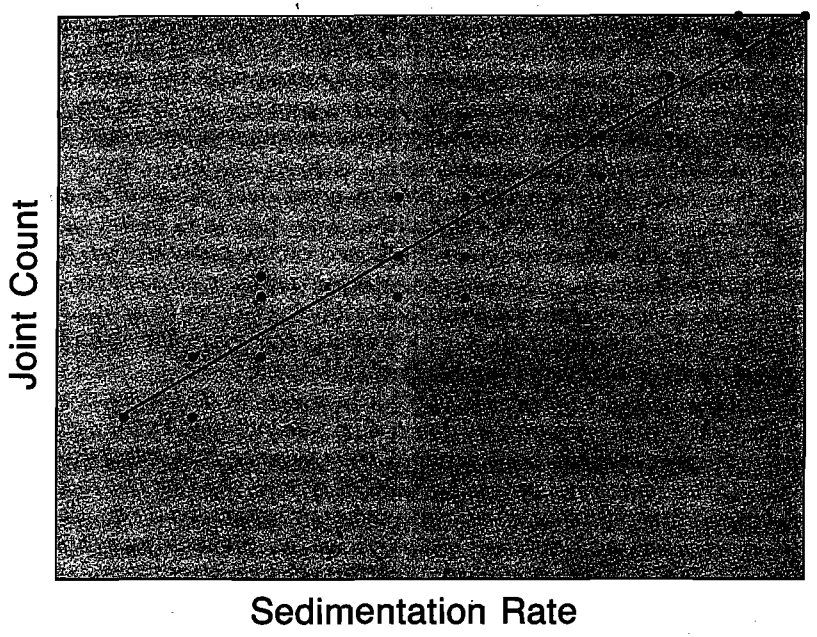


Figure 3-3 Association between erythrocyte sedimentation rate and a measure of active joints in patients with rheumatoid arthritis.

The Pearson correlation is a number between −1 and +1. It equals 0 if there is no relationship, and 1 if there is a perfect linear (straight line) relationship. There is one minor addition: if the slope is negative, that is, if the joint count decreases with increasing sedimentation rate, the correlation is preceded by a minus sign. Therefore a perfect negative relationship has a correlation of −1. Pearson correlations of various sizes are pictured in Figure 3-4.

As you can see, the more the individual points deviate from the straight line, the lower the correlation. With a perfect correlation (+1 or −1), all the points fall on the line. It should be evident from Figure 3-4 that a correlation of 0.8 indicates a fairly good association. Conversely, a correlation of anything less than 0.3 is hardly worth the excitement, statistically significant or not.

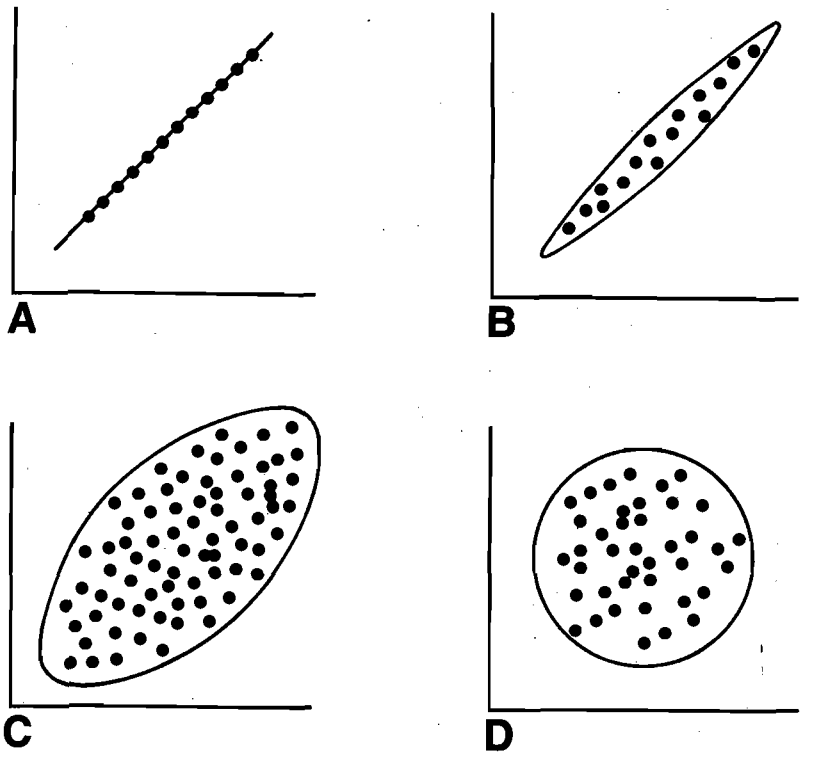


Figure 3-4 Correlations of various sizes. *A*, r = 1; *B*, r = 0.9; *C*, r = 0.5; *D*, r = 0.

INTRACLASS CORRELATION

The Pearson correlation is a perfectly appropriate measure of association to express the degree of linear relationship between two variables. However, under certain circumstances we demand a more stringent measure of association. This situation usually arises in the measurement of agreement between observers, when we don't simply want assurance that a patient scoring high by one observer will also be scored high by the other observer; we want to be sure that the observers are actually giving similar numbers.

Suppose we recruited two rheumatologists to examine hand joints on a series of patients with rheumatoid arthritis and work out the total number of inflamed or swollen joints (Fig. 3-5). It could happen that one observer set very much lower thresholds for what he chose to call "inflamed" than the other, so that for every patient his total was exactly two more (i.e., if one observer said 12 joints, the second said 10, and if one said four, the other said two).

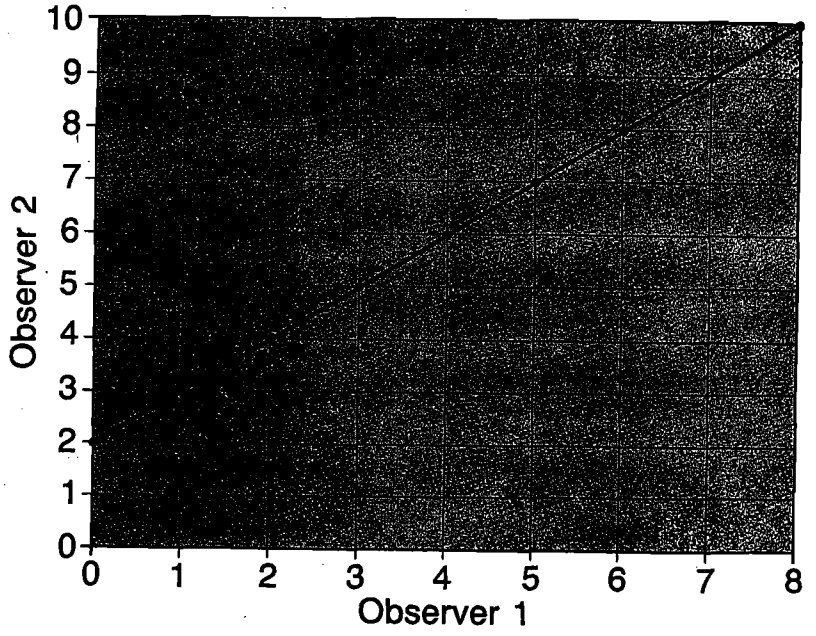


Figure 3–5 The number of hand joints judged as inflamed by two rheumatologists for various patients.

The Pearson correlation simply demands that there is a strong association between the raters — the highest scoring patients for Observer 1 are also the highest for Observer 2, and the lowest for Observer 1 are the lowest for Observer 2. Since this is the case and the points all lie exactly on a straight line, we would get a Pearson correlation of +1. However, by most standards the agreement is terrible; the observers never give the patient the same count.

To get around this problem, the Pearson correlation has been replaced in most circles by the **intraclass correlation (ICC)**. The intraclass correlation is still expressed as a number between 0 and 1; however, the ICC measures not only the *association* between the raters, but also the *agreement*.

Although much is made of the difference between association and agreement and the relative advantages of the intraclass correlation over the Pearson correlation, in most real-world situations the major variability in the data is from apparently random error. Under these circumstances the two measures give identical results. Furthermore, if we treat a 2×2 table as a series of points having values of (1,1), (0,0), (1,0), or (0,1), the intraclass correlation and Kappa yield identical results. For once we can get convergence among differing approaches.

RELIABILITY

Reliability is, as we indicated, a measure of the extent to which a measure is reproducible, or gives the same results, over different situations (e.g., different observers or different days). However, this reproducibility is defined in a very special way by comparing the variability across situations (error variance) to the true variability among patients (patient variance). The reliability coefficient is defined as follows:

$$\text{Reliability} = \frac{\text{Variance due to patients}}{\text{Variance due to patients + Error variance}}$$

In other words, the reliability expresses the proportion of the variability in the measures that is caused by true variability among patients. The implication of this definition is that if the patients we are studying are truly homogeneous with respect to the attribute of interest, the reliability of the measure will be near 0; conversely, if there is great variability among patients, there will likely be higher reliability. The reliability is a measure of the extent to which we can *differentiate* among patients on a particular attribute.

Although this definition is a bit hard for egalitarian folks to accept, it rests on the simple premise that the goal of measurement is to distinguish among people on a particular attribute. If all the people in the population have the same value of a particular quantity, why bother to measure it? Simply assume that the next person will have that value too.

It is not too difficult to demonstrate that the phenomenon is completely analogous to the discussion about the effect of prevalence on the performance of a diagnostic test. Reliability is like the chance corrected accuracy of a test. If the prevalence of disease drops, this is analogous to the patient population becoming more similar, and the reliability of the continuous measure and the accuracy of the test both fall.

There are some other terms usually associated with reliability, most of which are self-explanatory. **Interobserver** reliability examines the degree of agreement among different observers. **Test-retest** reliability involves administering a test or measure to a group of patients on two different occasions and examining the correlation. **Split-halves** reliability is used in longer tests, and involves splitting the test items into two halves at random and examining the correlations between subscores from the two halves of the test.

There are a number of other specific forms of reliability, but this should give you the idea.

VALIDITY

FACE AND CONTENT VALIDITY

Having demonstrated that a measure is reproducible, it remains to be shown that it is measuring what is intended. Sometimes this is straightforward and noncontroversial; for example, to show that a mercury manometer

is measuring blood pressure validly, one might compare blood pressure values obtained this way with direct measures of arterial blood pressure.

More frequently the situation is not so straightforward. How do you demonstrate that your new measure is really assessing self-concept, illness behavior, locus of control, or quality of life? One way is to argue that it measures trait X because trait X is what it measures, an argument that is invoked in a variety of forms for the measurement of intelligence. However, this approach is a little circular; a variation on the theme that is a little less egocentric is to approach a group of experts and ask them whether the measure looks like a reasonable measure of the concept as they understand it. This approach is termed **face validity**. You could also ask them if the measure appears to contain all the important concepts, behaviors, and elements of the concept. If the answer is "yes," you have also attained content validity.

There are better approaches to the measurement of content validity. For example, you might observe patients to see behaviors, interview them or review records, or base the instrument on previous reported measures. All of these strategies are appropriate to ensure that the measure contains the desired content. However, in the final analysis the assessment of face and content validity is, with rare exceptions, based on the opinion of experts. Since "old boy" networks are the norm in most academic disciplines, in that we associate with people who think like we do (i.e., correctly), these must be regarded as weak tests of validity.

CRITERION VALIDITY

As we indicated, measures of validity based on expert judgments are regarded in general as weak tests of validity. Perhaps the strongest approach to validity is the assessment of **criterion validity**, which involves comparison with a "gold standard." In turn, this is divided into two forms that differ only in time. If the comparison is made at the same time (i.e., both measures are administered together) the approach is called **concurrent validity**. If the measure is used to predict future status, such as confirmation of a disease at autopsy or admission to hospital, it is called **predictive validity**. The index of criterion validity is most often a correlation coefficient between the scores on the new test and on the old (or predicted) one.

The comparison of blood pressure reading with a mercury sphygmomanometer with arterial blood pressures is an example that highlights both the use of a "gold standard" and the reason for developing a new measure, namely, reduced cost or risk. However, such true "gold standards" are difficult to come by, and one is frequently left in the situation of comparing the new measure with another better accepted, but arguably inferior measure of the same attribute. One example of this is the measurement of depression. Although new measures proliferate, nearly all are compared with one of two scales — the Beck Depression Inventory or the CES-D scale. Since both standards are short and cheap, the only reason to develop a new measure is that it will be better; however, this is difficult to prove by simply

comparing with existing measures. Under these circumstances the expected correlation of the two measures should be high, but one would not anticipate correlations too close to unity; if it were nearly 1.0, the two tests are measuring almost exactly the same thing and there is little reason to develop the new one.

CONSTRUCT VALIDITY

Probably the most frequently applied, but poorest understood measure of validity is called **construct validity**. It is used in circumstances in which there is no other measure of the attribute under study. Instead of testing the relationship between the new measure and some other measure of the same thing, we invoke a theoretical construct that describes the relationship between the attribute under scrutiny and other attributes. We then examine the relation between these two measures, and if it is in the expected direction, we have evidence that both the measure and the hypothetical construct were right. However, if there is no relationship between the two measures, we have no way of determining whether our measure or our theory was wrong.

For example, if we are developing a measure of quality of life of patients with rheumatoid arthritis, we might hypothesize that the measure is strongly related to measures of function like morning stiffness or walk time, and relatively poorly related to measures of disease process like joint counts or sedimentation rate. Further, since we would hope that it is a relatively pure measure of the effect of the particular disease on perceived quality of life, we may further hypothesize that scores are uncorrelated with measures of depression. Finally, we can examine hypotheses about differences between groups, for example, that inpatients are likely to score poorer than outpatients.

It is evident that in the construct validity game there is no single study or hypothesis that clinches the case. Some hypotheses will be right and some will be wrong. Rather, the judgment of validity depends on the *weight* of the evidence being in the expected direction.

MEASUREMENT BIAS

In the previous section on research methodology we described how incorrect conclusions may result from design flaws. Biases such as the Berkson's or Neymann bias can yield estimates that are systematically higher or lower than the true value.

Unfortunately, research design errors are not the only source of bias. Large distortions can result from biases in measurement. There are innumerable sources of measurement bias; many psychologists have made careers out of cataloging how people can be induced to distort their

estimates one way or another. One of the most disturbing examples derives from choices resembling the following:

> "You are responsible for the care of 100 patients who have a fatal disease. You are given a choice between two drugs: Drug A has a 60 percent chance of saving everyone; Drug B will save 60 of the 100 patients. Which will you choose?"

Under these circumstances, most subjects choose Drug B. However, the question can be framed in the logically equivalent way:

> "You are responsible for the care of 100 patients with a fatal disease. You are given a choice between two drugs. Drug A has a 40 percent chance of killing all of the patients. Drug B will result in the death of 40 of the 100 patients. Which will you choose?"

When the question is framed in this way, most respondents choose Drug A. Obviously the way a question is asked can lead to radically different responses. There are many other ways that data can be willingly or unwillingly distorted by unsuspecting investigators. Our purpose will be served by illustrating a few.

However, before we illustrate a few different kinds of bias, let's distinguish between two concepts — bias and random error. **Bias** is a *systematic* deviation from the correct value of a particular variable. The effect of bias is to distort the estimate of that variable, for example, to increase the sample mean or decrease the prevalence of some trait. In **random error**, on the other hand, there is also a deviation from the true value, but because it is random the deviation sometimes adds to the estimate and sometimes takes from it. In the long run (i.e., with a lot of subjects) these deviations cancel each other out. The effect is to increase the variability of the scores, but random error does not affect the estimate of the variable. For this reason random error can be dealt with by statistics. Since bias is a *consistent* distortion from the true value, it cannot be corrected by any statistical manipulation, and thus is more insidious.

DIAGNOSTIC SUSPICION BIAS

Under certain circumstances the rate of occurrence of a diagnosis can depart from expectations simply because of an enhanced index of suspicion on the part of the diagnostician. This bias may be highly individualized and short term. One well-documented bias of individuals is illustrated by the clinical anecdote that goes something like this: "The funniest thing happened. Saturday night in the ER I diagnosed the first case of Somaliland Camelbite Fever I've seen in 20 years. This week I saw four more cases in my office. There must be a real epidemic going around!" A more likely explanation is the *availability bias*. The one case in the ER is readily available in memory, and is likely to be recalled when anything similar comes along.

A more long-term and widespread diagnostic suspicion bias is the *syndrome syndrome*. Over the decades it is easy to show how the popularity of certain diseases has waxed and waned. In the 1920s a common syndrome was "self-pollution", or masturbation. The clinical syndrome was well described, and there were literally institutions filled with depraved little self-polluters. In fact, W.K. Kellogg ran a sanatorium for these lost souls in Battle Creek.

Lest you feel this is a perversion of the early days of medicine prior to the advent of sophisticated diagnostic procedures, there are many current examples. Alzheimer's disease has apparently reached epidemic proportions. Some of the increased incidence is a result of better diagnostic tools and more old people around to get it. Nevertheless the syndrome was first described in the early 1900s. Presumably, until recently doddery old ladies were simply passed off as doddery old ladies. Now, if anyone over 65 forgets where they left their car keys, Alzheimer's is the first diagnosis to spring to mind.

We also alluded to the ureaformaldehyde foam insulation (UFFI) issue earlier. The interesting tale about UFFI is that it was installed for several decades in Europe prior to its arrival in North America. Once here, relatively few problems arose until the media announced all the lethal consequences of the stuff. Following that point physicians everywhere were diagnosing any number of complaints, from headaches to ingrown toenails, as resulting from UFFI poisoning.

In the *Preface* we mentioned one study in which physicians "found" tonsillitis requiring surgery in about 45 percent of kids, even when two other sets of physicians declared the kids clean (or at least not ill). Here again, the expectation of finding a disorder biased what was seen.

SOCIAL DESIRABILITY BIAS

Personality psychologists now routinely include a **social desirability** scale in many of their measures. The notion is that people, when asked sensitive questions about, for example, alcohol consumption or sexual practices, will consciously or unconsciously bias their responses toward the socially acceptable answer. If the bias is deliberate and conscious, it is called "faking good," and if unconscious, "social desirability." In either case the results are the same — an underestimate of the true prevalence of undesirable behaviors.

Several techniques have been developed to detect the presence of social desirability and to fix it if present. Many psychological scales contain imbedded social desirability scales; for example, only saints can truthfully answer "true" to the statement "I have never stolen anything." Alternatively methods such as the random response technique are designed to elicit better measures of the prevalance of unacceptable behaviors.

C.R.A.P. DETECTORS

C.R.A.P. DETECTOR III-1

Question An investigation of the usefulness of exercise ECGs was conducted using patients who had been admitted to a coronary care unit (CCU). The ECG was compared with findings from coronary angiography—a very expensive and risky procedure. For obvious reasons the researchers had difficulty recruiting a large number of "normal" subjects to undergo angiography. So they took 80 men off the street, did ECGs on them (which were normal of course), assumed that they would have normal angiograms, and added them to the negative ECG-negative angiogram category. The results looked very good indeed: sensitivity was 64 percent and specificity was 93 percent. Subsequent applications of exercise ECGs in ambulatory settings have shown that it is not what it was cracked up to be, and show a sensitivity of only 33 percent. Why does the discrepancy exist?

Answer The authors did two things to ensure that the results would look favorable. First, the positive cases were chosen from a highly select group of men in a CCU with confirmed cardiac disease, so they were more extreme than the usual suspected arteriosclerosis. Second, the initial study had too high a prevalence of disease. By including the "normal" volunteers and, better still, assuming that they had normal angiograms, they succeeded in messing the base rates in their favor.

Beware the "sample samba." By dancing around with prevalence, or by selecting extreme groups (e.g., phys. ed. students and 70-year-olds on their third myocardial infarction), anyone can make any test in the world look good.

C.R.A.P. DETECTOR III-2

Question A recent reanalysis was conducted of the Blair et al. National Cancer Institute study of the occupational effects of formaldehyde on cancer. They were unable to show any significant relationship between formaldehyde level and lung cancer, but did demonstrate a relationship between job class and cancer. They concluded that the retrospective measurement of formaldehyde was too crude, and that blue collar workers suffered more lung cancer as a result of occupational exposure to formaldehyde. The study was not published (thank goodness!). Why?

Answer The measurement of formaldehyde level may have been crude, but the use of job class as a surrogate for exposure ignores the many other variables that go along with job class. First, blue collar workers smoke more, and smoking causes lung cancer. Second, lower social class folks suffer more disease of all types, and live less long than upper class folks.

Correlation is not equal to causation. (See *Assessing Causation*).

C.R.A.P. DETECTOR III-3

Question In a study of the causes of cervical cancer one potential cause under investigation was whether or not the man was circumcised. The researchers approached 166 males and asked whether they were circumcised. This was then confirmed by a physical examination. Of the 44 men who said they were, 21 (48 percent) were not, and of the 122 men who said they were not circumcised, 50 (40 percent) were! Don't men know whether or not they are circumcised?

Answer Self-report may be a lousy lab test. If an investigator is using self-report data, there should be some assurance (other than faith!) that the data are valid.

C.R.A.P. DETECTOR III-4

Question For about two decades, patient management problems (PMPs) have been used as a component in the certification examination used to license physicians in Canada and the U.S. These are written simulations of a patient, on which the candidate selects options on history, physical, laboratory, and management and is rewarded (or punished) on the basis of the good options he selected and harmful options he avoided. Many studies demonstrated that candidates felt the method to be life-like (face validity), and care was taken to ensure that they were medically accurate (content validity). They have also been used as a measure of problem-solving skills. This was confirmed by a low correlation of PMP results with tests of knowledge, which suggested that they were measuring "something else" (construct validity). Can PMPs be considered to be good predictors of physician performance?

Answer Recent studies showed a very low reliability of the scores, which suggests that the "something else" they were measuring was simply noise. Other studies showed that candidates do about twice as much of everything (such as ordering lab tests) on the written problem as they do in real life. Both licensing bodies have subsequently dropped the requirement for performance on PMPs.

Face and content validity are poor substitutes for empiric forms of validity. Anyone can recruit some friends who will like his/her measure. The best test of validity is criterion-related validity. All others are relatively weak approximations.

REFERENCES

ISSUES IN CHOOSING A MEASURE

Lord Kelvin. Quoted in: Sears FW, Zemansky MW. College physics: mechanics, heat and sound. Reading, MA: Addison-Wesley, 1952.

MEASUREMENT WITH CATEGORICAL VARIABLES

Doll R, Peto R. The causes of cancer. Oxford: Oxford University Press, 1981.
Lipid Research Clinics Program. The Lipid Research Clinics coronary prevention trial results. JAMA 1984; 251:351-374.
Wynder EL, Graham EA. Tobacco smoking as a possible etiologic factor in bronchogenic carcinoma: a study of 684 proved cases. JAMA 1950; 143:329-336.

Diagnostic Tests

Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70:213-220.
Lusted L. Medical decision making. Springfield: CC Thomas, 1968.
Meehl PE, Rosen A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychol Bull 1955; 52:194-216.
Polesky HF. Serologic testing to human immunodeficiency virus. MMWR 1986; 36:833.

MEASUREMENT WITH CONTINUOUS VARIABLES

American Psychological Association. Standards for educational and psychological testing. 3rd ed. Washington: APA, 1985.
Crowne DP, Marlowe D. A new scale of social desirability independent of psychopathology. J Consult Psychol 1960; 24:349-354.
Eraker SA, Sox HC. Assessment of patients' preferences for therapeutic outcomes. Med Decis Making 1981; 1:29-39.
Warner SL. Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 1965; 60:63-69.

C.R.A.P. DETECTORS

Blair A. Mortality among workers exposed to formaldehyde. J Natl Cancer Inst 1985; 75:1039-1047.
Goldschlager N, Selzer A, Cohn K. Treadmill stress tests as indicators of presence and severity of coronary artery disease. Ann Intern Med 1976; 85:277-286.
Dunn JE, Buell P. Association of cervical cancer with circumcision of sexual partner. J Natl Cancer Inst 1959; 22:749-764.
McGuire CH, Babbott D. Simulation technique in the measurement of problem-solving skills. J Educ Meas 1967; 4:1-10.

TO READ FURTHER

MEASUREMENT WITH CATEGORICAL VARIABLES

McMahon B, Pugh TF. Epidemiology: principles and methods. Boston: Little, Brown, 1970.

Diagnostic Tests

McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 1984; 4:137-150.
Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown, 1985.

MEASUREMENT WITH CONTINUOUS VARIABLES

Anastasi A. Psychological testing. 5th ed. New York: Macmillan, 1982.
Kahnemann D, Slovic P, Tversky A. Judgment under uncertainty: heuristics and biases. Cambridge: Cambridge University Press, 1982.
Norman GR, Streiner DL. PDQ statistics. Toronto: BC Decker, 1986.
Norman GR, Streiner DL. Principles of measurement in health sciences. Oxford: Oxford University Press, 1989.