MEASURING AND INTERPRETING MORBIDITY

Steven T. Fleming and F. Douglas Scutchfield

> *"It is time to close the book on infectious diseases, declare the war against pestilence won, and shift national resources to such chronic problems as cancer and heart disease."*
> *US Surgeon General William H. Stewart, 1967*

The current surgeon general would certainly disagree with the above statement, made more than 50 years ago. Despite the massive shift in cause of death from infectious to chronic diseases over the last century, the "war against pestilence" has not been won. Enemies such as HIV/AIDS, methicillin-resistant *Staphylococcus aureus* (highlighted in case study 6.1 in chapter 6), and influenza, among others, are still formidable foes.

Population Health

Epidemiology is a population science and can therefore be distinguished from clinical medicine, which has a focus on the diagnosis and treatment of individual patients. A *population* can be defined as a group of people with at least one distinguishing characteristic—for example, they all reside in a particular region, or they all are of a particular race or gender, or they all were born in a particular year. Once we identify a specific population of people, we can investigate how disease is distributed among them, how common certain risk factors are among them, and whether and to what extent the risk factors increase the frequency of disease among them. We can also ascertain whether the population is a fixed or dynamic population. Membership in a *fixed population* is permanent and typically defined by a life event, such as the terrorist attack on the twin towers of the World Trade Center in New York on September 11, 2001. Membership in a *dynamic population* is transient, with an inflow of people (e.g., births, immigrants, new students), and an outflow of people (e.g., deaths, emigrants, student graduates). The population may be in "steady state," wherein some of the actual people change but the aggregate characteristics and total size

remain relatively constant. Alternatively, the population may be changing in either size or characteristics.

The term *population health* has been embraced by policymakers and academicians alike, although some uncertainty remains regarding the precise definition, and it is primarily a matter of scope. Kindig and Stoddart (2003) suggest that the term *population health* could refer to the multiple determinants of health, the outcomes of health, or both, with the latter being the preferable option and the one espoused by one Canadian advisory committee (Dunn and Hayes 1999, 57):

> Population health refers to the health of a population as measured by health status indicators and as influenced by social, economic, and physical environments, personal health practices, individual capacity and coping skills, human biology, early child-hood development, and health services. As an approach population health focuses on interrelated conditions and factors that influence the health of populations over the life course, identifies systematic variations in their patterns of occurrence, and applies the resulting knowledge to develop and implement policies and actions to improve the health and well-being of those populations.

Kindig is one of the prominent figures in population health in the United States. Drawing on previous work (Evans and Stoddart 1990), he has developed a model of population health planning (Kindig, Asada, and Booske 2008) that relates the medical care, individual behavior, social environment, physical environment, and genetic determinants of health to population health outcomes that include mortality and health-related quality of life (QOL), which is determined by the burden of **morbidity**. He also recognizes that mortality- and health-related QOL disparities by race/ethnicity, socioeconomic status, geographic location, and gender are important aspects for designing specific policies and interventions to improve population health.

morbidity
Any departure from a state of health.

The measurement of morbidity or disease is an important metric for population health, but it is not the only one. McDowell, Spasoff, and Kristjansson (2004) distinguish among four types of population health measures and their applications.

1. *Descriptive.* Health status measures, disability scales, and other measures describe the burden of disease or disability.
2. *Predictive.* A predictive or prognostic application requires measures that anticipate future morbidity burden, such as those derived from screening tests or other indicators of risk.
3. *Analytical.* Analytical epidemiology refers to the extent to which behavioral, social, economic, environmental, and other determinants increase the risk of disease.

4. *Evaluation*. Evaluation measures include health, disease, or functional status indicators that can be used to judge the success of programs or policies by measuring relatively small changes in health at the individual level.

Although the measurement of morbidity probably has some value for each of these applications, it is critically important for descriptive and prognostic applications.

This chapter focuses on the measurement and interpretation of disease statistics. We describe the nature, definition, and classification of disease while making a distinction between disease and illness. We discuss the natural history of disease, or how the disease plays out over time; the various sources of morbidity data; and how morbidity can be measured with incidence and prevalence statistics. Finally, we focus on screening and diagnostic tests; review their characteristics, such as sensitivity and specificity; and present two case studies regarding congestive heart failure investigation and breast cancer screening to illustrate these points.

The Nature and Definition of Disease

Disease can be defined in a number of ways. Merrill (2016, 50) defines disease as "an interruption, cessation, or disorder of body functions, systems or organs." Weiss and Koepsell (2014, 10) define it simply as "almost any departure from perfect health." The single term *morbidity* is defined by Porta (2014, 189) as "any departure subjective or objective, from a state of physiological or psychological well-being," a definition that includes disease and illness.

Many types of diseases affect human beings, thus posing a challenge to classify these diseases in a meaningful and efficient way. One could classify diseases by their means of transmission—for example, those that are airborne (such as influenza), vector-borne (such as malaria), or transmitted through intestinal discharge (such as cholera). One could also classify diseases by their source, such as those caused by microorganisms (such as plague) or inanimate sources (such as radiation or noise). The most comprehensive and widely used method of disease classification is the International Classification of Diseases (ICD), now in its tenth edition (ICD-10-CM), which classifies diseases into more than 65,000 categories. The diseases are categorized primarily by body system—there are separate chapters for the nervous, circulatory, respiratory, and digestive systems, with codes of up to seven digits. For example, with diabetes, the first three digits denote the type of diabetes (e.g., E10, insulin-dependent diabetes). The fourth digit classifies the type of complications (e.g., E102, renal), and a fifth digit denotes specific complications (e.g., E1023, diabetes

mellitus with diabetic renal failure). A sixth digit can further refine complications. A seventh digit is used in a number of situations, such as to distinguish between initial and subsequent encounters (in the case of injuries), to indicate which fetus is involved in the complication (in the case of pregnancies), or to indicate the severity of a coma. Although the tenth edition represents a substantial refinement over the earlier ICD-9-CM (for example, four times as many codes for diabetes alone), the changes in classification and expansion in number of codes pose problems for monitoring changes in morbidity over time.

The Natural History of Disease

natural history of disease
The course of disease over time from onset to resolution.

induction period
The period between exposure and onset of symptoms for noninfectious agents. Also called the *latency period*.

clinical horizon
The point at which signs and symptoms make a disease detectable.

The **natural history of disease** refers to the course of disease over time from onset to resolution, or more simply, how the disease unfolds over time.

The natural history of disease consists of a number of distinct phases (see exhibit 4.1). Exposure to the agent occurs during the *prepathogenesis phase* (before the disease process begins). The *susceptibility phase* denotes the period during which the host may be particularly susceptible to the agent, because of low resistance, poor nutrition, or other factors. If the host is successful in resisting disease, this would occur during the *adaptation phase*. The *pathogenesis phase* consists of the development of disease, with the early pathogenesis phase happening before symptoms occur. One can distinguish the early from the late *clinical phase*, with the former being the period during which diagnosis occurs. The period between exposure and onset of symptoms is referred to as the *incubation period* for infectious agents and as the **induction period**, or *latency period*, for noninfectious agents.

Because most diseases have pre-clinical and clinical phases, we can identify the line of demarcation as the **clinical horizon** (Mosby 2009). Presumably there would also be a "subclinical horizon," where the disease is detectable only through laboratory testing, imaging, or other technologies. Celentano

EXHIBIT 4.1
Natural History of Disease
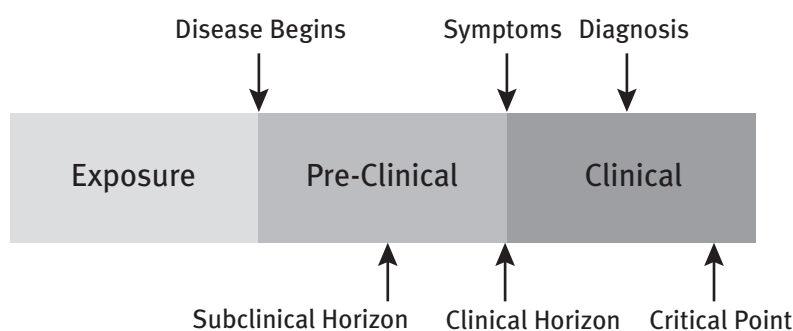
This is an unedited proof.
Copying and distribution of this PDF is prohibited without written permission.
For permission, please contact Copyright Clearance Center at www.copyright.com

and Szklo (2018, 356) define the *critical point* as "a point in the natural history before which treatment is more effective and/or less difficult to administer" and the *detectable pre-clinical phase* as the interval between when the disease could be detected by screening and when symptoms appear. Clearly the length and position of the detectable pre-clinical phase in relation to the critical point have major implications with regard to the efficacy of screening. For a screening program to be effective, the clinical horizon must occur before the critical point. In other words, we must be able to detect disease at a point in its natural history before major consequences occur. If the critical point occurs before the clinical horizon, as is the case with some cancers (e.g., pancreatic cancer), serious consequences occur before the disease can be detected, so screening is ineffective. In such cases, the challenge of medical technology is to move the clinical horizon to an earlier point in the natural history of the disease.

> **critical point**
> The moment in the progression of a disease before which treatment is more effective and after which major or severe consequences occur.

Each disease has a characteristic natural history, though some diseases may share similar manifestations over time. An **acute diseases** is a relatively severe, treatable disease of short duration, with an outcome of either recovery or death. A **chronic disease** is typically less severe and of longer duration, and it often does not conclude with complete recovery but rather progressive disability. A **subacute disease** is intermediate in both severity and duration.

> **acute disease**
> Severe, treatable disease with short duration resulting in either recovery or death.

With these definitions in mind, another way to conceptualize natural history is to graph the progress of disease over time in terms of severity of disease or departure from wellness or health, as postulated in the classic work by Donabedian (1973). With this approach, the natural history would show obvious distinctions among acute but fatal disease; chronic disease leading to progressive disability; acute, self-limited disease that normally concludes with the patient at complete health; chronic disease with relapsing episodes (e.g., multiple sclerosis); and acute asymptomatic disease (i.e., a disease that infects a person but produces no symptoms).

> **chronic disease**
> Less severe disease than acute disease, with a longer duration often resulting in progressive disability.

> **subacute disease**
> Disease with intermediate severity and duration.

Sources of Morbidity Data

Morbidity data can be obtained from a number of sources in the United States and elsewhere. In the United States, physicians and laboratories submit data to the public health departments, which in turn submit these data to the Centers for Disease Control and Prevention (CDC).

Specific notifiable diseases must be reported within certain time frames, depending on the degree to which each is a threat to public health. For example, botulism (a deadly foodborne illness that causes paralysis) must be reported by telephone immediately; hepatitis A, malaria, and measles must be reported within one day; AIDS, mumps, and Lyme disease must be reported within a week. CDC (2020a) publishes regular reports of these diseases and others through the weekly publication *Mortality and Morbidity Weekly Report*.

Clinical records from physicians and hospitals are a second source of morbidity data. Medical records in both settings are a rich repository of morbidity data, though the 1996 Health Insurance Portability and Accountability Act regulations protect the confidentiality of such data, making it more difficult to access. Some observers would argue that medical records represent a biased source of morbidity statistics because people who are treated in healthcare settings may not be representative of the community at large. Some people are sick but do not seek medical attention for various reasons, such as problems accessing care or a lack of insurance. Financial claims derived from patient encounters are a related source of morbidity data, inasmuch as each claim is tagged with one or more diagnoses. Claims are tied to reimbursement by third-party providers, such as Medicare and Medicaid, and as such are driven by financial incentives. The diagnoses associated with each claim may be those that maximize reimbursement rather than those that are most clinically relevant.

Morbidity registries are a third source of morbidity data. Perhaps the most mature registry system in the United States is the network of registries coordinated by the North American Association of Central Cancer Registries. The National Cancer Institute administers the Surveillance Epidemiology and End Results program, which collects registry data from states and cities representing about 25 percent of the US population.

A fourth source of morbidity data is the periodic surveys that are undertaken by various federal, state, and local agencies. The National Center for Health Statistics (NCHS), a part of the CDC (2020b), compiles morbidity information through its numerous surveys, including the National Health Interview Survey, the Health and Nutrition Survey, the National Hospital Discharge Survey, the National Nursing Home Survey, and the National Ambulatory Medical Survey. Much of this information is available to the public and can be downloaded from the center's website.

Measuring Morbidity

Three classes of mathematical parameters are useful to epidemiologists: ratios, proportions, and rates. A *ratio* is obtained by dividing one independent number by another; in other words, the numerator is not a subset of the denominator. For example, among 1,000 motorcycle fatalities, there were 950 men and 50 women. The ratio of fatalities by gender would be 950/50 = 19/1. This doesn't necessarily tell you that men are at higher risk than women, because they may drive motorcycles more frequently, but rather that one might expect 19 times as many men as women to die from motorcycle accidents. The *proportion* is a measure where the numerator is a subset of the denominator. Using the same example, the proportion of male fatalities in motorcycle accidents would be

950/1,000 = 0.95. Finally, a *rate* includes (1) frequency of the event (e.g. disease, death, accident) in the numerator (X), (2) population at risk in the denominator (Y), and (3) time during which the event occurred. Rates are usually expressed as some factor of 10 (k); for example,

$$X/Y \times k$$

For example, suppose that during 2020 there were 12,000 heart disease deaths in Kentucky (X), and the population at risk in Kentucky was 4,400,000 (Y), expressed as a rate per 100,000 (k) as follows:

$$12,000/4,400,000 \times 100,000 = 273 \text{ per } 100,000$$

The burden of illness can be expressed by two kinds of rates: **prevalence** and **incidence**. Prevalence measures cases of disease in a defined population and period. Prevalence rates come in two varieties—point prevalence and period prevalence—and the definitions of these terms vary by source. Porta (2014), for example, insists that *point prevalence* is the proportion of people in a population who have a particular disease or condition at a particular point in time—say, January 1, 2020. Celendano and Szklo (2018), on the other hand, suggest that point prevalence can be expressed as a rate as shown in the foregoing equation, representing 273 heart disease deaths per 100,000 people in Kentucky during 2020. *Period prevalence* measures the number of people with a particular disease during some period (e.g., one year) and expresses this figure as either a proportion or a rate.

On the other hand, *cumulative incidence* examines the number of members of a cohort who have developed disease over a defined period. The *incidence density*, also known as the *incidence rate*, measures how many new cases of a given disease occur in a defined population which is at risk over a certain period and is calculated by dividing the cumulative incidence by a measure of person-time. It is important that the denominator measure the actual population "at risk." (Cumulative incidence and incidence density will be discussed in greater detail in chapter 11. See also Capstone Case B.)

Ideally we should exclude five groups of people who are not or may not be at risk of newly acquired disease, further described in the following ways. First, people who died during a previous period obviously are no longer at risk of incident disease, nor should they be included in the denominator of prevalent disease. Second, if the disease is a chronic disease, people who develop the disease are no longer at risk of incident disease. Third, people who die within the time frame expressed by the measure may be at risk for only part of that period. For example, people who died near the beginning of the year were at risk for

prevalence
Existing cases of disease in a defined population and period.

incidence
New cases of disease in a defined population and period.

only a small part of that year. Convention would have us subtract one-half the number of those who died during the period, with the assumption being that people die randomly throughout the year and, on average, those people were at risk for only one-half of the period. Fourth, some measures pertain only to certain populations. For example, only men are at risk for prostate cancer, and only women are at risk for uterine cancer. Fifth, some people may take steps that remove them from the at-risk population. For example, people who are immunized may no longer be at risk of getting the disease, and women who have undergone a hysterectomy are no longer at risk of getting uterine cancer. Epidemiologists should make the denominator of incidence density as accurate as possible within the constraints of available data.

Consider the example of diabetes mellitus (DM). In 2015, newly diagnosed cases numbered 1.53 million (M), but an estimated 23.6M people were living with DM. The first figure is used to calculate the incidence rate and the second the prevalence rate. So with these numbers, the incidence rate for 2015 was 6.19 per 1,000 adults aged 20 and older, and the prevalence rate was 95.5 per 1,000 adults aged 20 and older:

1.53M/247M (est. US pop. 18 and older) × 1,000 = 6.19/1,000

23.6M/247M (est. US pop. 18 and older) × 1,000 = 95.5/1,000

Clearly, incidence and prevalence are related. Exhibit 4.2 illustrates this relationship. The faucet dripping into the bowl represents the incidence. That is how the addition of new cases is accumulated in the prevalence rate. The volume in the bowl represents the prevalence. The faucet representing the outflow from the bowl reflects the fact that people with the illness either die or get well and no longer are in the prevalence bowl. Thus, prevalence and incidence are related by duration. The relationship is specified as prevalence (P) = Incidence (I) × duration (D). Although this simple relationship does make some assumptions regarding the stability of these rates, it can be a useful metric to predict the average life duration of people with the disease. For example, assume that the incidence of lung cancer is 610.8 per 100,000 and the prevalence is 306 per 100,000. If $P = I \times D$, then $D = P/I$, or 306/610.8 = 0.5. In other words, using the assumptions in this example, the average survival of lung cancer patients is 0.5 years, or 6 months.

Exhibit 4.2 illustrates that an increase in prevalence may be the result of increases in incidence, duration, or both. If you know two terms of the equation, you can solve for the third term.

When collecting morbidity statistics, the epidemiologist must decide what to do with *recurrent* cases of disease—that is, disease that occurs more than once in the same person. Both the nature of disease and the period between
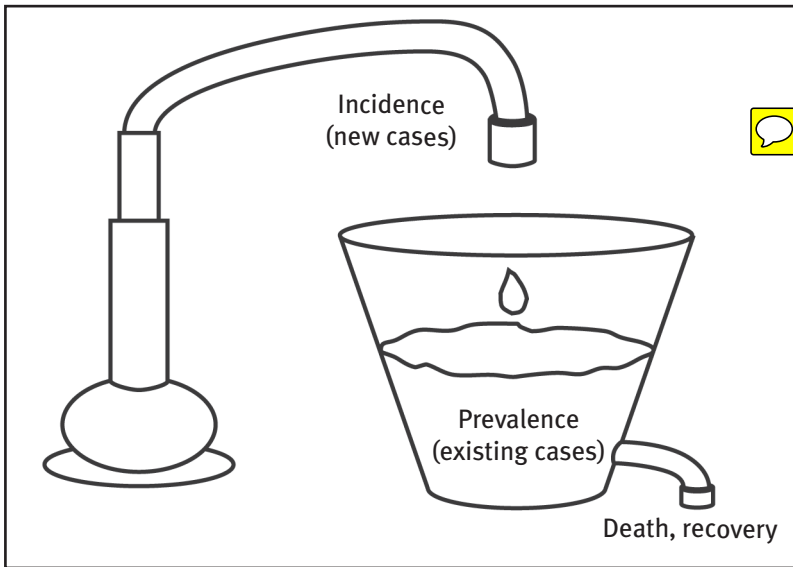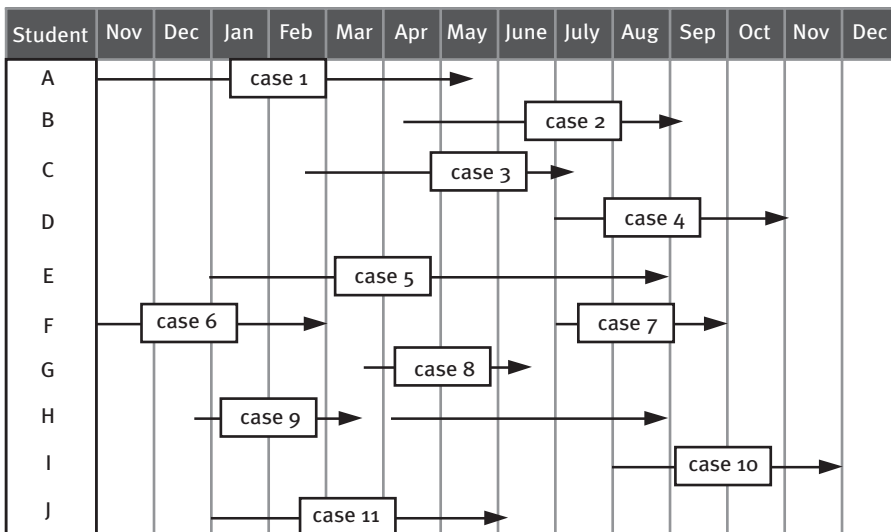
episodes are important. If the nature of the disease lends itself to recurrences (e.g., gonorrhea), then recurrent cases are usually counted as new cases, as long as the period between disease episodes is long enough to treat the recurrent case as a new case rather than a relapse of an earlier illness.

To illustrate these points, consider exhibit 4.3, which summarizes the occurrence of a particular disease (say, gonorrhea) among 1,000 college

Student	Nov	Dec	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
A				case 1										
B							case 2							
C					case 3									
D									case 4					
E				case 5										
F		case 6						case 7						
G						case 8								
H			case 9											
I										case 10				
J				case 11										

students. Gonorrhea often recurs, so we can specify the time window between episodes (say, one month). The incidence rate is calculated as the number of new cases over a specified period (say, one year) divided by the population at risk (in this case, 1,000 students). In exhibit 4.3, we see that the new cases for 2020 include cases 2, 3, 4, 5, 7 (recurrence), 8, 10, and 11. (Student F suffered a recurrence [case 7], because the second episode occurred outside the specified time window. This would count as a new incident case rather than a continuation of the earlier case [case 6]. Student H suffered an episode within the time window. We presume this to be the same case [case 9], rather than a new case.) Thus, there are eight new cases, divided by an at-risk population of 1,000, for an incidence rate of 8 per 1,000.

The period prevalence rate is the number of existing cases during a specified period, regardless of whether they are continuing, new, or recurring. Because cases that recur within the specified time window count, this calculation would include all 11 cases, for a period prevalence rate of 11 per 1,000. The point prevalence rate is evaluated at a particular point in time (say, July 15, 2020). At that point, there were five existing cases, for a point prevalence rate of 5 per 1,000, or 0.5 percent.

Incidence and prevalence rates can be used to characterize patterns of disease by age group. The work of Valanis (1999) describes the kinds of disease and related issues and problems by age groups including (1) pregnancy/infancy, (2) childhood and adolescence, (3) young to middle adulthood, and (4) the elderly.

For example, among infants, respiratory distress, congenital malformations, sudden infant death syndrome, and low birth weight are the critical morbidities. For pregnant women, AIDS, toxemia (pregnancy-induced high blood pressure), ectopic pregnancies, and hemorrhage are the critical morbidities. Related issues include declining fertility, pregnancy among older women and adolescents, and the timing and spacing of pregnancies. Acute conditions such as head colds, influenza, and injuries are replaced by chronic conditions such as hypertension, heart disease, cancer, and diabetes as a person gets older.

CASE STUDY 4.1. Epidemiologic Investigation of Congestive Heart Failure

An epidemiologic investigation that began on January 1, 2020, identified a population of 1,000 people among whom four were found to have congestive heart failure on this date. During the year of the study, six additional new cases were found. Among the 10 cases, there were seven deaths during the year.

(continued)

(continued from previous page)

There were no deaths among the remaining 990 people. These findings are summarized in exhibit 4.4.

Case	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
1	■	■	■	■	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■	Died			
3	■	■	■	■	■	■	■	■	■	■	■	Died
4	■	■	Died									
5		■	■	■	■	■	■	■	■	■	■	■
6			■	■	■	Died						
7				■	■	■	■	■	■	■	■	Died
8					■	■	Died					
9						■	■	■	■	■	■	
10						■	■	■	■	■	■	Died

■ = disease
▦ = at risk of disease

Questions

1. What was the point prevalence on January 1, 2020?
2. What was the point prevalence on July 31, 2020?
3. What was the point prevalence on October 31, 2020?
4. What was the cumulative incidence during 2020?
5. What was the incidence density during 2020?
6. What was the mortality rate during 2020?
7. What was the case fatality rate during 2020?

Answer Guide

1. Cases 1, 2, 3, and 4 = 4/1,000 or 4 per 1,000
2. Cases 1, 2, 3, 4 (died), 5, 6 (died), 7, 8 (died), 9, 10 = 7/(1,000 − 3) = (7/997) × 1,000 = 7.02 per 1,000: note that 3 deaths need to be subtracted from the denominator because those people were not alive and present on this date.
3. 1, 2 (died), 3, 4 (died), 5, 6 (died), 7, 8 (died), 9, 10 = 6/(1,000 − 4) = (6/994) × 1,000 = 6.04 per 1,000 Note that 4 deaths need to be subtracted from the denominator because those people were not alive and present on this date.

(continued)

(continued from previous page)

4. There are 6 new cases in the denominator. The population at risk = $1{,}000 - 4 - (1/2) \times 7$, $[6/(996 - 3.5)] \times 1{,}000 = 6.05$ per 1,000. Note that in the denominator we subtracted from the 1,000 cases the 4 prevalent cases at the beginning of the year and one-half of the number of deaths that occurred during the year.

5. For incidence density (ID), the numerator represents 6 new cases. There were 990 people who did not get congestive heart failure or die during the year, so they were at risk for a total of $990 \times 1 = 990$ person-years. Cases 1–4 were not at risk for any time during the years because they already had the disease. Cases 6–10 were at risk for a total of 18 months, or 1.5 person-years. That is, ID = $[6/(990 + 1.5)] \times 1{,}000 = 6.05$ per 1,000 person-years.

6. For the mortality rate, there would be 7 deaths in the numerator and 1,000 people at risk of dying in the denominator = $7/1{,}000$ or 7 per 1,000.

7. Case fatality measures the percent of all cases that result in death within a period = $7/10$ or 70 percent.

Screening

Secondary prevention typically involves the use of screening tests to identify individuals with a particular disease in the hope of finding the disease at an early and treatable stage. *Screening* is the "identification of unrecognized disease or defect by the application of tests, examinations, or other procedures that can be applied rapidly and inexpensively to populations" (Valanis 1999).

screening test
Test performed on healthy individuals to identify disease before symptoms occur.

Screening and diagnostic tests differ in a number of ways. A **screening test** is typically done before, rather than after, symptoms occur. Screening tests do not need to be ordered by a physician and may be obtained in non-medical settings, such as health fairs. A **diagnostic test** must be ordered by a physician and typically requires expensive, specialized equipment; is more time consuming; and may incur pain, discomfort, or risk. Screening tests are usually simpler, quicker, and painless. Screening tests are applied to healthy populations to identify disease before symptoms occur so that it can be treated early. Diagnostic tests are applied to patients with symptoms to determine an accurate diagnosis.

diagnostic test
Test performed on patients with symptoms to identify disease.

validity
The accuracy with which a measure represents a particular phenomenon.

Two measures of the quality of a test are validity and reliability. **Validity** refers to the accuracy with which a measure, such as a screening test, represents a particular phenomenon. For example, suppose a five-question screening test

was developed for clinical depression; the test would have validity to the extent that the questions accurately characterized this mental disorder. As another example, the prostate-specific antigen (PSA) screening test for prostate cancer has validity if a measurement of PSA accurately portrays prostate cancer.

Reliability is a measure of consistency. Reliable screening tests yield the same results regardless of the number of times they are repeated. *Interrater reliability* measures the degree to which different reviewers get the same results with single or multiple applications of a test. *Intrarater reliability* refers to the consistency of test results found by the same reviewer. For example, one could assess the interrater reliability of a blood pressure screening test by determining the consistency of results with multiple tests by different nurses. The intrarater reliability of mammography could be assessed by measuring the consistency of test results when a single radiologist interprets a number of films on multiple occasions.

Ideally, the screening test should distinguish individuals who have the disease (the true positives) from those who do not (the true negatives). The test should minimize the number of individuals who do not have the disease but test positive (the false positives) and those who have the disease but test negative (the false negatives).

The logic of the screening test involves the choice of a particular level (e.g., blood pressure level, PSA level) as the critical juncture between the positive and false negative test result, in an effort to minimize false positives and negatives. Unfortunately, this represents a trade-off. If one wants to be sure that the test identifies all true positives, then one must accept a higher level of false positives. By the same token, to ensure that those who test positive for the disease do in fact have the disease, one must accept more false negatives. **Sensitivity** measures the proportion of those who have the disease and test positive, whereas **specificity** measures the proportion of those who do not have the disease and test negative. Thus, sensitivity is the ability of the test to identify those who are truly sick, whereas specificity is the ability of the test to correctly identify those who are well.

This relationship is illustrated mathematically in the 2×2 table of exhibit 4.5. The columns represent those with and without the disease, labeled "reality." The rows represent those who test positive and negative. Cell *a* represents the true positives, *b* represents false positives, *c* represents false negatives, and *d* represents true negatives.

Assume that 2,000 people are tested—1,000 with the disease and 1,000 without. Exhibit 4.6 is an example of what might be found when the test is done. The sensitivity is calculated by dividing *a* by $a + c$, and the specificity is calculated by dividing *d* by $b + d$. In our hypothetical case, the sensitivity is 80 percent and the specificity is 90 percent. This represents a less-than-ideal test, but one consistent with several commonly used laboratory tests. One can also

reliability
The consistency with which a measure yields the same results in multiple applications.

sensitivity
The ability of a test to correctly identify individuals who have the disease the test is meant to detect.

specificity
The ability of a test to correctly identify individuals who do not have the disease the test is meant to detect.

Reality

	Yes	No
Yes	(a) True positives	(b) False positives
No	(c) False negatives	(d) True negatives

Test (vertical label on left)

calculate a false positive ratio as *b* divided by *b* + *d* and the false negative ratio as *c* divided by *a* + *c*. With a sensitivity of 80 percent and a specificity of 90 percent, the false positive and false negative ratios are 0.10 (10 percent) and 0.20 (20 percent), respectively.

The number of false positives and false negatives will vary with the prevalence of the disease, even though the sensitivity and specificity of the test will remain the same. The hypothetical example just presented assumed a prevalence of 50 percent—that is, of the 2,000 people tested, half had the disease. Exhibit 4.6 also illustrates the case with the same specificity and sensitivity in a group of 2,000 individuals where the prevalence is 10 percent. With a 50 percent prevalence, 100 people are incorrectly labeled as positive. With a 10 percent prevalence, the number rises to 180 of the 2,000 population. As you can see,

	Disease (50%)		Disease (10%)		Disease (1%)	
	Yes	No	Yes	No	Yes	No
Yes	800	100	160	180	16	198
No	200	900	40	1,620	4	1,782
Totals	1,000	1,000	200	1,800	20	1,980

Test (vertical label on left)

Sensitivity = 80%	Sensitivity = 80%	Sensitivity = 80%
Specificity = 90%	Specificity = 90%	Specificity = 90%
PPV = 0.89	PPV = 0.47	PPV = 0.07
NPV = 0.82	NPV = 0.98	NPV = 0.998

the number of false positives increases as the prevalence rate decreases. If the prevalence rate were 1 percent, 198 people would be mislabeled as false positives.

The measure that examines the ability of a test to predict disease is called the *predictive value*. One can measure both the positive and the negative predictive value of a test. The **positive predictive value (PPV)** is the proportion of those who test positive and who actually have the disease. The **negative predictive value (NPV)** is the proportion of those who test negative and who actually do not have the disease. Returning to the 2×2 table of exhibit 4.5, the positive predictive value is a divided by $a + b$ and the negative predictive value is d divided by $c + d$. Note the PPV and NPV in exhibit 4.6 in the circumstance where the prevalence of the disease is 50 percent, 10 percent, and 1 percent. The PPV falls to less than 10 percent in the situation where the prevalence is 1 percent.

When more than one test is available for a particular disease, the clinician may choose one of two techniques to improve the testing characteristics: simultaneous testing or sequential testing. *Simultaneous testing* involves ordering more than one test at the same time. It is a "believe the positive approach" because a diagnosis is confirmed with a positive result on either or both tests. Conceptually, simultaneous testing involves estimating the total area of a Venn diagram where circles A and B represent the number of true positives identified by each test and the overlap represents the number of true positives identified by both tests (see exhibit C.6 in Capstone Case C). For example, if a clinical test (A) for asthma had a sensitivity of 80 percent, then the test should identify 80 out of 100 people who really had asthma (number of people in circle A). If a laboratory test (B) for asthma with a sensitivity of 85 percent was ordered simultaneously, this test would identify 85 of those 100 people (number of people in circle B). The number of people in the overlap would be calculated as the sensitivity of one test multiplied by the number of people identified by the other test—in this instance, $0.85 \times 80 = 68$ or $0.80 \times 85 = 68$. The number of people who only tested positive on one test would be $80 - 68 = 12$ for test A, and $85 - 68 = 17$ for test B. The total area of the Venn diagram would be $12 + 17 + 68 = 97$, so the sensitivity would be 97/100, or 97 percent. Thus by ordering two tests simultaneously, the net sensitivity improves significantly above the sensitivity of either test alone. The trade-off of this approach, however, would be a reduced specificity because one would have to test negative on both tests to be ruled out. Suppose the specificity of the tests were 0.8 for test A and 0.7 for test B. The net specificity would be $0.8 \times 0.7 = 56$ percent. Positive predictive value would also be lower.

Sequential testing typically involves a "believe the negative approach" whereby patients are ruled out with a negative test result on the first test (stage 1), and patients who test positive are moved on to a second, usually different test (stage 2). For example, Capstone Case C illustrates this approach for colorectal

positive predictive value
The chance that a person who tests positive for a disease actually has that disease.

negative predictive value
The chance that a person who tests negative for a disease actually does not have that disease.

cancer with a fecal occult blood test (FOBT) in the first stage followed up by a colonoscopy in the second stage for only those who test positive on the FOBT. Case study 4.2 illustrates this technique of sequencing two screening tests to improve both specificity and positive predictive value.

Case Study 4.2. Breast Cancer Screening

This case study evaluates the changes in test characteristics for breast cancer screening, with the first stage being magnetic resonance imaging (MRI) and the second stage being core needle biopsy (CNB). Assume sensitivity of 92 percent and specificity of 70 percent for MRI and sensitivity of 87 percent and specificity of 98 percent for CNB. Assume that the prevalence of breast cancer among women aged 20–40 is 0.2 percent, among all women aged 20 and older is 2 percent, and among women with symptoms is 10 percent.

Questions

1. How successful is the MRI in identifying women with breast cancer?
2. How successful is the MRI in ruling out women without breast cancer?
3. Of all those without breast cancer, what percentage will incorrectly test positive (false positives) with the MRI?
4. How would you calculate and interpret the PPV for a sample of 100,000 women aged 20 and older?
5. How would you calculate and interpret the PPV for a sample of women aged 20–40?
6. How would you calculate and interpret the PPV for a sample of 100,000 women with symptoms?
7. What is the relationship between prevalence of disease and PPV?
8. Suppose that we were to do sequential testing on 100,000 women aged 20 and older with the MRI first and a follow-up CNB only on those women who test positive. What would be the net sensitivity, specificity, and PPV of this two-test sequence?

Answer Guide

1. The MRI would identify 92 percent of women with breast cancer.
2. The MRI would rule out 70 percent of women without breast cancer.
3. 100% – 70% = 30%, i.e., 30 percent would incorrectly test positive for breast cancer.
4. From the information, one can derive the following 2 × 2 table (exhibit 4.7):

(continued)

(continued from previous page)

	Cancer	No Cancer	Total
MRI+	1,840	29,400	31,240
MRI-	160	68,600	68,760
	2,000	98,000	100,000

EXHIBIT 4.7 MRI: Women Aged 20 and Older, 2×2 Table

Step 1: If there are 100,000 women who are screening and the prevalence of breast cancer is 2 percent, then $0.02 \times 100,000 = 2,000$ actually have breast cancer.

Step 2: Of the 2,000 women who actually have cancer, 1,840 will test positive with the MRI (sensitivity $0.92 \times 2,000 = 1,840$). Therefore there are $2,000 - 1,840 = 160$ false negatives.

Step 3: If there are 2,000 women with breast cancer, then there are $100,000 - 2,000 = 98,000$ without breast cancer.

Step 4: With a specificity of 70 percent, there would be $98,000 \times 0.70 = 68,600$ true negatives and $98,000 - 68,600 = 29,400$ false positives.

Step 5: The total who test positive would be $1,840 + 29,400 = 31,240$.

Step 6: The total who test negative would be $160 + 68,600 = 68,760$.

Step 7: The PPV would be $1,840/31,240 = 5.9$ percent. This means that a woman who tests positive with the MRI would have a 5.9 percent chance that she actually has breast cancer.

5. Exhibit 4.8 summarizes the use of MRI screening on women aged 20–40. The PPV would be $(184/30,124) \times 100 = 0.61$ percent. This would mean that a woman aged 20–40 who screens positive with the MRI would have a 0.61 percent chance that she actually has breast cancer.

	Cancer	No Cancer	Total
MRI+	184	29,940	30,124
MRI-	16	69,860	69,876
Total	200	99,800	100,000

EXHIBIT 4.8 MRI: Women Aged 20–40, 2×2 Table

(continued)

(continued from previous page)

6. Exhibit 4.9 summarizes the use of MRI screening on women with symptoms. The positive predictive value would be (9,200/36,200) × 100 = 25.4 percent. This would mean that a women with symptoms who screens positive with the MRI would have a 25.4 percent chance that she actually has breast cancer.

EXHIBIT 4.9
MRI: Women with Symptoms, 2 × 2 Table

	Cancer	No Cancer	Total
MRI+	9,200	27,000	36,200
MRI-	800	63,000	63,800
Total	10,000	90,000	100,000

7. As prevalence of disease increases, the PPV of a test will increase.
8. Exhibit 4.10 summarizes the use of sequential screening, MRI first and then CNB. The first stage is the same as exhibit 4.7. Notice that in the second stage, the patients who are MRI+ become the totals in the second stage, and the sensitivity (87 percent) and specificity (98 percent) are applied to the patients in stage 2 (e.g., 1,840 × 0.87 = 1,601 true positives, and 1840 − 1601 = 239 false negatives). Net sensitivity would be 1,601/2,000 = 80 percent. The total patients who test negative would be 68,600 (ruled out in stage 1) + 28,812 (ruled out in stage 2) = 97,412. Net specificity would be (97,412/98,000) × 100 = 99.4 percent. Net PPV would be (1,601/2,189) × 100 = 73.1 percent.

EXHIBIT 4.10
MRI and CNB Sequential Testing

	Cancer	No Cancer	
MRI+	1,840	29,400	31,240
MRI-	160	68,600	68,760
Total	2,000	98,000	100,000

	Cancer	No Cancer	
CNB+	1,601	588	2,189
CNB-	239	28,812	29,051
Total	1,840	29,400	31,240

The difference between sensitivity and PPV, on the one hand, and specificity and NPV, on the other, is not entirely intuitive. Sensitivity and specificity are intrinsic characteristics of the test itself and of the ability of that test to make a distinction—on the basis of some measurable characteristic, such as blood sugar or blood pressure levels—between those who are have a particular disease and those who do not. Positive and negative predictive values are derived not only from characteristics of the test but also from the prevalence of the disease in the population. As shown earlier, it can be demonstrated empirically that PPV increases as disease prevalence does. The test itself has not changed, but one can place more trust in a positive test result from a screening test applied to a population wherein the disease is highly prevalent. On the other hand, NPV falls as disease prevalence rises. One can trust a negative test result more with rare diseases than with common diseases.

In assessing the usefulness of a screening test, ensuring that the test is both sensitive and specific is necessary; however, examining the predictive value of a test is also important, as discussed earlier. A classic example of this issue is HIV testing. When the causative agent for AIDS was identified and a test was developed to screen for HIV, a great deal of pressure came into play to test various groups, frequently not of high risk. The obvious downside to testing a low-prevalence population for HIV is that many individuals would falsely test positive, which would result in incorrectly labeling many people as having the virus. It would also result in substantial time and energy to evaluate the extent to which the positive test was correct for a specific individual and would create a great deal of unnecessary anxiety in those who tested positive but did not have the virus. The better strategy was to use the test only in high-risk populations, where the prevalence would be high enough to ensure a better PPV. The current HIV-1 enzyme immunoassay test has such high sensitivity and specificity (99.9 percent and 99.85 percent, respectively) that the PPV is high enough, even in a low-risk population, if a repeat test is administered. A good screening program is one that is relatively simple to implement and relatively inexpensive. It employs a test with high enough specificity and sensitivity to detect a disease of sufficient importance (in terms of prevalence or severity) at an early enough stage that prompt and available treatment significantly improves outcomes. The test is relatively safe and acceptable to patients.

For example, let us compare the mammography and colonoscopy screening for breast and colorectal cancer, respectively. The former is moderately simple and inexpensive, and relatively safe and acceptable to patients, though there is some discomfort. The test has moderately high sensitivity and specificity. It detects the most prevalent cancer among women, and early detection can improve survival. The colonoscopy is an invasive, relatively expensive, surgical procedure carrying a small, but significant, risk of complications. Patient acceptability is a major problem due to the one- or two-day unpleasant preparation

that is required. Yet early detection for this leading cause of cancer among both men and women can significantly improve outcomes.

Screening programs must also seek to address the racial, geographic, and socioeconomic disparities that exist among screening programs because of physical, financial, educational, or cultural barriers to care. Within a population health framework that seeks to address disparities in health outcomes, disparities in secondary prevention must likewise be considered.

Summary

Each disease has a characteristic natural history or "course over time" from onset to resolution. This natural history may involve multiple stages and specific junctures, such as the clinical horizon (when the disease can be detected) and the critical point (the point after which severe consequences occur). In this chapter we distinguished between screening and diagnostic tests; discussed sensitivity, specificity, and positive and negative predictive value; and outlined the characteristics of a good screening program. The case studies focused on epidemiologic investigation of congestive heart failure and on screening for one of the most common types of cancer among women, breast cancer.

The breast cancer case study illustrates the problems faced by most screening tests—they are imperfect. Screening tests can incorrectly report positive results for those without the disease. This type of error can lead to additional follow-up tests, costs, or complications. False negatives occur when patients with cancer are incorrectly given a negative test result, leading to delay in treatment, at the very least. The study also shows the relationship between prevalence and predictive value. In the breast cancer case study, for example, the positive predictive value of the MRI was much higher for symptomatic women (25.4 percent) than for women aged 20–40 (0.61 percent), as determined by the difference in prevalence of these two populations.

End-of-Chapter Case Exercises

1. Suppose that you are funded to engage in a study of chronic obstructive pulmonary disease (COPD) in a fixed population of 1,500 people. The study lasts for three years (January 1, 2018–December 31, 2020). At the beginning of the study (January 1, 2018), 11 people are identified as already having COPD. Six additional people develop COPD during 2018, and two of those six people with COPD die (these six people are at risk for a total of four person-years during 2018); five additional people develop COPD in 2019, and two of those people with COPD

die (these five people are also at risk for a total of nine person-years); seven additional people develop COPD in 2020, and two of those people with COPD die (these seven people are at risk for a total of 19 person-years during 2019). *Hint*: Incidence denominators exclude pre-existing cases of disease from prior years and deaths from prior years. Point prevalence denominators at the beginning of the year exclude deaths from previous years. Period prevalence denominators exclude any deaths from previous years. To simplify, assume that deaths during the "current" year are ignored in the denominator of incidence and period prevalence rates.

   a. What is the point prevalence rate (per 1,000) of COPD on January 1, 2018?
   b. What is the point prevalence rate (per 1,000) of COPD on January 1, 2019?
   c. What is the period prevalence rate (per 1,000) of COPD during 2019?
   d. What is the incidence rate (per 1,000) of COPD for 2018?
   e. What is the incidence rate (per 1,000) of COPD for 2020?
   f. What is the incidence density (per 1,000 person-years) for COPD for the entire three years?

2. The Pap test has been used for many years to screen for cervical cancer. Suppose we ran this test on 100,000 patients, and the rate of actual cervical cancer among those patients was 1 per 1,000. Of those with cervical cancer, 56 tested positive; of those without cervical cancer, 1,998 tested positive. Identify the term (word) for each of the following definitions:

   a. Proportion of diseased people correctly identified as diseased by the test?
   b. The probability of having the disease if you test positive?
   c. Number of disease-free patients who are identified as positive by the test?
   d. The probability of not having the disease if you test negative?
   e. The proportion of disease-free people who are correctly identified as disease-free by the test?
   f. The number of patients who actually have the disease and are identified as positive by the test?
      With regard to this specific test, calculate the following:
   g. Number of true positives
   h. Number of false negatives
   i. Sensitivity of this new test
   j. Specificity of this new test

    k. PPV

    l. NPV

    m. Probability of actually having the disease if you test negative

3. You want to do a sequential colorectal cancer screening in a population of 10,000 subjects. For the first stage, assume that the FOBT has a sensitivity of 50 percent and a specificity of 60 percent. For the second stage, assume that the colonoscopy test has a sensitivity of 95 percent and a specificity of 90 percent. Assume a prevalence of colorectal cancer in this population of 10 percent.

    a. How many true positives are in this population using only the FOBT?

    b. What is the PPV using only the FOBT?

    c. How many subjects will be given the colonoscopy test?

    d. What is the net sensitivity using both tests?

    e. What is the net specificity using both tests?

    f. What is the net PPV using both tests?

4. A new screening test for diabetes was evaluated on a group of 800 subjects, 20 percent of whom had diabetes, to determine the accuracy of the test. Of those with diabetes, 140 were identified by the procedure as having diabetes, and 500 of the people without diabetes were identified by the test as not having diabetes.

    a. What is the sensitivity of this new test?

    b. What is the specificity of this new test?

    c. What is the term for the probability of not having a disease if you test negative?

    d. If you test negative, what is the chance that you will actually have diabetes?

    e. Suppose that you gave the test to a population with a 10 percent prevalence of diabetes. What would happen to the PPV and NPV?

    f. You decide to do sequential testing with another test that has 90 percent sensitivity and 90 percent specificity. If you believe the negative, how many will be tested a second time with the other test? Assume 20 percent prevalence.

    g. What is the net sensitivity of the two-test sequence?

References

Celendano, D. D., and M. Szklo. 2018. *Gordis Epidemiology*. Philadelphia, PA: Elsevier.

Centers for Disease Control and Prevention (CDC). 2020a. *Morbidity and Mortality Weekly Report*. Reviewed April 1. www.cdc.gov/mmwr.

———. 2020b. "National Center for Health Statistics." Reviewed March 30. www.cdc. gov/nchs/.

Donabedian, A. 1973. *Aspects of Medical Care Administration*. Cambridge, MA: Harvard University Press.

Dunn, J. R., and M. V. Hayes. 1999. "Toward a Lexicon of Population Health." *Canadian Journal of Public Health* 90 (Suppl): S7–S10.

Evans, R., and G. C. Stoddart. 1990. "Producing Health, Consuming Health Care." *Social Science and Medicine* 31 (12): 1347–63.

Kindig D., and G. Stoddart. 2003. "What Is Population Health?" *American Journal of Public Health* 93 (3): 380–83.

Kindig, D. A., Y. Asada, and B. Booske. 2008. "A Population Health Framework for Setting National and State Health Goals." *Journal of the American Medical Association* 299 (17): 2081–83.

McDowell, I., R. A. Spasoff, and B. Kristjansson. 2004. "On the Classification of Population Health Measurements." *American Journal of Public Health* 94 (3): 388–93.

Merrill, R. M. 2016. *Introduction to Epidemiology*, 7th ed. Burlington, MA: Jones and Bartlett Learning.

Mosby. 2009. *Mosby's Medical Dictionary*, 8th ed. St. Louis, MO: Mosby Elsevier.

Porta, M. 2014. *A Dictionary of Epidemiology*, 6th ed. New York: Oxford University Press.

Valanis, B. 1999. *Epidemiology in Health Care*. Stamford, CT: Appleton & Lange.

Weiss, N. S., and T. D. Koepsell. 2014. *Epidemiologic Methods: Studying the Occurrence of Illness*. New York: Oxford University Press.