CHAPTER

2

Descriptive Methods

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

- Identify and describe descriptive epidemiologic study designs and methods
- Describe and provide examples of the characteristics of person, place, and time
- Develop case definitions and understand their importance
- Work with line listing of individual data
- Review basic measures of disease frequency, including incidence and prevalence, standardization, absolute counts, proportions, rates, ratios, and incidence and prevalence measures
- Describe and calculate specialized measures, including case fatality rates and attack rates

Background

This chapter covers one of the most important foundations of epidemiology and public health—descriptive methods. Descriptive epidemiology was one of the earliest methodologies in the field, and it continues to be a key way in which our methods are used. With increased use of rapid computing resources that make iterative modeling and statistical analysis of multivariable relationships available at the desktop, the past several decades have seen a decreased emphasis on the importance of descriptive epidemiology. However, many researchers are now revisiting and expanding descriptive techniques. Those epidemiologists who specialize in outbreak investigations never ceased using and improving on

these important methods; the rest of us have a newfound appreciation for the richness of descriptive methods in concert with newer techniques. Descriptive methods help us understand public health phenomena and are critical in understanding whatever field we are studying. For some of you, this will be a review of the descriptive methods that you learned in an introductory epidemiology course; for others this may represent a fresh look at the material. Whether you are already comfortable with descriptive techniques or this is a new topic for you, this chapter will convey the material with a conceptual approach to descriptive methods and provide a foundation for the future.

Let us begin with an example to provide some context for thinking about the importance of describing data. Consider your reaction to the following scenario: you are healthy and young, seldom sick. You might not have an extensive education in biology or medicine, but you have a basic layperson's knowledge of disease. You attend an orientation for your new job at a hospital, where you are hoping to gain experience in hospital information systems. During lunch, you eat with other attendees at the hospital cafeteria. Nutrition conscious, you opt for the salad bar and have tossed salad plus a small scoop of pasta salad. By the early afternoon session, you are vomiting and have moderate to severe diarrhea. Fortunately, you do not have to leave early, but by the time the orientation comes to a close, you are quite exhausted. You nurse your sickness at home for 36 hours before feeling fully recovered from the gastrointestinal upset and associated dehydration.

Based on your sample size of only one, you make the following observations:

- You were healthy upon arriving at the orientation.
- You felt sick within 2 hours of eating lunch.
- Others who ate at the salad bar may have also become ill and just like you, braved the afternoon sessions; you do not know.

What do these observations tell you? They could mean a few things:

- There is something of a noninfectious etiology occurring, but it would be new and something you are unaware of.
- You were coming down with an infectious disease before you got to the orientation but remained asymptomatic until the afternoon.
- Something you ate at lunch made you ill.
- Some other exposure during the day made you ill.

If you opt to go with the idea that something you ate made you ill, what information can assist you in evaluating this question? You decide to consult some books at the library on communicable diseases, and you find out that staphylococcal food intoxication produces similar symptoms to yours and is caused by *Staphylcoccus aureus*, which can grow and produce an enterotoxin in meats, egg products, macaroni and potato salads, and cream-filled pastries. The toxin has a brief incubation period, with as little as 30 minutes to 8 hours passing between consumption and the development of the symptoms of the type you experienced.

Then you look at the local public health department website to see its listings of public health inspections of restaurants and institutions. You see that your new employer has had five reported incidents of foodborne outbreaks in the past 2 months in the cafeteria, including two of staphylococcal food intoxication. You are increasingly suspicious that something at the salad bar caused your illness, though you are still uncertain which element made you ill. Was it the salad? The pasta? The dressing? The drink? Something that touched something else, such as a dish with residue from another meal on it?

Your observations parallel the basics of descriptive epidemiology:

- Characterization of person, place, and time: You described yourself (person), where you were first ill (place) and what you were doing immediately preceding the symptoms (the environment), and when it occurred relative to other events as well as in absolute terms (the time). These are the three most salient features of an outbreak investigation or of any descriptive epidemiologic study: person, place, and time.
- Dependent variable (the outcome): gastrointestinal upset characterized by vomiting and diarrhea.
- Potential independent variable (the putative agent): staphylococcal food intoxication from the pasta salad.
- Establishment of a working case definition: You put together the person, place, and time characteristics that you identified to create a description of what you experienced: moderate to severe vomiting and diarrhea relatively soon after exposure to the putative agent.
- Time from putative exposure to symptoms: Two hours passed before you got sick after eating lunch.
- Assessment of potential causes: You performed a miniature literature review based on your symptoms and timing to investigate organisms that could have been associated with the illness.
- Suggested hypotheses: You examined publicly available public health data to see whether food poisoning is a reasonable explanation given the hospital's history. In many descriptive studies it is not possible to gather specimens. As in this case, a person does not usually know she or he is going to get sick and cannot take samples because of situational constraints. Here, biological specimens were not obtained because you self-treated and did not self-refer for care or diagnosis. Still, even without the specimens, you identified a hypothesis. It is important to remember though that simply a history of many outbreaks in a location does not prove that this is the situation at hand, and likewise, no prior outbreaks does not mean there was not one.
- Intervention: No intervention took place, but your findings suggest that hospital staff may be able to improve food management techniques and ultimately reduce the risk of foodborne outbreaks in the future. One step that has not yet occurred in this scenario is that you would need to ensure that your experience is communicated to your local health department so that it can assess the situation in the future and if there is a problem, prevent further illnesses. At the very least, the cafeteria staff can be reminded of hygienic food preparation techniques.

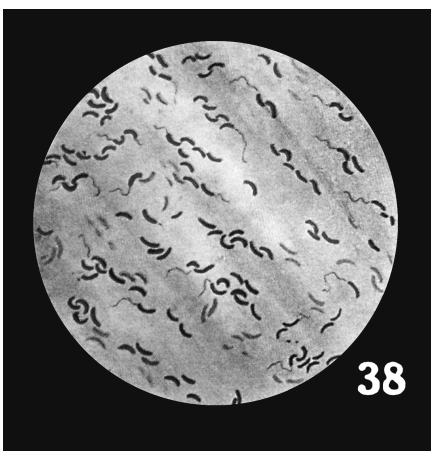
The following week, when you begin work, another new employee who also was at the orientation sheepishly inquires whether you got sick with vomiting and diarrhea at orientation following lunch. You discuss what each of you ate and find that your friend had a turkey sandwich with mustard and a side of pasta salad. The only common food between you was the pasta salad. You feel your mystery may be solved. Though your sample size is still only two and you lack substantial or biological evidence, you and your new friend feel fairly confident that you both had food poisoning and that it might have been from the hospital cafeteria's salad bar, specifically the pasta salad. It's important to remember, though, that because of confounding as well as myriad other design challenges, not the least of which is absence of biological confirmation, you cannot be sure this is what made you sick. As you develop more skills, you will notice how many methodological challenges enter into a full understanding of this problem. Please remember this is only an example for the sake of instruction. For now, let us turn our focus to the information gained in this example and how it was synthesized. The method, no matter how simple, is instructive in understanding descriptive epidemiology as a core concept.

Some of the most exciting and important discoveries have been initiated through descriptive epidemiology. John Snow used mapping to inductively identify the source of cholera in London during two outbreaks during the mid-1800s. Through mapping and thorough description of the people who became ill, and the exposures of those who did and did not become ill, Snow collected sufficient evidence to convince city officials to enact a public health intervention—the iconic removal of the Broad Street pump (**Figure 2-1, Figure 2-2, Table 2-1**).

John Snow Mapping Cholera

John Snow used a quintessential epidemiologist skill: mapping.

A serious outbreak of cholera took place in 1854 in London, England, following on the heels of an equally dangerous one in 1849. Cholera is an acute bacterial enteric disease caused by *Vibrio cholerae* (see **Figure 2-1**), primarily serogroups O1 and O139. Cholera has a case fatality rate of 50% in the absence of treatment (meaning that approximately half of those infected will die if not treated). Death is primarily due to dehydration caused by severe diarrhea and vomiting. Cholera is one of three diseases requiring notification under the International Health Regulations, and is strongly linked to living conditions and access to clean water. Although epidemics and pandemics of cholera occur even today, often related to emergencies, disasters, and other impediments to accessing clean water, cholera



Courtesy of the Centers for Disease Control and Prevention.

Figure 2-1 *Vibrio cholerae* bacteria.

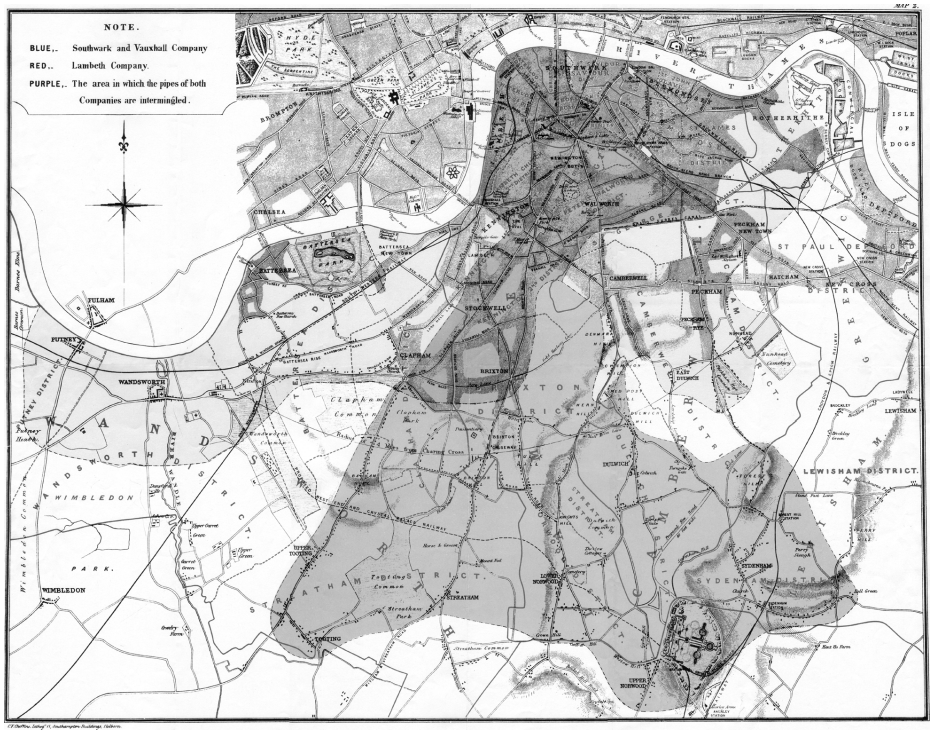


Figure 2-2 Snow's London.

is preventable and treatable. John Snow used many of the descriptive and outbreak epidemiologic skills you will learn in this book to identify the cause of the outbreak of cholera in London. In the 1849 outbreak, there were more than 500 deaths due to cholera in a matter of 10 days. At that time, water from the two water companies, Lambeth and the Southwark and Vauxhall, used water from a polluted part of the Thames river (see **Figure 2-2**). The sewage pollution there was thought to be the cause of disease. After 1849, the Lambeth Company began using water from a less densely polluted part of the river. During the second outbreak in 1854, Snow identified that people supplied by the Lambeth Company were now much less likely to contract and die from cholera. Snow used maps to assess the water sources supplying those individuals with cholera and compared them to neighborhoods where there was less disease. Taken together, he was able to identify the source. Without the sophisticated testing that we have today, he went to the Broad Street pump and removed the handle—an iconic public health action. This done, water from the contaminated source was no longer available. Quickly, the outbreak subsided. In addition to saving lives, this action supported Snow's hypothesis: if he had been mistaken about the cause, this intervention would have been unlikely to be effective.

The Broad Street pump is iconic in epidemiology in general and infectious disease epidemiology in particular. Snow used tools that we have available to us and that are the cornerstones of infectious disease epidemiology. He identified a potential source (water from the Southwark and Vauxhall Company), counted cases in relation to the outcome (contraction of cholera), mapped it (now we have geographical information systems [GIS] to help us in this endeavor), identified a hypothesis, tested the hypothesis, took public health action, and documented his findings. The data may be found in **Table 2-1**, so you may see them yourself.

(*continues*)

Table 2-1 Cholera Data Analyzed by John Snow

Proportion of deaths to 10,000 houses, during the first 7 weeks of the epidemic, in the population supplied by the Southwark and Vauxhall Company, in that supplied by the Lambeth Company, and in the rest of London.

Water source	Number of houses	Deaths from cholera	Deaths in each 10,000 houses
Southwark & Vauxhall Company	40,046	1,263	315
Lambeth Company	26,107	98	37
Rest of London	256,423	1,422	55*
Where do these data come from?	*Service record*	*Death records*	*(Deaths/houses served) × 10,000*

* This number was originally published as 59 in Snow's table in *On the Mode of Communication of Cholera*. It may have been the result of a typo, changes in underlying data from the census department, or the actual number of houses in each area.
Also noted in article by Carvalho FM, Lima F, and Kriebel D. Re: on John Snow's unquestioned long division. *Am J Epidemiol* 2004;159:422. See Literature Cited for additional resources regarding John Snow and cholera.

Reproduced from Snow, J., *On the Mode of Communication of Cholera*, Second Edition, much enlarged, 1855, Table IX.

Many diseases have been identified through the work of skilled and observant clinicians noticing anomalies in terms of person, place, or time and using descriptions of observations of evidence. For example, in 1981, a rare type of pneumonia was identified among five young, healthy, active homosexual males in Los Angeles. Even this small number of patients was sufficient to point toward a new disease, the disease that later came to be known as acquired immunodeficiency syndrome, or AIDS. Keen observations and their juxtaposition with "the usual case" allowed the gravity of these symptoms and diagnoses to be recognized; it would have been easy for these cases to go undetected by the clinicians. The key observation was that *Pneumocystis carinii* pneumonia was common among severely immunocompromised individuals, such as the elderly, or those with health conditions that harmed their immune systems, but rare among young, healthy individuals. **Figure 2-3** through **Figure 2-5** provide additional information.

There are many other examples of clinicians, epidemiologists, lab technicians, and patients themselves noticing conditions and rare diseases. In each case, observant people were able to identify a public health threat by examining the relation between outcomes (the dependent variable or illness) and potential causes (the independent variables, risk factors, or causative agents) by way of meaningful comparison between people with and without the disease. Descriptive epidemiology provides a method for systematically examining data points by the use of specific methodologies. The formal study methods that use descriptive epidemiology include case studies, case reports, ecological studies, and outbreak investigations to formalize and document observations.

Descriptive Epidemiology and Early HIV/AIDS

Noticing aberrations in population health and potential differences in presentation of disease are often the first steps towards stopping disease; recognition is critical. Case reports are often the product of one or two astute individuals noting that something is different, unusual, wrong, or just plain out of sorts. Almost always, what is amiss is in the person, place, or time characteristics of the events: a common disease in the wrong type of person (older, younger, sicker, healthier than usual); a disease common in the West but rare in the East appears in a new place; a disease that usually occurs in the winter is found in the summer. When someone notices this, it may be because they see patients in a clinic, like a healthcare provider, or perhaps they work at the emergency department and notice an influx of a certain type of patient who is not like the norm. Or a public health worker at the department of public health may notice increased surveillance reports that look too similar when submitted via passive reporting, or calls to a disease-specific help desk that raise suspicion of an emerging, re-emerging infectious disease, or an outbreak situation. There are many "clues" that can help us identify potential health scares.

In the summer of 1981, the world changed with the notice of five cases of *Pneumocystis carinii* pneumonia (PCP) in June. These were disclosed in the *Morbidity and Mortality Weekly Report* (MMWR) as the case reports shown in **Figure 2-3** and the images in **Figures 2-4** and **2-5**. This pneumonia is cased by a fairly ubiquitous parasite and is seen most commonly among immunocompromised and elderly

Pneumocystis Pneumonia—Los Angeles

In the period October 1980–May 1981, 5 young men, all active homosexuals, were treated for biopsy-confirmed *Pneumocystis carinii* pneumonia at 3 different hospitals in Los angeles, California. Two of the patients died. All 5 patients had laboratory-confrimed previous or current cytomegalovirus (CMV) infection and candidal mucosal infection. Case reports of these patients follow.

Patient 1: A previously healthy 33-year old man developed *P. carinii* pneumonia and oral mucosal candiasis in March 1981 after a 2-month history of fever associated with elevated liver enzymes, leukopenia, and CMV viruria. The serum complement-fixation CMV titer in October 1980 was 256; in May 1981 it was 32.* The patient's condition deteriorated despite courses of treatment with trimehoprim-sulfamethoxazole (TMP/SMX), pentamidine, and acyclovir. He died May 3, and postmortem examination showed residual *P. carinii* and CMV peneumonia, but no evidence of neoplasia.

Patient 2: A previously healthy 30-year-old man developed *P. carinii* pneumonia in April 1981 after a 5-month history of fever each day and of elevated liver-function tests, CMV viruria, and documented seroconversion to CMV, i.e., an acute-phase titer of 16 and a convalescent-phase titer of 28* in anticomplement immunofluorescence tests. Other features of his illness included leukopenia and mucosal candidiasis. His pneumonia responded to a course of intrravenous TMP/SMX, but, as of the latest reports, he continues to have a fever each day.

Patient 3: A 30-year-old man was well until January 1981 when he developed esophageal and oral candidiasis that responded to Amphotericin B treatment. He was hospitalized in Febuary 1981 for *P. carinii* pneumonia that responded to oral TMP/SMX. His esophageal candidiasis recurred after the pneumonia was diagnosed, and he was again given Amphotericin B. The CMV complement-fixation titer in March 1981 was 8. Material from an esophageal biopsy was positive for CMV.

Patient 4: A 29-year-old man developed *P. carinii* pneumonia in February in 1981. He had had Hodgkin's disease 3 years earlier, but had been sucessfully treated with radiation therapy alone. He did not improve after being given intravenous TMP/SMX and cortico-steroid and died in March. Postmortem examination showed no evidence of Hodgkins disease but *P. carinii* and CMV were found in lung tissue.

Patient 5: A previously healthy 36-year-old man with a clinically diagnosed CMV infection in September 1980 was seen in April 1981 because of a 4-month history of fever, dyspnea, and cough. On admission, he was found to have *P. carinii* pneumonia, oral candidiasis, and CMV retinitis. A complement-fixation CMV titer in April 1981 was 128. The patient has been treated with 2 short courses of TMP/SMX that have been limited because of a sulfa-induced neutropenia. He is being treated for candidiasis with topical nystatin.

The diagnosis of *Pneumocystis* pneumonia was confirmed for all 5 patients antemortem by closed or open lung biopsy. The patients did not know each other and had no known common contacts or knowledge of sexual

Figure 2-3 First case reports of *Pneumocystis carinii* pneumonia.

(*continues*)

partners who had similar illnesses. The 5 did not have comparable histories of sexually transmitted disease. Four had serologic evidence of past hepatitis B infection but had no evidence of current hepatisis B surface antigen. Two of the 5 reported having frequent homosexual contacts with various partners. All 5 reported using inhalant drugs, and 1 reported parenteral drug abuse. Three patients had profoundly depressed in vitro proliferative responses to mitogens and antigens. Lymphocyte studies were not performed on the other 2 patients.

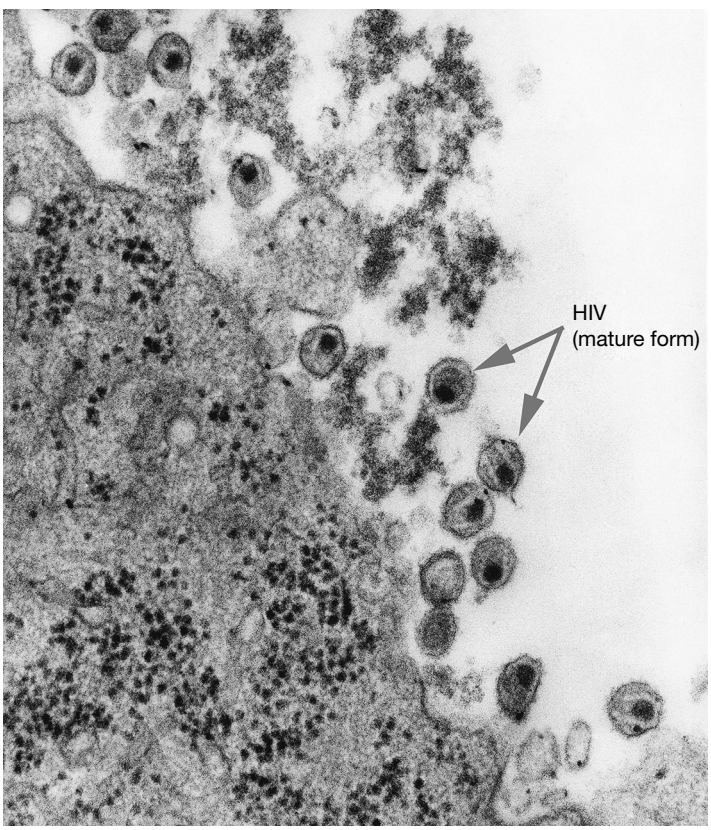
Reported by MS Gottlieb, MD, HM Schanker, MD, PT Fan, MD, A Saxon, MD, JD Weisman, DO, Div of Clinical Immunology-Allergy, Dept of Medicine, UCLA School of Medicine; I Pozalski, MD, Cedars-Mt. Sinai Hospital, Los Angeles; Field Services Div, Epidemiology Program Office, CDC.

This is the actual report describing the first PCP cases that ultimately informed our recognition of the HIV/AIDS epidemic. This was in the June 5, 1981 issue of the *Morbidity and Mortality Weekly Report*.

Reproduced from CDC, Pneumocystis pneumonia—Los Angeles. *Morbidity and Mortality Weekly Report*, 1981. 30: p. 250–2.

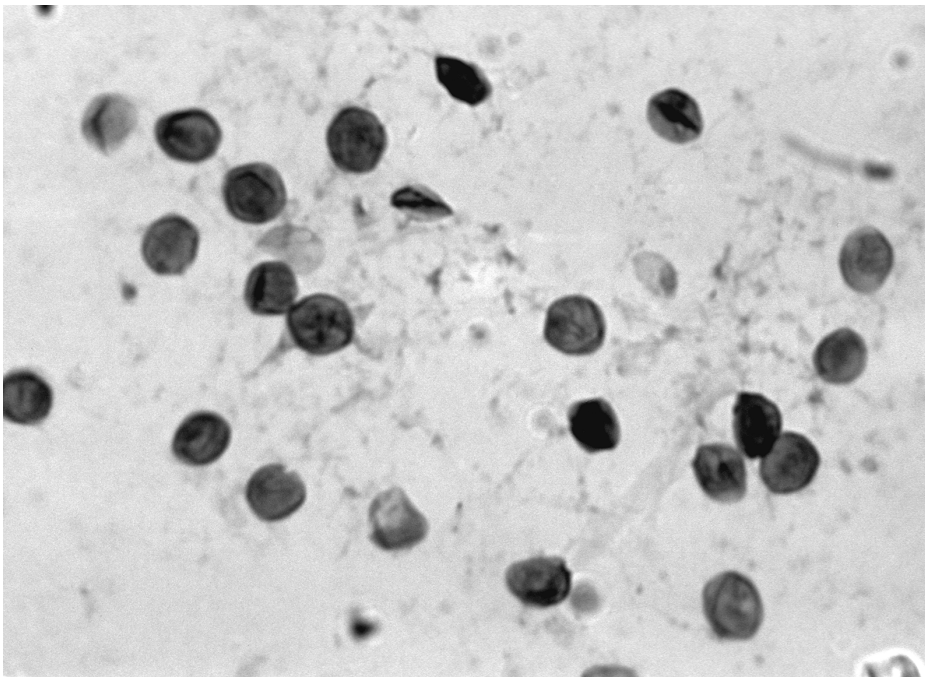
Figure 2-3 First case reports of *Pneumocystis carinii* pneumonia—the ushering in of an era (*continued*).

persons. For PCP to be in young, health individuals was very rare. These five cases ushered in the era of HIV/AIDS (**Figure 2-4**). Five young men, all "actively homosexual," were identified as having PCP (**Figure 2-5**)—a small but significant cluster of a rare disease among healthy persons of this age cohort.



Courtesy of the Centers for Disease Control and Prevention.

Figure 2-4 HIV.



Courtesy of the Centers for Disease Control and Prevention.

Figure 2-5 *Pneumocystis carinii* pneumonia.

Just a month later, in July, the MMWR reported 26 cases of Kaposi's sarcoma (KS), a rare cancer, among young, healthy, homosexual males. This is another case where the disease did not fit the usual characteristics, this time for person and place. The usual KS patients are elderly and immunocompromised. As with PCP, young, healthy cases in the United States were very rare. In addition, these cases had a very high case fatality rate (20%); usually cases are more chronic and do not result as often in death.

And these cases were clustered unusually on two coasts of the United States, in New York and California.

In the second report, more information was provided, making its approach a case series study design. This report provides some denominator data, though limited: a historical comparator was used. "A review of the New York University Coordinated Cancer Registry for KS in men under age 50 revealed no cases from 1970–1979 at Bellevue Hospital and 3 cases in this age group at the New York City Hospital from 1961–1979." This information highlighted that the 26 cases of KS occurring in such a short period of time and in such a small geographic area were, indeed, unusual. By August 1981, a report, indicating 108 people with one or both conditions (i.e., KS and/or PCP) and their basic demographics where available, was released.

A follow-up study was performed of cases of KS and/or PCP between June 1, 1981, and April 12, 1982. In this study, detailed demographic, clinical, and behavioral data were collected from each patient or their proxies (for those who had died); this study found the data consistent with the possibility of a new sexually transmitted infectious organism that leads to acquired immune deficiency (as happened to be the case), but also the possibility that it could be another factor, such as drugs, commonly used by homosexual men at the time (i.e., amyl nitrate or "poppers"). These steps follow directly from those in all descriptive epidemiologic studies, where the cases were evaluated, described, and then the description suggests hypotheses and appropriate next steps to the researchers.

(*continues*)

HIV/AIDS has since become a pandemic, with its worst human tolls now among heterosexuals in most parts of the world also among injection drug users and men who have sex with men. The fastest rising incidence rates in the United States are among African American and Latino men who have sex with men, particularly young men.

The intensive detective work to solve the mystery of AIDS epidemiologically, as well as in the laboratory, has been well chronicled in a variety of books, articles, films, and other documents.

As we watch the future unfold, we hope that we will have new and improved methods of control to add to those that we already have (such as highly active antiretroviral treatment, perinatal prophylaxis, circumcision, and others)—including the long-hoped for promises of being able, one day, to prevent HIV transmission with vaccines. Describing the natural history of disease as well as factors associated with both HIV acquisition and transmission and access and adherence to treatment remains critical to this day in halting the epidemic.

Describing Data

The first step of any epidemiological task is to carefully consider the data. The overarching goal of descriptive analysis is to become well acquainted with each data point, alone and in relation to other data points. Becoming familiar with the data allows a profile of the relationship between exposure and outcome variables to emerge, with analytic study methods following to quantify these relationships. The intent of descriptive epidemiology is to describe a health outcome exposures, or risk factors, characterizing the distribution of person, place, and time data; generating hypotheses; and identifying potential confounders and effect modifiers. These insights are necessary to direct more-advanced analyses, such as testing hypotheses, measuring the strength of associations after adjusting for confounders and effect modifiers, and stating causality. Data from descriptive studies can help determine the most effective study designs to detect associations.

One important quality of epidemiologic methods is that we can use them to describe anything, not just diseases. We can describe any health outcome of interest, including known, emerging, reemerging, or previously unknown disease entities. We can characterize health behaviors; healthcare utilization; and other public health phenomena, such as exposures to environmental factors, poverty access, and more. Descriptive epidemiology can include describing what characteristics are found in affected individuals, including looking for hypotheses to later test regarding potential causal agents. Descriptions of healthcare utilization may lead to recognition that specific subpopulations are not receiving or accessing care, healthcare disparities are occurring, or negative health behaviors are increasing (e.g., increased reports of unsafe sex, failure to vaccinate, etc.). Information ascertained about the people who are affected by a condition of interest, including where and when and for how long they have been affected, then may be used for healthcare facility planning and resource allocation. For example, in the event of an outbreak of disease, estimates of how many individuals are infected can assist in planning the proper number of personnel required to respond, number of hospitalizations expected, central locations of highest need, and expected duration of the epidemic. In the early moments of an outbreak or any health-related acute situation, these estimates are crucial.

Person, Place, and Time

Descriptive epidemiology is straightforward: our goal is to comprehensively characterize the people, places, and times of the events under study by quantifying their attributes.

Person

Person refers to the characteristics of the individuals affected by the outcome of interest. Who are they? Are they mostly male? A specific age group? A certain race/ethnicity? Religion? Are certain types of people specifically not affected by the outcomes? What unites the affected individuals? Which characteristics do they have or not have in common? We want to describe the type of person who has the condition under study in as much detail as possible. Standard demographic, clinical, and behavioral person characteristics include:

- Age, gender (at birth and self-identified), and marital status, past and current other demographics.
- Living arrangements, including homelessness status and number of children/adults in the home.
- Religion, spirituality, and cultural norms.
- Behavior: This is particularly important to the study of specific risk factors, such as foods consumed, sexual behavior, drug use, or healthcare-seeking behavior.
- Access to healthcare services and health status, including comorbidities and characteristics of the disease of concern.
- Socioeconomic status, including educational level, access to care and healthcare insurance, and employment status.
- Race/ethnicity: It is important to recognize that these are often proxy variables. Race and ethnicity are broadly defined, and the outcomes with which they are associated are frequently more strongly associated with other factors, such as behavior, economic status, and location of residence, rather than being meaningful in and of themselves.

Place

Where did the events take place? Where were the events in relation to each other? Were there other events proximal to each other, or were they in isolation? In a time rich with accessible travel, it is important to identify the place where the exposure occurred, because it may not be where the outcome was identified. For example, someone traveling might be feeling well upon leaving Africa but be direly ill upon landing in Sweden. The necessary characteristics of place would be those in Africa, where the disease was acquired, not in Sweden, where it was diagnosed, unless the disease was contracted on the plane. Rapid air, train, and car travel can make determining the place of an exposure difficult, but it is essential for descriptive epidemiology. For chronic diseases or conditions, place may be important in identifying risk factors for disease. For example, certain immune disorders, such as multiple sclerosis, are more common among people who live in colder, more northern environments; this relationship may provide insight about etiology or future intervention possibilities. Standard place characteristics include:

- Physical location where the exposure occurred, with information on country, state, city, zip code, block, etc.

- Type of location, including the type of housing, such as house or apartment; whether the exposure occurred at a school, job, restaurant, or other venue; rural, urban, or suburban neighborhood; proximity to factories or toxic waste; and presence of running water and sewage disposal.
- Surrounding characteristics of the environment, such as a desert, forest, or humid climate; industrial or rural setting; elevation; level of smog, pollen, or toxins in the air; and known infectious agents or vectors in the region.

Time

When did the events occur? When did symptoms first appear? When was the first diagnosis made? The timing of exposures and outcomes can tell us the type and source of an epidemic. Standard time characteristics of exposures and outcomes include:

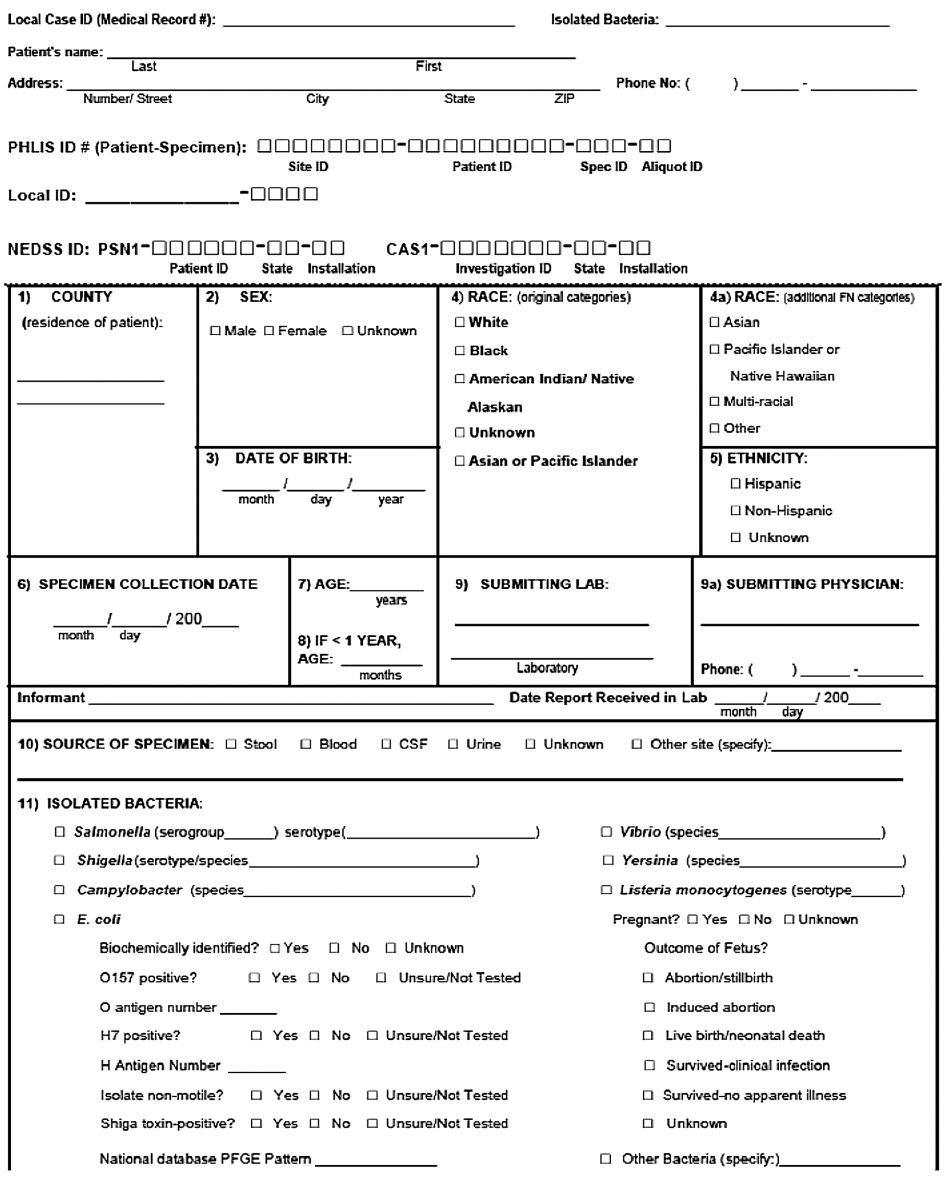
- Date (month, day, year) of event, including day of week.
- Clock time of event, first appearance of signs and symptoms, and diagnosis.
- Relationship to sunshine or to darkness.
- Relationship in time between outcome and other events, such as sewage release, large social gatherings, or natural disasters.
- Relationship to cycles, calendars such as flu season or agricultural season.
- Relationship of each event to other events in time and space. Geospatial clustering is when several cases occur in one area or have one geographic characteristic in common. For example, cases might be spaced far apart but all occur along the same interstate freeway. Temporal clustering is when several cases occur in a relatively brief period of time. How closely grouped in time and space the cases must be to be related is dependent upon the disease being investigated.

To facilitate descriptive analysis, data on person, place, and time characteristics are generally collected and documented in a systematic fashion. Following an outbreak at an event, for example, trained interviewers from the local office of public health may contact all the guests by phone and discuss them about what they ate, hoping that the interviewees are able to recall all the food items possible and any possible symptoms they may have had and when they had them. Interviewers may use a systematic inventory of all the foods present at the event, with the same questions administered to each and every guest (see **Figure 2-6a and b**). Questions in the inventory could cover specific foods and drinks consumed, list amounts in a standard fashion, identify specific combinations of food, and include other foods eaten in the suspect time period somewhere other than the catered event. A structured symptom inventory can then be obtained, eliciting specific symptoms, severity, timing of the first onset, treatments (e.g., over-the-counter medications for diarrhea, hospitalization for dehydration), and resolution.

Instruments for collecting descriptive data are developed and pretested for validity and reliability to the extent possible, given that in a public health emergency, speed in response is sometimes more important than perfection. This is one benefit of using standardized and validated forms from central sources such as the Centers for Disease Control and Prevention (CDC). These data collection forms are the basis for collecting descriptive information, and the

information that is then turned into one (or more) line of data per person, called a line listing as shown in **Figure 2-7**. Data are obtained through a variety of means, including interviews, self-interviews, and computer-assisted self-interviews, from potentially exposed individuals.

The following box shows an example of how part of a line listing might analyze data collected on a form used in an outbreak investigation. The box also displays ways that line-listed data can then be analyzed to help us understand the data.



Local Case ID (Medical Record #): _____ Isolated Bacteria: _____

Patient's name: _____ (Last, First)

Address: _____ (Number/ Street, City, State, ZIP) Phone No: ( )_____ - _____

PHLIS ID # (Patient-Specimen): □□□□□□□□-□□□□□□□□□□-□□□-□□ (Site ID, Patient ID, Spec ID, Aliquot ID)

Local ID: _____-□□□□

NEDSS ID: PSN1-□□□□□□-□□-□□ (Patient ID, State, Installation) CAS1-□□□□□□□-□□-□□ (Investigation ID, State, Installation)

1) COUNTY (residence of patient):	2) SEX: ☐ Male ☐ Female ☐ Unknown	4) RACE: (original categories) ☐ White ☐ Black ☐ American Indian/ Native Alaskan ☐ Unknown ☐ Asian or Pacific Islander	4a) RACE: (additional FN categories) ☐ Asian ☐ Pacific Islander or Native Hawaiian ☐ Multi-racial ☐ Other
	3) DATE OF BIRTH: _____/_____/_____ (month, day, year)		5) ETHNICITY: ☐ Hispanic ☐ Non-Hispanic ☐ Unknown

6) SPECIMEN COLLECTION DATE: _____/_____/ 200____ (month, day)	7) AGE:_____ years; 8) IF < 1 YEAR, AGE: _____ months	9) SUBMITTING LAB: _____ _____ Laboratory	9a) SUBMITTING PHYSICIAN: _____ Phone: ( )_____ -_____

Informant _____ Date Report Received in Lab _____/_____/ 200____ (month, day)

10) SOURCE OF SPECIMEN: ☐ Stool ☐ Blood ☐ CSF ☐ Urine ☐ Unknown ☐ Other site (specify):_____

11) ISOLATED BACTERIA:

- ☐ *Salmonella* (serogroup_____) serotype(_____)
- ☐ *Shigella* (serotype/species_____)
- ☐ *Campylobacter* (species_____)
- ☐ *E. coli*
  - Biochemically identified? ☐ Yes ☐ No ☐ Unknown
  - O157 positive? ☐ Yes ☐ No ☐ Unsure/Not Tested
  - O antigen number _____
  - H7 positive? ☐ Yes ☐ No ☐ Unsure/Not Tested
  - H Antigen Number _____
  - Isolate non-motile? ☐ Yes ☐ No ☐ Unsure/Not Tested
  - Shiga toxin-positive? ☐ Yes ☐ No ☐ Unsure/Not Tested
  - National database PFGE Pattern _____
- ☐ *Vibrio* (species_____)
- ☐ *Yersinia* (species_____)
- ☐ *Listeria monocytogenes* (serotype_____)
  - Pregnant? ☐ Yes ☐ No ☐ Unknown
  - Outcome of Fetus?
    - ☐ Abortion/stillbirth
    - ☐ Induced abortion
    - ☐ Live birth/neonatal death
    - ☐ Survived-clinical infection
    - ☐ Survived-no apparent illness
    - ☐ Unknown
- ☐ Other Bacteria (specify:)_____

Courtesy of CDC/FoodNet; (CDC 2000; CDC 2005).

Figure 2-6a Foodborne Disease Active Surveillance Network (FoodNet) Case Report Bacterial Form.

Part III. General information

Did you attend a large gathering the week before your illness? (e.g., wedding reception, showers, church events, clubs, school events, athletic events, office parties or banquets, parties, festivals, fairs) Y N

If yes, what events?
Event 1: _____location:_____ When? ____ / ____ / ____
Event 2: _____location:_____ When? ____ / ____ / ____
Event 3: _____location:_____ When? ____ / ____ / ____
Event 4: _____location:_____ When? ____ / ____ / ____

Do you know anyone else in your neighborhood/school/office/business/health club/church/synagogue etc. with the same illness? Y N
If yes: Where?_____
How many people?_____ Name_____ Tel_____
Name_____ Tel_____
Name_____ Tel_____

Did you travel anywhere during the seven days before your illness? Y N
If yes, where? _____When? ____ / ____ / ____ to ____ / ____ / ____
If airline travel, what airline? _____
Outgoing flight no. _____ Return flight no._____
Foods eaten on plane going there: _____
return:_____ If you stayed at a resort please provide resort name:_____
If cruise ship, name of ship_____ Destinations _____

Have you had contact with children in a childcare setting during the seven days before illness? Y N
If yes, when: ____ / ____ / ____ Name of facility:
_____
Location_____Phone: _____
Are you aware of any other illness in the daycare? Y N DK

During the seven days before your illness, did you have any pets at home, have contact with household pets elsewhere, or visit a household with pets? (including reptiles) Y N
If yes, what type of pets? _____
If your own pets, where do you buy your pet foods? _____brand: _____

Did you live on a farm, visit a farm, or visit a petting zoo in the seven days before your illness? Y N

If yes: what kind of animal(s) did you have contact with?_____
When? ____ / ____ / ____ Where?
_____
From what sources of water did you drink during the seven days before your illness?
Municipal tap water Y N DK
Private well water Y N DK
Untreated surface water (river, pond, lake) Y N DK
Bottled water Y N DK
Other _____

Did you drink any untreated/raw water during the seven days before your illness? Y N
If yes, where? _____

Did you swim during the seven days before your illness? Y N

Part IV. Specific food questions

In the week before your illness, did you eat any dish containing store-purchased ground beef (that is, cooked at home)? I'm referring either to bulk ground beef or pre-made beef patties purchased in a store by you or a relative/house-mate? Y N DK

If yes: where purchased? _____
When?_____
What was the brand name?_____
What type of ground beef was it (extra lean, lean, % fat, etc.)?_____

In the week before your illness, did you consume meat originating from any place other than a grocery store or restaurant, such as from hunting, a butcher shop, custom butchery? Y N
Where: _____ What:
_____

In the week before your illness, did you make or eat any dish that involved breaking and mixing four or more eggs? Y N DK

If yes: Where did you buy the eggs?_____ When?
_____
What was the brand?_____

Have you done any baking that used a raw egg in the preparation? Y N
Did you taste any of the uncooked batter? Y N

Reproduced from Centers for Disease Control and Prevention, Foodborne Disease Outbreak Investigation and Surveillance Tools, National Hypothesis Generating Questionnaire. Available at: http://www.cdc.gov/foodsafety/outbreaks/surveillance-reporting/investigation-toolkit.html.

Figure 2-6b Excerpts from standard foodborne disease outbreak case questionnaire.

ID number	Case definition status Case (1) or Control 0)	Gender Male (0) or Female (1)	Age (continuous, years)	Consumed agent A No (0) or Yes (1)	Quantity (if yes) 1–2 (1) 3–4 (2) 5–6 (3) >6 (4)	Hospitalized No (0) or Yes (1)
415	1	0	22	1	3	1
416	0	0	23	0	1	0
417	0	1	24	1	1	0
418	1	0	45	1	1	0
420	1	1	40	0	–	0
421	0	1	41	0	–	0
422	0	1	32	1	3	1
423	0	1	22	0	–	0
424	0	0	18	0	–	0
425	0	0	15	0	–	0
426	1	0	24	1	1	0
427	1	1	11	1	4	1
428	1	1	35	1	2	1
429	0	0	66	1	3	1
…	…	…	…	…	…	…
444	0	1	67	0	–	0
445	0	0	45	0	–	0
446	0	1	13	0	–	0
447	0	1	11	1	1	0

Figure 2-7 Example of a line listing.

Describing Data

An example of how data on each person may be translated from the data collection form into analyzable data, in the line listing in Figure 2-7.

What do you do once you have collected line listing of data? In order to get to know your data:

1. *Calculate the frequencies of categorical variables.* This will inform you of how the sample is distributed among different categories of independent variables. Summary data of your outcomes are especially important because they reveal the proportion of missing data, which can impact your study enormously. Some examples of frequencies:

(*continues*)

Demographic and clinical characteristics of women diagnosed with sepsis postoperatively (N = 110)

	n	%
Gender		
Female	73	66.4
Male	37	33.6
Age (years)		
<18	15	13.6
18–35	28	25.5
36–45	56	50.9
>45	11	10
Past medical history		
No significant medical problems	8	7.3
Mild	59	53.6
Moderate	41	37.3
Severe	2	1.8
Past surgical history		
No abdominal surgeries	38	34.6
One prior abdominal surgery	39	35.4
Two or more prior abdominal surgeries	33	30.0
Body mass index (BMI)		
Underweight (<18.5 kg/m^2)	4	3.6
Normal (18.5 to 24.9 kg/m^2)	48	43.6
Overweight (25 to 29 kg/m^2)	34	30.9
Obese (>30 kg/m^2)	21	19.1
Unknown	3	2.7

2. *Calculate measures of central tendency (mean, median, mode) and dispersion (standard deviation or variance) for continuous variables.* How are variables distributed? Do they follow a normal distribution (that is, like a bell curve)? Or are they skewed left or right? Are the tails heavy or skinny? This can be assessed visually to some degree, and tested quantitatively as well. Some examples of measures of central tendency and dispersion:

Characteristics of participants with cryptosporidium (N = 136)

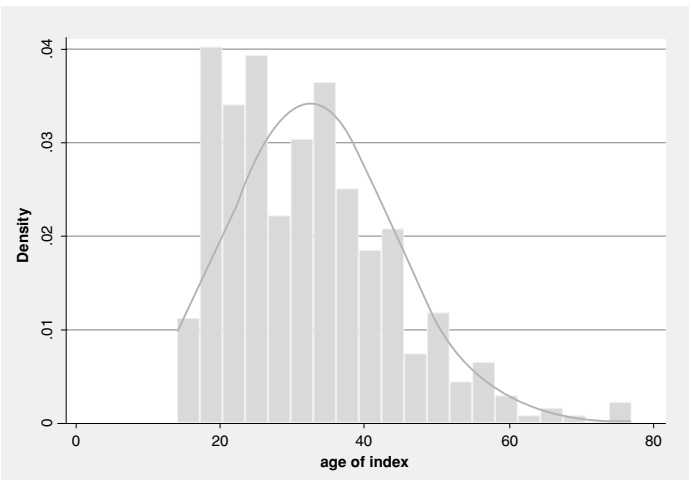
	Mean	Median	Mode	Standard deviation
Age (years)	25.4	24.0	23.0	5.79
BMI (kg/m^2)	23.7	23.5	23.4	21.25
Baseline CD4 (absolute count) at study entry	419.0	365.0	368.0	331.25

3. *Plot the continuous data one variable at a time, using box plots, stem-and-leaf plots, or other graphic displays at your disposal*. This describes the data variable by variable. In addition, it helps identify where there are out of range values or missing values, and gives a general description of your continuous data.

Stem-and-leaf plot for age (age of index)

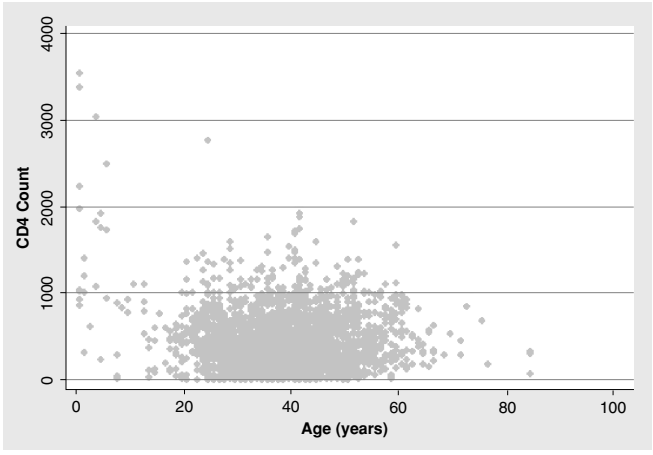
1f | 45555
1s | 6667777777
1. | 88888888888888888899999999999999
2* | 000000000000000000000000111111111111111
2t | 22222222222222222222233333333
2f | 44444444444444444444444555555555555555
2s | 66666666666666667777777777
2. | 88888888899999999999
3* | 00000000000011111111111
3t | 22222222222222222333333333
3f | 44444444444555555555555555555
3s | 66666666677777777777
3. | 88888888888888889999999
4* | 000001111111111111
4t | 222222233333333333333
4f | 44444444555555
4s | 666777
4. | 8888999
5* | 0000000011111
5t | 333
5f | 444
5s | 6666777
5. | 8899
6* | 11
6t | 3
6f |
6s | 66
6. | 9
7* |
7t |
7f | 55
7s | 7

A histogram describing the age of the index patients.



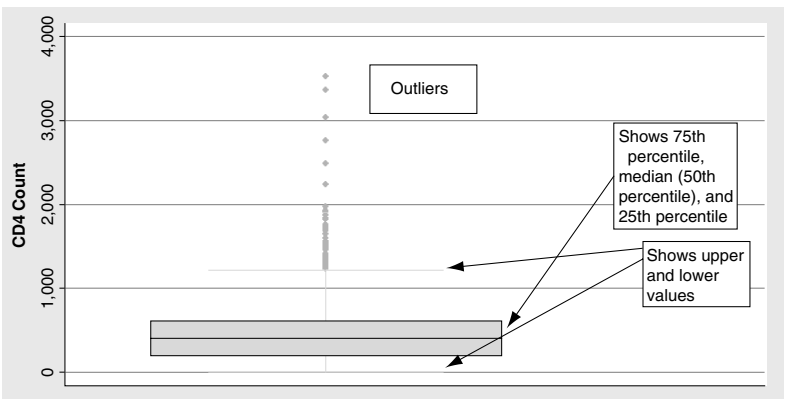
4. *Plot the data in a scatterplot, placing the dependent (outcome) variable on the Y-axis (the vertical axis) and the independent (potential predictor) variable on the X-axis (the horizontal axis)*. What do the data look like? How do they relate to each other? Is there any discernable pattern or relationship between the independent variables and dependent variable under study? Is there any discernable pattern between independent variables? (Remember that we are still looking for "clues"; not seeing a pattern does not mean there is not one, just as seeing one does not mean there is one!) Using a scatterplot showing the relationship between age and CD4 count, we can see here that it looks as if CD4 counts are higher the younger the participant is. This is not enough to prove anything or provide statistical testing on its own, but it does give one a feel for the relationship. If you plot this first, you will know what to expect when you analyze your data. For example, if you found the opposite from what this picture suggests (i.e., CD4s increase with age) you might want to do some quality checking.

Data in a Scatterplot



5. *Look at outliers, datapoints that stand out from the rest of the distribution.* For continuous variables, this can be quantified by looking at datapoints that extend beyond a set level (e.g., two standard deviation above or below the mean). What are they? Get to know each of these outliers. Investigate them. Are they data entry errors? Documentation problems? Or are they true? There are a number of techniques available to diagnose outliers and treat them appropriately. However, sometimes, if the data are correct, the outlier can be a substantive "clue" towards figuring out the problem at hand. Each might represent an acute case, a pronounced relationship, or something "different" that can be extremely useful.

Box and whisker plot of baseline CD4 count.



Descriptive Methods

Systematic data collection is necessary for almost every study design method, descriptive and analytic alike. Descriptive study designs include case reports, case series, ecological studies, and outbreak investigations. Together they comprise the core of descriptive epidemiologic methods and transform the data collected into interpretable information. Outbreak investigations are treated later in the text. Outbreak investigations are a specialized form of descriptive epidemiology and harness the power of describing person, place, and time.

Case Reports

Case reports are an essential link between clinical medicine and public health. Case reports often result after an astute clinician notices something odd in the presentation of a specific patient or in the appearance of a cluster of unusual events. What makes something odd or unusual, though? There are several hallmarks that might capture one's attention:

- Presentation of a known disease in an unusual population
- Identification of a previously unrecognized syndrome or disease
- Presentation of a disease that is more or less severe than previously seen or has a new characteristic, such as genetic resistance to a drug or failure to respond to the standard of care treatment
- Disease transmission by a mode not generally seen or suspected
- A cluster in time or space of diseased individuals that is unexpected in some way, such as may be seen with increased cancer risk from an environmental exposure

Case reports are often communicated within a facility but may also be disseminated to peer-reviewed journals, governing bodies overseeing clinical care (e.g., hospital review boards), or government entities (e.g., CDC, Ministries of Health). These case reports are then shared as needed with other practitioners and public health agencies. Similar cases may be sought or diagnostic procedures recommended in the event that providers see cases in the future. This helps launch public health into action should the need arise.

Case Series

Case series are similar to case reports except that they usually describe more than one case of the disease. In addition, case series frequently attempt to identify denominator data, though the method of obtaining these data is often relatively crude. For example, a provider who is seeing six cases of *Trichomonas vaginalis* (a sexually transmitted disease) that did not respond to metronidazole (the usual treatment) after adherent patients underwent several cycles of treatment may prepare in-depth case reports on the six patients. The provider then might expand the investigation to count how many cases that appeared to be metronidazole-resistant *T. vaginalis* occurred over a specified time period at that clinic. This number becomes the numerator and the number of people with *T. vaginalis* exposed to metronidazole becomes the denominator. Together these data might suggest a rate (though it is not actually one) and might give direction for future study. It is important to remember limitations for this type of study. Specifically,

the data are retrospective in nature; documentation was not likely to be systematic, as it was conducted for other purposes; and other patients may have had the same clinical presentation and not been identified or may have been included in the data set and not documented in the same way. Further, the clinic records may be less than optimal, including access to those records. It may not be possible to extract all patient records with regard to *T. vaginalis* and metronidazole treatment. For example, medical charts may be unavailable to the researcher because the patients are frequently sick (often in clinic), because they are never sick (may be archived), or for unknown reasons. Individuals on whom data can be collected may differ from those on whom data cannot be collected, and they may differ in ways that cannot be measured that are related to the exposure or outcome being investigated.

Clinic-based studies, such as case reports and case series, provide only an estimate of the individuals who accessed care at that particular clinic. There may be others who did not have access to the clinic, had no insurance, or were unable to seek care. Others may not have had signs or symptoms that encouraged them to seek care, or they may have been previously treated with metronidazole but did not return for follow-up care despite actual need. The observed rates at the clinic may not represent those of the underlying population. This is a nuance that is important to remember and will frequently serve as a limitation to your own inquiries. Still, estimation of metronidazole-resistant *T. vaginalis* rates at the clinic, no matter how crude, can help assess whether resistance is a rare event and whether it is actually increasing or only appears to be. This type of study is valuable in that it can give us clues to future studies we need to do.

Ecological Studies

Ecological studies differ from other types of descriptive epidemiology in that individuals are not the unit of analysis. In this type of study, we analyze data at the *group* level. These studies are important for several reasons:

- They frequently generate important hypotheses for further analytic research.
- They allow analysis comparing large groups of people, such as inhabitants of different countries, that would otherwise be impossible.
- They can be done without the benefit of substantial resources; publicly available information is often sufficient to conduct an ecological study.

Geographic comparisons are common with ecological studies, but they are not the only possible approach. This underscores the importance of mapping (using GIS or other software) in all descriptive studies. Just like John Snow did, mapping data so we can see things visually is a very powerful tool! It is also used in ecological studies. Other ecological comparisons include classes, schools, genders, races, and other grouping variables. Descriptive data about the outcome or exposure are collected and then linked with additional descriptive data on the other variables of interest.

Here is an example of how ecological studies work. Imagine we have statistics on the number of cartons of ice cream sold by county in two states for a period of 10 years, as well as the reported rates (adjusted for the age differences of the underlying population) of obesity for the same time period. Associations can be calculated between ice cream sales, the independent variable, and obesity, the dependent variable. This is valuable information. These data may shed some light on the relationship between ice cream and obesity: Do counties with high sales rates have increased

or decreased obesity rates? They also may inform us about changes seen in this relationship over time. As the ice cream sales increased, did the obesity rates increase or decrease? Over time, ecological studies can be useful in evaluating various hypotheses about the relationship under study.

While ecological studies are not without limitations, they are important and have stimulated important public health research studies and subsequent accomplishments. Ecological designs have been integral in understanding overall relationships in a number of areas of research. We can see an important historical example in the studies of blood lead levels during the 1970s and 1980s. The second National Health and Nutrition Examination Survey (NHANES) (a cross-sectional study), which took place from 1976 to 1980, found that 4.0% of children aged 6 months to 5 years had elevated blood lead levels. Data from the study revealed that childhood blood lead levels had declined by 37% from 1976 to 1980 and that this change appeared to have been accomplished by decreased lead in gasoline. The EPA estimated that consumption of lead in gasoline went down by 40% from 1970 to 1979 and concentrations of lead in ambient air decreased by 41% in the same time. Other studies had similar findings; one found that lead concentrations in umbilical cord blood specimens in a Boston hospital decreased from April 1979 to April 1980 in correlation with monthly gasoline lead sales. Another study examined gasoline lead emissions in New York City from 1970 to 1976 and found a correlation with trends in blood lead levels of children.

These are ecological studies because associations between the exposures and outcomes of interest were measured at the population level. There was no way to connect individual children's exposure to lead in gasoline and other sources with the level of lead found in their blood or with the subsequent mental disability and behavioral issues that arose. Randomized trials assigning participants to different sources and levels of lead exposure would have been clearly unethical, and it would not have been feasible for a large observational study to monitor children's cumulative exposures to lead from all sources from birth. Ecological studies had to be enough, and they were. These findings were sufficient to exert legal pressure that ultimately restricted the amount of lead permitted in gasoline and paint. The amount of lead used in gasoline has dropped from the 205,810 tons used in 1976 to the 520 tons used in 1990, a 99.8% reduction.

The primary limitation of ecological studies is called the *ecological fallacy* (also known as the *ecological inference fallacy*). This fallacy emerges because we do not know whether the association seen on the aggregate (group) level is true at the individual level. For example, despite our statistics that characterize the *group's* behavior, we know nothing about the individuals making up the group. Returning to our first example, suppose that counties with the highest ice cream sales have the highest rates of obesity. This supports the hypothesis that increased sales are associated with increased weight, but we can never know whether the people with the highest obesity rates are the ones eating all the ice cream. Many other explanations exist, including more people with obesity living in one area because of the availability of public health programs for weight loss, differences in the populations' income or activity level, and other factors that make the population-level data discordant with the individual level. There could be many reasons for the differences between the two counties that the data would not reveal because of the ecological fallacy. Another problem with ecological studies is that it is not possible to be sure about temporality: Which came first, the ice cream or the obesity, on the population level? This study cannot inform us about temporality or causality. Whether the dependent or the independent variable came first usually remains unknown until a stronger study design can evaluate the research

Table 2-2 Descriptive Epidemiology Summary Table

Design	Description
Descriptive studies	Studies that describe public health events with a detailed investigation into person, place, and time characteristics and generally involve no formal comparisons between groups. Hypothesis is generated, not tested.
Case reports	An unusual event is identified in one person. Generally report a new disease, a disease in a new type of patient or population, or something otherwise unusual. Description of the particular person with regard to person, place, and time and details of the condition are provided. Hypothesis is generated, not tested.
Case series	A group of case reports is assembled, with cases that represent a similar condition or situation. Where possible, data are provided to indicate usual frequency of event through estimates of rates of historical data. Hypothesis is generated, not tested.
Ecological studies	Collect aggregate data on exposures and aggregate data on outcomes provided for geographic areas or other population-level groupings. Hypothesis is generated, not tested.
Outbreak investigations	Specialized type of descriptive epidemiologic design that assesses acute disease or other public health events. Outbreak investigations attempt to identify the source of emerging or reemerging disease (or other public health emergencies), stop them, and prevent them in the future. Hypothesis is generated, not tested, although information derived from these investigations can often be used to immediately stem threats to public health.

question on the individual level. Still, ecological studies are important for generating hypotheses and can be valuable in suggesting associations that merit further study.

A summary of our primary descriptive epidemiology toolkit may be found in **Table 2-2**.

Taking Public Health Action

In addition to being the first part of any analytic study and to helping us generate hypotheses for future studies, descriptive studies alone can sometimes provide enough information to suggest a rapid intervention to protect public health. Although public health practitioners like to be sure about the cause of a particular outbreak, one important premise of public health is that action can and should be taken when it is evident that doing so can protect the health of individuals and the public. Whenever in the investigative process it becomes clear that there is a strong likelihood that a specific cause or behavior is associated with a specific health-threatening outcome, public health action should be undertaken to reduce exposure to the source. For example, thalidomide is a sedative now well-known for the birth defects it caused in children of women who took it for morning sickness while pregnant. In the early 1960s, the FDA blocked approval of the drug because of concerns about its safety, even though it was available in Europe (www.fda.gov). These concerns were confirmed when clusters of limb malformations in newborns of mothers who had taken thalidomide emerged in Canada and Europe, where the drug had been widely used. Thalidomide has been cited as the tragedy that increased research into drug safety during pregnancy and caused Congress to give the FDA greater authority to require more thorough drug testing before approval can be given.

Studies can be undertaken to provide evidence for causal relationships between exposures and negative health outcomes while actions are simultaneously conducted to immediately stop exposure. In an acute epidemic, rarely is it advisable to wait for a lengthy study to prove the exposure conclusively: people are sick and additional cases must be prevented. The urgency of public health action should be tempered, though, with appreciation for the rights of individuals and the fact that not all the information may be known at the time the action is taken. Public health action that can harm the rights of individuals in any way should be carefully considered before implementation. If the public health action could restrict freedom, cause discrimination or stigma about the disease or the exposure, or physically or emotionally harm those at risk for the outcome, care must be taken to determine the best next steps.

One outbreak that illustrates both the urgency of public health action and the necessity of accurate research is the epidemic of *E. coli* O104:H4 that occurred in Germany and other European countries in 2011. In Germany alone, 3,816 cases and 36 deaths were attributed to the epidemic. Early efforts to trace the bacteria back to an exposure produced misleading results: in a case-control outbreak investigation of one restaurant, cases were more likely to report having eaten a salad that contained leaf lettuce, tomatoes, and cucumbers. That salad also contained sprouts, but enough people failed to remember that ingredient during the survey that official suspicion mistakenly fell on cucumbers grown in Spain. By the time this hypothesis was rejected, Spain was reporting revenue losses of $286 million per week, and farmers' crops were rotting in their fields. Further epidemiologic studies eventually traced the source to locally grown fenugreek sprouts purchased from a specific seed distributor in Germany. This incident highlights the challenges we encounter while collecting data and how methods are integrated into what we find: remembering small but important details can be difficult, and we need improved, innovative approaches to facilitate the best possible data collection in every case. We also need methodological approaches to confirm findings rapidly, to avert this type of negative situation.

In the next section we will discuss ways of describing frequency of epidemiological events. Then, once we understand these methods of describing the public health issue at hand, we can proceed to testing hypotheses and beginning to assess the relationships between exposures and disease.

Taking Rapid Public Health Action

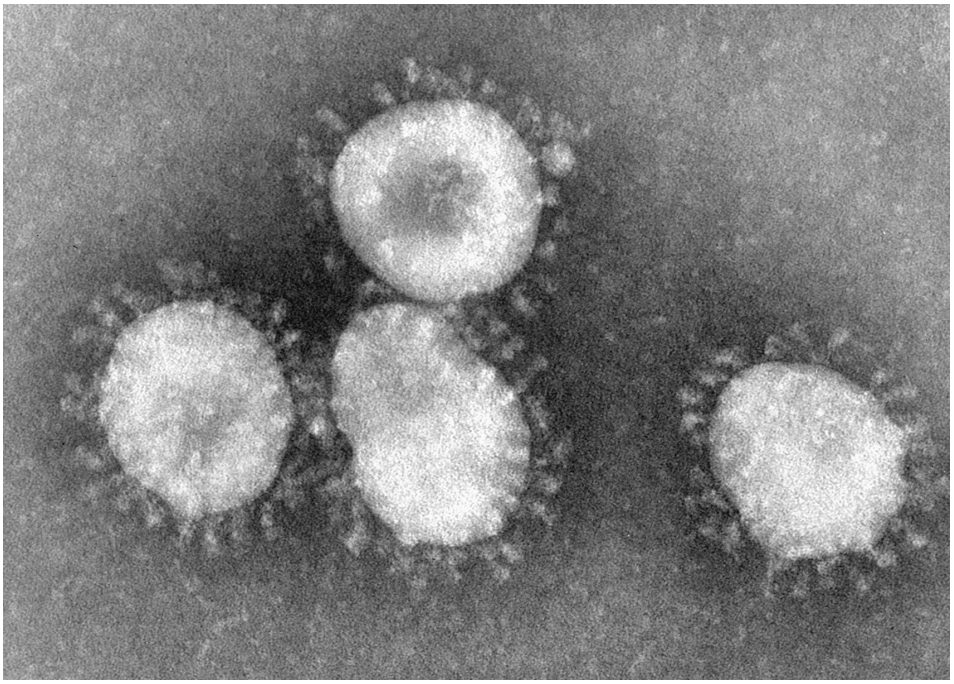
Taking public health action as quickly as possible is important. However, equally important is recognizing that sweeping actions can have negative as well as positive consequences. A good example is the actions surrounding discovery of severe acute respiratory syndrome (SARS).

When people became suddenly and severely ill with respiratory symptoms in 1999, the syndrome and method of transmission (droplet) were rapidly identified some time prior to a complete understanding of the causative agent, a coronavirus (see Figure 2-8). Several public health interventions were quickly taken: affected and potentially exposed individuals were sought, quarantined, and symptomatically treated. Certain flight restrictions from geographically affected regions were implemented. This rapid response may have limited the number of deaths from SARS. Once the causative agent was found in the laboratory and the disease better understood, some of these restrictions were lifted.

However, it is essential that public health personnel be also aware there can be hazards of too rapid a response. Because of generalized hysteria, for a short time people stopped flying in commercial

(*continues*)

airplanes, began wearing masks whenever they were in public in affected parts of Asia, and scorned those with coughs or SARS-like symptoms. Some of these actions were reasonable in close proximity to an ill individual: hand washing, masks, and quarantine make sense in case some are already ill. But the majority of persons coughing were ill with garden-variety colds, not SARS. The high level of intensity of response could have been capable of scaring individuals and prevent those who may have been ill from seeking care than reduce transmission of this serious virus. A tempered public health response might have been negative—reacting too late to the situation resulting in more deaths—but it also may have provided time to understand the virus's characteristics and develop responsible social marketing campaigns to help people into care without stigmatizing them. To be sure, hindsight is always "20/20," and at the time, it was impossible to know what we were dealing with and taking a conservative approach, most likely to protect the public's health is the best approach. But recognition that rapid action can have negative consequences as well as positive is important to note. At the same time, experience with SARS has a direct bearing on how we will handle avian flu, should this become pandemic. Active engagement in public health protection, using information gained from outbreak investigations, is necessary. In addition, ecological studies, based on trends in non-avian flu following September 11, 2001, suggest that the normal patterns of flu may have been significantly altered when air travel was curtailed in the United States following 9/11. In a case such as SARS or avian flu, stopping air travel, quarantine, and more extreme measures may make sense as far as stemming the epidemics. The limitations of ecologic studies mean that hypotheses regarding the efficacy of airline travel regulations, for example, may need more in-depth hypothesis testing before we understand for certain their benefit.



Courtesy of CDC/ Dr. Fred Murph and Sylvia Whitfield.

Figure 2-8 Coronaviruses are a group of viruses that have a halo, or crown-like (corona) appearance when view under a microscope. The coronavirus is now recognized as the etiologic agent of the 2003 SARS outbreak. This virus is a cousin of what we consider the common cold. Only in SARS, it is substantially more acute and deadly, and equally contagious.

Rates and Measures

How we count cases of disease and communicate them is a critical skill. In your introductory epidemiology course, you likely learned the basic formulae of rates and measures and what they mean. At this point in your professional career, you can already read the literature and interpret epidemiological findings. The next skill you will develop is to look at how the data were collected and limitations inherent in their collection so that you can better understand what the studies mean. In this section we are going to review straightforward skills on rates and measures, organically considering how the measures are built. This approach will hopefully allow you to use the measures fluidly, not as abstract letters and notation, but as concepts with which you are completely comfortable. Then we will layer on new ways to consider them and limitations that might not have been discussed previously, concepts that will allow you to better consider the limitations of the measures themselves and how they were established. Rates and measures are the nuts and bolts of how we understand and communicate disease frequencies.

Let us consider an example involving hepatitis C. Hepatitis C is most commonly spread via injection drug use (IDU). Approximately 90% of those with hepatitis C have no symptoms while the disease is in its acute state, so many people go without their disease being detected and as many as 80% develop chronic infection. Of those, half may develop cirrhosis of the liver or liver cancer. Hepatitis C is highly infectious compared to HIV: comparing single needle stick exposures, the estimated risk of acquiring hepatitis C is approximately 3% compared to a 0.3% chance of acquiring HIV and a 30% risk of hepatitis B. Studies of IDU suggest that between 50% and 95% of them have hepatitis C. In view of recent recognition regarding the elevated prevalence of hepatitis C, even among those without traditional high-risk characteristics, the CDC now recommends that persons born between 1945 and 1965, even without other risk factors, be tested for hepatitis C.

Imagine there are two counties that appear to have outbreaks of hepatitis C. In the past year, local hospitals in County A have reported 15 cases of acute hepatitis C, and those in County B have reported 13 cases. You begin with descriptive epidemiology and describe the characteristics of all individuals with hepatitis C with respect to person, place, and time. You count them; plot them on maps; and describe their gender, age, income level and occupation, use/nonuse of injection drugs, comorbidities (illnesses they have along with hepatitis C), and family members' or household contacts' hepatitis C status. You use laboratory studies to confirm they have hepatitis C and determine whether it is acute or chronic and other characteristics of the infection, such as the strain. You confirm the date and time of diagnosis and the proposed mode of infection. Now what?

There are several things to consider. Two outbreaks, or what seem to be outbreaks, in two different counties indicate a public health situation. You are not sure why each of the outbreaks has occurred, and specific public health action at this point is uncertain at best. To identify the cause of these disease counts and what can be done about them, it is necessary to move beyond a basic description. The first thing to do is establish the existence of an outbreak—just because someone labeled it as such does not mean it is one! You will need a way to compare the two counties to one another and to their individual baseline rates, both in the immediate past and over similar time periods in previous years. To make comparisons, you need information about

the underlying populations—in this case, the counties—as well as the cases themselves. This comparison depends on the use of a common denominator for each of the counties. We use rates and measures to create a common denominator and evaluate the existence of outbreaks. It is common for students to find these terms daunting, so rather than provide formulae at this point, let us continue with the example and explore it conceptually.

Investigating further, you find that the population of County A is 784,712 and that of County B is 1,500,546 (**Table 2-3**). To calculate the county-specific rates, you divide each county's rate (specifically, annual cumulative incidence) by its population and multiply by a common factor, such as 1,000 (for more frequent outcomes) or 100,000; then you can compare rates between the counties. By including the denominator in your evaluation, you can now see that County A actually has quite a big problem and County B's rates are within the county's normal range (0.5 to 1.0 per 100,000 annually), not much different than prior years'. While County B's rates of hepatitis C are high, it has a large IDU population, and hepatitis C has been fairly common there. Contrarily, in County A there were very few cases of hepatitis C in previous years, with rates ranging from 0.2 to 0.4 per 100,000. Its current rate of 1.91 cases of hepatitis C per 100,000 is alarming.

You begin your investigation and focus on determining what is happening in County A. After meeting with local infectious disease physicians, you find out that a new drug treatment center has opened up in County A. Upon intake, the new clients (who came from areas within and outside of County A) were screened for hepatitis C. Thus the increase in cases reflects the increased access to care of drug users who were already infected with hepatitis C and have now been diagnosed. This increase would be considered *artifactual*, that is, appearing to be an epidemic when in fact it is not. An artifactual epidemic is different from a true epidemic because it arises from the way cases are either classified or detected, not from an actual increase in the number of people with the outcome. In this example, the increase was because of a change in the way cases were identified (i.e., routine screening by the treatment facility). It is important to note that even though this is an artifactual hepatitis C epidemic and not a true epidemic, the cases are all real.

In this example, rates were a necessary part of comparing each county with previous periods within the county and to one another, evaluating the presence of a real or artifactual epidemic, identifying the location of the outbreak, and measuring changes and trends over time. If only absolute counts of cases had been available, they would have been useful in terms of quantifying resource need but not in determining whether a true outbreak were taking place. Absolute numbers cannot be used to project whether case numbers can be expected to increase, stay the same, or decrease, making future resource allocation difficult to determine. A true outbreak might be caused by a new, more infectious strain of the virus or an increase in the number of injection drug users in a network spreading the disease. It might have necessitated a rapid response, such as distribution of sterile

Table 2-3 Hypothetical County-Specific Rates for Hepatitis C, 2000

	Cases in 2000	Midyear population in 2000	Cases/Population	Rate per 100,000
County A	15	784,712	0.00001912	1.91
County B	13	1,500,546	0.00000866	0.87

needles for IDU. Without rates to identify whether there is an actual outbreak, it would not be possible to identify the source or take public health action to stem it. Rates also provide a means of communicating your findings to others in a common language that is quickly understandable by all.

Before we move on to discuss standardization, let's take a moment and consider the assumptions that went into construction of these rates: we assume the population figures, often midyear estimates (or sometimes mid-decade estimates) are correct. We assume that the diagnoses of the cases are correct and neither under- nor overestimate true disease. We make an assumption by presenting these data as crude (not adjusted by age, gender, race, etc.) that there are no significant differences based on these characteristics. This is not to suggest that the rates are incorrect but only to remind us that the statistics we use are only as good as the methods that are used to collect them. Often rates are constructed from the best data available to us, as we do not have access to estimates that might be more accurate; such estimates may not even exist. Whenever we look at figures such as rates we need to be mindful of all that goes into them. This allows us to see the richness of the values and their challenges and be able to better use them when we are making epidemiological or policy decisions.

Direct and Indirect Standardization

Differences in underlying population structure can create problems for us as we try to compare rates. Crude rates do not tell the whole story or adjust for how populations are constructed. The story of this problem is described clearly when we address confounding: underlying characteristics of a population can obscure our understanding of relationships between two variables. You can think of this most clearly in an example of where people live and die.

Imagine you have an assisted living facility in which all the residents have at least one serious morbidity and are at least 80 years of age. Down the street is an artists' loft living environment, where people have to have at least one visual art project currently being developed and must be under 25 years of age at the time they take residence in the loft. Without knowing anything further, which living environment would you guess has a higher rate of death, the assisted living facility or the artists' loft? Barring some dangerous exposure or another common factor that would put the loft dwellers in jeopardy, the elderly would be more likely to have a higher risk of death over, say, one year than have the younger individuals. This is simply because of the strength of the underlying association between age and death, certainly having nothing to do with the environment. Thus we have to be sure that we take the distribution of confounders—here, age, but it could be anything that is associated with both the exposure and the independent variable—into account when looking at associations.

Standardization is a similar construct. Denominators allow us to account for differences in underlying population size to identify the magnitude of an outbreak or of any public health measure. Similarly, standardization allows us to take into account differences in underlying population structure. Rather than focus on the calculations, which are covered in nearly every introductory epidemiology textbook, we will briefly discuss the purpose of standardization to make the process more intuitive.

To compare groups, we need to ensure we are looking at directly comparable measures (e.g., incidence rates and prevalence, *not* absolute numbers). We also need to standardize the

measures against a norm to ensure the natural makeup of subpopulations is not weighting your findings in the direction of a large subgroup that simply has more representation. Standardization is the application of one population's case rates to another's. There are two techniques: direct standardization, whereby subpopulation rates are applied to a standard population, and indirect standardization, whereby standard population rates are applied to a subpopulation.

These techniques allow comparison between two or more groups that differ with respect to their population distribution, a good example of which is age, though this is by no means the only one. Certain variables can distort relationships between the independent and dependent variables. When data are available, adjusting for additional characteristics, such as gender, race, and ethnicity, can also add clarity. Failing to account for the underlying population structure differences when comparing two areas produces misleading results, and we will not be able to understand the predictors or outcomes under study.

Here is another example, again involving age, with a quantitative illustration: imagine we want to compare the rates of death from hypertensive disease between two states, Louisiana and California. Louisiana has a much smaller population than California and has a different distribution of age groups. Comparing cause-specific rates (not absolute numbers) of death attributed to hypertension between the two is helpful but still does not help us see past the effect of age on death rates. Because an older population is more likely to die simply by virtue of age, it is difficult to determine what is happening or whether the rates are out of the ordinary. Only after adjusting for age can we compare the two states.

The data for this example are provided in the Discussion Questions section, where you can work the full example. We can see in this example that in 2009 there were more deaths attributed to hypertension among residents of California. If, however, all deaths in each state were summed and divided by the total population of that state and multiplied by a common factor, such as 100,000, Louisiana has a higher overall death rate attributed to hypertension. This trend is important to assess because it may indicate that Louisiana has a more significant hypertension problem than California has, or it may be that this is an artifact of the differences in age distribution between the two states.

Because we have data available about the standard population of the United States through population census measures, we will use direct standardization for this example. As you will recall, in general, we use direct standardization when we can and reserve indirect standardization for when the number of deaths or other outcomes is too small in some age groups to be reliably standardized, such as if the number of outcomes in that subgroup is fewer than 20 or we want to compare one population to a standard population.

Direct standardization yields the expected number of events (illness, mortality, etc.) in the standard population (such as the United States) if that same population had the event rate of the sample of interest. Direct standardization can be used to calculate an adjusted incidence (or mortality) rate. To use direct standardization to compare rates of mortality from hypertensive disease between the two states, you would follow this procedure:

- Determine the number of cases in the two states, grouped by age category.
- Obtain the populations of the two states and the projected population for the United States (for instance, from census data), broken down into the same age groups.

- Calculate the observed age-specific rate in each state by dividing the number of deaths in the age group by the state population of that age group, and multiply by a common multiplier, such as 100,000.

$$\frac{64 \text{ deaths in California}}{5{,}385{,}409 \text{ population in California}} \times 100{,}000 = 1.2 \text{ deaths per } 100{,}000 \text{ California state population}$$

- Multiply the age-specific rates from both states by the standard population size (for this age group it is 37, 233, 437) for each age category, and divide by the multiplier.

$$\frac{1.2 \times 37{,}233{,}437}{100{,}000} = 446.8 \text{ deaths in U.S. population}$$

- The result is the number of expected deaths if the standard population had the same death rate as did California allowing the age-specific rates to be compared. You will do this same process for Louisiana, and come up with 835.7 deaths in U.S. population as your directly standardized figure.
- After calculating these rates, it is then possible to calculate the adjusted incidence for each state using the total number of expected deaths (or cases) divided by the sum of the standard population.

$$\frac{\sum \text{Expected cases in standard population}}{\sum \text{Standard population}}$$

Indirect standardization achieves the same end, but produces the expected number of events in the sample of interest if it had the event rate of the reference population. Indirect standardization can be used to calculate the standardized mortality, incidence, or prevalence ratio to compare study populations with one another and the reference population. Indirect standardization works well if we do not have a stable population standard or cannot find a relevant data source by which to directly standardize. This approach begins with a comparison of two groups (cities, states, or other populations of interest) to the age-specific case or death rates in another standard population, such as state data. The next step is to compare what rates we would have found if each group had the same population distribution as the state standard and apply the standard's rates to the populations to determine the age-adjusted death rates from the disease or outcome of interest. Indirect standardization yields a standardized ratio (of mortality, prevalence, or incidence depending on the research question) by which populations can be compared, using the following methods:

- Multiply the age- and disease-specific death or case rate from the reference population by those of the study populations (cities, states, etc.) to calculate the expected death rate for each age group in those populations. If the reference rate is given per 100,000 or another common factor, divide by that factor for the expected event count.

$$\text{Age group}\left(\text{Ref rate} \times \text{Group n}\right) = \text{Expected \# of deaths}$$

- Add up the observed number of deaths from the two study populations.

$$\sum \text{Expected \# deaths}$$

- To calculate the standardized mortality ratio (SMR), divide the observed deaths by the expected deaths for each study population.

$$\text{SMR} = \frac{\sum \text{Observed \# deaths}}{\sum \text{Expected \# deaths}}$$

If the SMR is >1.0, then the number of deaths exceeds those expected; <1.0 means the SMR is less than expected. In some instances, the SMR is multiplied by 100; in those situations >100 is greater than expected and <100 is less than expected. If we have two populations, as in this example, then we can compare each city's SMR so we have a state-by-state comparison to each other in addition to a comparison with the standard.

Types of Measures

There are four basic types of measures: absolute counts, proportions, rates, and ratios.

Absolute Counts

In general, proportions, rates, and ratios are the most informative, but absolute counts can be valuable in assessing need and service utilization, as well as in the initial investigation of a public health concern. For example, imagine two hospitals are planning infectious disease units for their new wings. Tuberculosis (TB) patients and others with serious and transmissible airborne infections require rooms with specific types of ventilation systems and isolation. The hospital administrators and architects are meeting and need to know how often this sort of room will be required per year. If it is an infrequent event, only one such room might be added; if it is a common event, more rooms will be required. This clearly has cost implications and so cannot be taken lightly. Too few rooms would be a public health threat; too many rooms would be an unnecessary expenditure that would deprive other patients of services.

The administrator of one hospital checks her database and sees that over the past 10 years, the hospital's one isolation room was used an average (mean) of 2.3 times per year (standard deviation 0.35, range 0–4) and at no one time were two patients in need of the same room. The administrator of the other hospital checks his database and sees that over the past 10 years, his hospital's one external ventilation room was used a mean 5.8 times per year (standard deviation 1.2, range 4–10) and there were 17 instances in which multiple patients needed the room simultaneously, requiring them to be transported to alternate hospitals.

Clearly, this information is useful, even though it is only absolute counts. Although we still do not have a sense of how big a problem there is with TB or other respiratory infections in each of the locations and we could not yet compare their underlying rates to examine trends, the hospitals can now build new rooms appropriately. Absolute counts can additionally be used to

make some inferences. Imagine that one year each of these hospitals sees five TB patients. Even without knowing any more information, which of these hospitals would be concerned, the first or the second? Clearly the first. Five TB cases seem beyond its expected norm, whereas for the second hospital, five is at the lower end of its annual cases.

Proportions

Proportions describe variables in a context of a denominator. For example, if 54 children attended a day care where a rotavirus outbreak occurred and there were 16 cases, we could say that 16/54, or 29.6%, of the students became ill. This is called the attack rate: of those exposed, the number that became ill, a very intuitive measure although not a true rate. Note that the numerator is always included in the denominator. This proportion should be refined further to include elements of person, place, and time, for example, "within a 2-week period [time], 29.6% [proportion] of the day care attendees ages 6 months to 2 years [person] were newly infected [new cases] with laboratory-confirmed rotavirus [level of diagnostic certainty] at the Daycare Center [place]."

Rates

Many proportions describe time, but rates must include a time period in the measure. Rates state the number of cases divided by the population at risk within a given time frame. The population at risk is an important element of identifying the appropriate number for the denominator. For example, the hepatitis C scenario, described previously, defined a rate. Rates are typically expressed per a unit of population, a factor of 10 that makes the rate more understandable. Rates are especially useful because they describe the risk of the outcome under study of happening over a specific period of time. Thus, in the hepatitis C example, the risk of newly contracting hepatitis C cases in the 8-week time period described is 1.9 per 100,000 in City A and 0.87 per 100,000 in City B. Note that in this scenario we assume that all members of the population are at risk of acquiring hepatitis C for the entire time. This may not be a good assumption, because not all persons may be at risk if they are immune or protected in some other way.

Ratios

Finally, ratios are a measure whereby the numerator is divided by the denominator but the numerator is *not* contained within the denominator. For example, let us consider autism spectrum disorder (ASD). Autism affects some groups more than others, occurring more frequently in males than in females and varying by race/ethnicity as well. One study found that depending on the state, the male:female autism ratios can be from 2.7:1 to 7.2:1, indicating that for every one affected female child, there are 2.7 to 7.2 male children with autism. It was also observed that the state with the most male cases had the lowest male:female ratio and the state with the lowest number of male cases had the highest male:female ratio. Note that the numerator in each of these is not contained within the denominator; that is, male and female categories are mutually exclusive. This is the salient distinguishing characteristic between a proportion, or rate, and a ratio.

How can a ratio help us? In this example, these ratios can give researchers material to consider as they work to determine risk factors and the mechanisms by which autism occurs. Other health

issues might make use of ratios such as outdoor:indoor occupation or urban:suburban:rural residence. A high outdoor:indoor occupation ratio might indicate that an illness is caused by a tick or a mosquito that prefers to feed outdoors. Ratios can provide insight into the etiology of an agent, what it might be, and how to address confounders in the analysis of the data.

Incidence Measures

Incidence measures are among the most important types of measures in epidemiology. Incidence measures tell us how many *new* cases of a disease occurred in a given time frame instead of how many cases exist. They allow us to examine what is happening at the moment with regard to cases instead of everything that happened up until the study period. When a problem is first identified, incidence is all we have: the number of new cases in the past month. But after that, things get increasingly complex, because in addition to newly diagnosed cases, there are individuals who have had the disease for some time, those who had it but recovered, and others who have it but are undergoing treatment, recurrences, and so forth. Incidence measures are rates and as such, involve two key elements: the number of *new* cases during a specified period of time and the number of people *at risk* of developing the disease in the same time period.

When we calculate incidence, the number of new cases during a specified period of time is the numerator. Distinguishing between new and existing cases is a crucial determination in the development of useful and accurate rates. In a textbook, the difference is generally quite straightforward, with exact numbers of new cases during a given time period so the reader can easily calculate the incidence, whereas in practice, the distinction between new and existing cases can be difficult.

One example of the difficulty of this measure is seen in sexually transmitted disease (STD) research. Imagine calculating the incidence of *Chlamydia trachomatis*, a bacterial STD especially common among adolescents and young adults. Unlike many viral infections, individuals may have repeated cases of this infection, so counting them can be tricky. One can count initial new infections as well as existing infections (the prevalence), but accurately counting only new infections can be challenging. Yet the accuracy of our measures depends on our being precise in our counts and measures and differentiate new from existing and repeating cases.

The number of people in the population at risk of developing the disease over the same period of time is the denominator of an incidence rate. For some diseases, this is straightforward. For example, in measuring the risk of ovarian cancer, the population at risk clearly includes only women and among women, only those with ovaries. In other diseases, however, we face challenges in identifying an appropriate denominator. Those at risk of developing the disease over the specified period of time need to be *susceptible* to the disease. How do we estimate that? It depends on the disease under study. Diseases, such as measles, that confer immunity following illness and generally, following vaccination (94% to 98% vaccine efficacy following the first immunization and 99% with a subsequent booster) would render a different at-risk denominator than that of a bacterial infection, which a person could contract repeatedly.

As we saw with the *C. trachomatis* example, we could have multiple incidence measures with fluctuating numerators and denominators. Incidence—new first cases—of the STD

could be calculated using the denominator of those who have never had the disease before and the denominator of all sexually active individuals. This, too, could be considered in multiple ways. For example, if we had the data, we could evaluate only those sexually active individuals having unprotected intercourse. A different incidence could be calculated looking at the number of new reinfections during the specified time period as the numerator and the number of sexually active individuals who had contracted *C. trachomatis* once before during the specified time period. This could be calculated repeatedly, because repeated STD infections are extremely common. Thinking through the population at risk and the definition of a new case is critical.

As you move forward in your studies, it will be increasingly necessary to differentiate between new and existing cases and between populations at risk and not at risk. You need to be able to think through the nuances of each data source and the meaning those nuances will impose on each measure. The ingredients in an incidence measure are fairly straightforward:

- New cases: These are cases that count as incident cases among those at risk during the time period of interest.
- Time period: The time period specified is crucial to incidence. Like other rates, the time period must be the same for the numerator and the denominator and must be specified clearly for the incidence rate to be correctly interpreted.
- Multiplier: In general, to make comparisons simple and avoid very small or noninteger incidence rates, a multiplier is often used, usually a factor such as 1,000 or 100,000. The result is the division of the numerator by the denominator, multiplied by this factor. This simplifies presentation of the incidence rates and comparisons with other rates. Because the multiplier is applied to both the numerator and denominator, it does not affect the incidence itself.

There are two types of incidence measures: cumulative incidence (CI) and incidence rate (IR), also known as incidence density (ID) or incidence density rate (IDR). Conceptually, both types of incidence measures are similar: both express the number of new cases in a population; they differ in their denominator.

Cumulative Incidence

This measure of risk assumes that all people in a population over a specific period of time were observed for the entire time:

$$\text{CI} = \frac{\begin{array}{l}\text{the number of new cases of disease}\\ \text{among those at risk of the disease in a}\\ \text{population over the given time period}\end{array}}{\begin{array}{l}\text{the number of individuals in that}\\ \text{population over the same time period}\end{array}} \times \text{multiplier}$$

We are seldom able to assume that each individual in a population was followed, but the measure can still be used if the basic tenets are met. That is, we need to be able to make some assumptions about the stability of the population. Most populations are dynamic, with people constantly entering and leaving the population, so estimating the true denominator at risk can

be an impossible task. We often use the CI even when we know that we cannot meet this assumption. Yet it is important to be aware of the limitations that exist in the data that help create these measures. There are better measures to use, such as incidence rate, when we can.

Incidence Rate

We can obtain a true rate if we follow each member of our study population and record how long each person is observed for the outcome of interest and when she or he meets the end point (outcome) of the study. This measure directly integrates time into the denominator. Our measure of risk is then described by the number of new cases of disease among those at risk of the disease in a population over the given time period:

$$\text{IR} = \frac{\begin{array}{l}\text{the number of new cases of disease}\\ \text{among those at risk of the disease in a}\\ \text{population over the given time period}\end{array}}{\begin{array}{l}\text{person-time followed over the same}\\ \text{time period}\end{array}} \times \text{multiplier}$$

It is important to note the differences between these two measures. IRs are commonly used in longitudinal studies, such as natural history studies, cohort studies, or clinical trials. CIs are used when information about the midyear population at risk is present but no individual-level information is available.

Here is an example of how to calculate CI and IR. City H is a hypothetical city in the United States. In 2013, there was an increasing concern about a new, emerging disease:

Midyear population of City H, 2013	950,000
New cases of emerging disease under study between 1/1/2013 and 12/31/2013	1,020
Prevalent cases of disease on 1/1/2013	1,403
Estimated population of City H on 1/1/2013	876,449

Everyone is at risk of the new disease, and it appears that those who have been ill once can become ill again. Risk of the disease appears to vary geographically within the city. The public health physicians develop a protocol that they submit to the relevant institutional review boards and local agencies. They invite clinic attendees at high risk of the disease to participate in a study to find out factors associated with the disease. They follow participants over 3 years with quarterly visits. At each visit, detailed clinical and behavioral data are collected, as well as laboratory specimens, to evaluate exposure to the disease of interest. Participants accrue follow-up time as long as they do not miss a study visit. Those who develop the disease or leave the study no longer accrue person-years of follow-up.

People in cohort	Person-years under study of cohort members	New cases over 3 years of follow-up
750	2,070	59

CI for the disease in City H in 1999 can be calculated as follows:

$$\text{CI} = \frac{\begin{array}{l}\text{the number of new cases}\\ \text{of disease among those}\\ \text{at risk of the disease in a}\\ \text{population over the given}\\ \text{time period}\end{array}}{\begin{array}{l}\text{the number of individuals}\\ \text{in that population over the}\\ \text{same time period}\end{array}} \times \text{multiplier} \quad \begin{array}{l}\text{(here we will use}\\ 100,000\text{, but it could}\\ \text{be any factor applied}\\ \text{to both numerator}\\ \text{and denominator)}\end{array}$$

$$= [1,020 / 950,000] \times 100,000 = 107.4 \text{ per } 100,000$$

This is how the IR for this study could be calculated. IR =

$$\frac{\begin{array}{l}\#\text{ of new}\\ \text{cases in}\\ \text{at-risk}\\ \text{population}\\ \text{in time}\\ \text{period}\end{array}}{\begin{array}{l}\text{Person-time}\\ \text{followed over}\\ \text{time period}\end{array}} \times \text{multiplier} = \frac{59}{\begin{array}{l}2,070\\ \text{person-}\\ \text{years}\end{array}} \times 100,000 = 2.85 \text{ per } 100 \text{ person-years}$$

$$\text{IR} = \frac{\#\text{ new cases in at} - \text{risk population during time period}}{\text{units of person} - \text{time followed during the time period}} \times \text{multiplier}$$

$$\text{IR} = \frac{59}{2,070} \times 100,00 = 2.85 \text{ per } 100 \text{ person-years}$$

Here we used 100,000 as the multiplier, but we could have applied any factor to both the numerator and denominator. Note that if a person were rendered no longer susceptible to the disease after having it, the number of people at risk in the population would change from year to year as those who had already been ill were removed from the denominator, thus reducing the number of people at risk in the population. Note that this differs, too, from the CI that would come from the cases on 1/1/13 (that is, 1403/876,449) as well as from the percentage of the cohort followed that had the outcome (that is, 59/750).

Kaplan-Meier Methods for Calculation of Incidence Rates

Accounting for varying individual time under follow-up as well as for situations where the outcome of interest is unknown is an important skill in descriptive epidemiology. The latter

situation called *censoring*. Censoring occurs when information on a given outcome is not available, for any number of reasons, such as loss to follow-up (one of the most common biases in prospective studies), inadequate information on the outcome, or removal from the risk set prior to occurrence of the outcome. Since not all the people in the group under study are followed for the same time period and have available data, we have to account for and properly calculate person-time to assess the risk estimates. There are two primary ways to calculate individual incidence using person-time: classic life table methods (also known as actuarial or interval-based life tables) and the Kaplan-Meier approach. The two methods are similar, yet because of the latter's strengths we will focus on the latter.

In **Table 2-4** you will see a hypothetical cohort study examining time to recurrence of chlamydia in women diagnosed with pelvic inflammatory disease who were hospitalized for treatment. What each column stands for is shown in the second row. Note that the Kaplan-Meier approach can be performed for any event, outcome, or survival depending on the study of interest.

There are three important things to notice:

1. In the time column, the time intervals are not arbitrary or preset: they reflect the actual times of events. We could have many more or fewer, as dictated by the data themselves. This allows us to reflect the instantaneous force of morbidity, the risk estimates based on the time the event occurred. We do not have to use an artificial correction factor and

Table 2-4 Kaplan-Meier Approach to Calculating Incidence

i	n_i	d_i	Conditional Pr(recurrence) $q_i = \frac{d_i}{n_i}$	Conditional Pr(no recurrence) $p_i = 1 - q_i$	Cumulative probability of no recurrence S_i
Time in months of follow-up	**Number of women at risk during each interval**	**Number of recurrences of chlamydia**	**Given a woman survived until this interval, probability of recurring through the interval**	**Given a woman survived until this interval, probability of not recurring through the interval (counterfactual of having the event)**	**Cumulative probability of not having the event (survival function)**
1	300	2	0.007	0.993	0.993
3	289	5	0.017	0.983	0.993 × 0.983 = 0.976
4	288	4	0.014	0.986	0.963
6	285	4	0.014	0.986	0.949
9	286	8	0.028	0.972	0.923
11	284	9	0.032	0.968	0.893
12	280	3	0.011	0.989	0.884

guess when the event happened or the number of people in the denominator changed; it is incorporated into the measure.

2. The number of people (n_i) allows incorporation of the changing denominator. The number per row is allowed to fluctuate based on the number of people studied and this censored.
3. As with all statistical measures, we need to be mindful of the required assumptions. For the Kaplan-Meier approach, we have to make the assumption that survival (or not having the outcome of interest) is independent of censoring. Another way of saying this is that people who are censored should have the same probability of having the event under study as those not censored. This relates to a bias we are concerned about, especially in regard to loss to follow-up. It is possible that people are censored as a result of a study-related issue (e.g., not satisfied with the study, have side effects to or do not like the treatment, etc.), which can create bias. But the question of the assumption is not only whether those people are different but whether those people who are censored have the same probability of the event as those people who stay in the study and are not censored. This must be true for this method to apply. In addition, we need to have an absence of secular trends strongly associated with the variables under study during the study period. In this example, such a secular trend might be a change in treatment modalities or new prevention opportunities.

Graphing the results of Kaplan-Meier estimates is often helpful since graphing highlights visually the number of people on whom risk estimates are being based at each step and allows clear visualization of the differences (if there are any) with regard to survival for the groups being compared. Here is a graph that looks at a study over 3 years and compares group A to group B with regard to survival.

For completeness, it is good to understand the differences in approach between the Kaplan-Meier method and the life table calculations. The life table method uses preset intervals instead of having events drive the timing of the intervals. By dividing the study period into even intervals, say 2 years, the life table approach provides conditional probabilities of

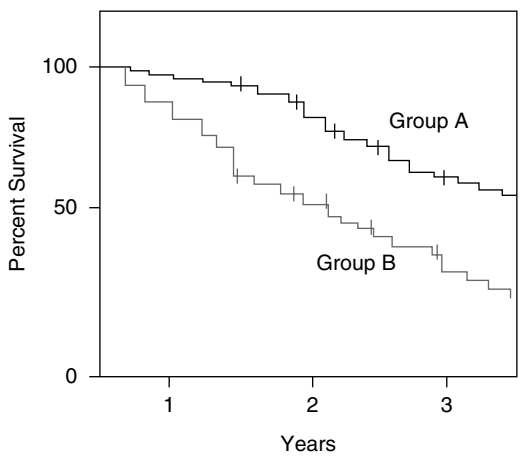


Figure 2-9 Graph x (years) vs. y (percent survival) comparing Group A to Group B.

having the outcome through each interval. Because the life table approach uses arbitrarily assigned intervals, if censoring occurs in the middle of an interval, it is imprecisely accounted for using a correction factor, which assumes that the censoring (or event) happens halfway through the interval, even if that is not for certain. Because the Kaplan-Meier method calculates the exact times of events, it is not only easier to use but also accounts for events at the time of each event's occurrence. Finally, there must be an additional assumption beyond those given previously, that there is uniformity of when the losses to follow-up and the events occur during each interval. This is what allows the correction factor to be used as well. If you cannot make that assumption, the life table method is usually not appropriate and a Kaplan-Meier method is preferred.

Prevalence Measures

Prevalence measures are commonly used and represent a merging of information; they merge existing cases with new cases, describing how many people there are with the characteristic. There are two types of prevalence measures: point prevalence is the number of existing and new cases at a given moment, and period prevalence is the number of existing and new cases over a stated period of time. The denominator for each measure is the number of persons in the population over that same period of time. For example, to find the point prevalence of prostate cancer on January 1, 2000, we would identify all cases on that day, irrespective of when they had been diagnosed, and divide that number by the number of persons in the population on that same day.

$$\begin{aligned}\text{Point prevalence}_{1/1/1999} &= \frac{\begin{array}{l}\text{new and existing}\\ \text{cases on } 1/1/1999\end{array}}{\begin{array}{l}\text{population on}\\ 1/1/1999\end{array}} \times \text{multiplier}\\ &= [1,403/876,449] \times 100,000\\ &= 160.1 \text{ per } 100,000\end{aligned}$$

To calculate period prevalence between January 1, 2000, and December 31, 2000, we would identify all new and existing cases during that time period and divide that number by the number of persons in the population during that time period.

$$\begin{aligned}\begin{array}{l}\text{Point}\\ \text{prevalence}_{1/1/1999 \text{ through } 12/31/1999}\end{array} &= \frac{\begin{array}{l}\text{new and existing}\\ \text{cases on } 1/1/1999 +\\ \text{new cases } 1/1/1999\\ \text{through } 12/31/1999\end{array}}{\begin{array}{l}\text{estimated mid-year}\\ \text{population for } 1999\end{array}} \times \text{multiplier}\\ &= [(1,403+1,020)/950,000] \times 100,000\\ &= 255.1 \text{ per } 100,000\end{aligned}$$

Similarly, this is how to calculate the period prevalence of disease in a study cohort:

$$\text{Prevalence of disease in cohort} = \frac{\text{number of cases over follow up}}{\text{number of participants}} \times \text{multiplier}$$
$$= [59/750] \times 100$$
$$= 7.9 \text{ per } 100 \text{ participants}$$

Why bother with prevalence when we have incidence? They both contribute valuable and differing information. Incidence can help identify epidemics, healthcare disparities, access issues, diagnostic changes, and much more. Prevalence helps quantify needs for care, such as how many people are living with a disease, and observe changes in treatment (improvements, declines). Often we might wish to measure incidence, but we are not able to ascertain new cases or the number of persons at risk, so prevalence must suffice.

Relationship between Prevalence and Incidence

One characteristic of prevalence and incidence is that their relationship is both intuitive and quantifiable. As more individuals become sick, they increase the prevalence, with new and existing cases increasing together, that is, until one of two things happens: there is a treatment such that the prevalence goes down as people are cured, or the disease is sufficiently serious and the sick die such that the prevalence goes down. In both of these cases—cure or death—the same thing happens, though for different reasons: the duration of the illness is shortened. Prevalence is a function of both incidence and duration. It is mathematically expressed as follows:

$$\text{prevalence (P)} = \text{incidence (I)} \times \text{duration (D)}$$

We can think of the relationship between prevalence and incidence this way: incident cases of any disease or condition occur. Once this happens, affected persons have the following outcomes: cure/resolution/remission, continuing illness, or death. Individuals who either die or are cured do not become prevalent cases; those with continued disease do. Thus, as duration is shorter as a result of either death or cure, prevalence is lower, thus holding incidence the same. Similarly, if duration is longer because of longer times of living with the disease, improved treatment, or decreased likelihood of death, prevalence is higher, thus holding incidence stable.

Consider a fictional example of a newly discovered, rapidly terminal cancer (**Figure 2-10**). In 1980, this hypothetical new form of cancer was characterized. For many years, there was no specific diagnostic test, treatment, or preventive intervention. The mortality rate was therefore high and the prevalence was low. Then in 1990, an improved diagnostic test reached the market, causing a sudden rise in the observed incidence. The mortality rate continued to be high and, as the usual cancer treatments were not effective, the prevalence increased in direct relation to the increased incidence. In 1995, after much research, the first drug targeting the cancer was

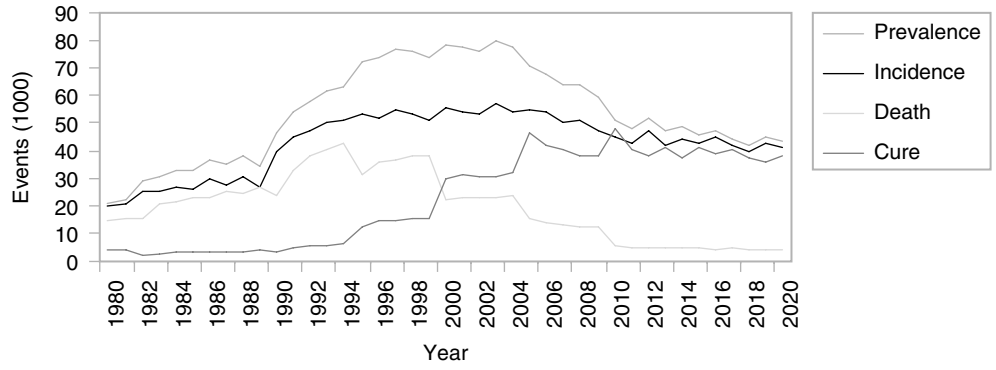


Figure 2-10 Hypothetical cancer prevalence as a function of incidence and duration.

approved by the FDA and was put into use, increasing the rates of cure and survival with cancer. Subsequently, the prevalence dramatically increased as more people developed and survived cancer instead of dying. Another drug was approved in 2000. This one further decreased cancer mortality and increased cure rates, but because the same proportion of patients left the population as before, the prevalence stayed about the same. During this time, it had become clear that the cancer was associated with exposure to an environmental toxin that had become common in treated water, and legislation was signed in 2005 banning use of the chemical. Over time, this move reduced the cancer incidence, concurrent with further improvements in early treatment, and the prevalence began to steadily decline.

Specialized Measures

Several specific measures are often used to describe the effect of disease: Case fatality rate (CFR), attack rate, and vaccine efficacy. Each of these measures has the same underlying assumptions that exposures and outcomes are carefully operationalized using a common case definition, individuals are comprehensively followed forward in time, and there is a specified time period under study.

Case Fatality Rate

The case fatality rate (CFR) is a measure of disease severity expressing as a proportion how many individuals with a disease die of the disease. Early in the epidemic just mentioned, 1,750 of the 2,500 people who had been diagnosed with the novel cancer in 1982 died that year, a 70% CFR. Why is this measure a rate and not just a proportion? Though it can be indistinct without all the specific information explicitly provided, note that the CFR does contain an explicit or implicit measure of time. The time frame may be hard to find if it is in a narrative articulation of the CFR, but it will be there. The article or report may say something such as "between January 1 and June 25, 1995, there were 100 people infected and 4 died" so that the specific time period is made clear. CFRs assume that all individuals were diagnosed with the same disease, using the same case definition of that disease, and that they were all followed for the duration of the time period so that their ultimate vital status could be ascertained. An example may be found in **Table 2-5** and in the CFR box.

Case Fatality Rates

When we currently think of influenza pandemics, we often think first of the 1918 pandemic of the Spanish flu, which killed between 20 million and 50 million people globally, vastly more than the number killed in the world war that was just ending. Of course, in between there have been other pandemics (e.g., the 1957 Asian flu and 1968–1969 Hong Kong flu, most notably). And every year, thousands die of pneumonia and influenza, with the National Center for Health Statistics in 2000 ranking it seventh in all cause mortality for all ages and races (with 65,313 deaths), and fifth in the leading causes of death among persons ages 65 years and older. One relatively recent flu concern surrounds Avian Flu, H5N1. This strain carries the fear of what could be the next strain to turn into a pandemic.

Avian flu, a highly pathogenic avian influenza A (H5N1) virus, has occurred primarily among poultry in Asia; direct cases of transmission from poultry to humans have been seen, but very few human-to-human cases. Still, a strain that we humans are not well-acquainted with could yield us relatively helpless: our shield of immunity is not primed against this strain, as with what happened in the 1918 flu (H1N1). If the avian flu shifted slightly to simplify human-to-human transmission, a swift pandemic could ensue; this remains a concern globally. The World Health Organization (WHO), CDC, and other public health organizations are on the alert with active surveillance and intervention (e.g., reduction of infected poultry populations, vaccinations in development, strategic plans written, etc.). Statistics are kept by the WHO and are available. These provide a good opportunity to see case-fatality rates in practice (**Table 2-5**).

How does one read this table? First, notice that the time period (required for a case-fatality rate) is annually, and the regions are countries. This is a very large geographical unit as well as a large time period, yet these are the data available to us, so we can calculate a case-fatality based on them, provided we are clear about what data we have. As we have seen, the case-fatality rate is calculated as:

Case-fatality rate = number of deaths during a specific time period/ number of cases during the same time period

So if we are interested, for example, in the case-fatality rate of Indonesia in 2006, our figure would be:

Case fatality = 45/55 = 81.8%

Note the notations at the bottom of the table, and that this only includes lab-confirmed cases. Avian flu is a serious diagnosis, yet many of its symptoms are like those of other viruses influenzas in particular. If you had to guess, do you think this number of cases under- or overestimates the true number of cases of avian flu to date?

Major flu epidemics and the responsible strains are shown in **Figure 2-11**. Influenza is a virus with three recognized types: A, B, and C. Two antigenic properties of the surface glycoproteins, hemagglutinin (H) and neuraminidiase (N), help to classify the influenza A subtypes; thus, when we note the combination, say H5N1, it denotes a specific subtype of influenza A. Due to the recombining of antigens, called antigenic drift, these viruses shift constantly, changing over time. This is why there need to be new vaccine compositions developed for the flu shot each year, providing the correct subtypes for each year. Though as humans we have developed immunity to subtypes that are common, we do not have defenses against avian flu, perhaps making us more vulnerable. A good example of the importance of surveillance for influenza occurred in 2009 with the pandemic of H1N1 (Swine Flu). Because of rapid public health response and vaccine development and distribution, significant impact of the virus was prevented. This is important not only because of the widespread nature of the disease but also because it affected certain populations, including pregnant women and children, more severely than typical flu strains. **Figure 2-12** shows the global distribution of H1N1 early in the epidemic in 2009.

(*continues*)

Table 2-5 Cumulative number of confirmed human cases for avian influenza A(H5N1) reported to WHO, 2003-2009*

	2003		2004		2005		2006		2007		2008		2009	
Country	**cases**	**deaths**	**cases**	**deaths**	**cases**	**deaths**	**cases**	**deaths**	**cases**	**deaths**	**cases**	**deaths**	**cases**	**deaths**
Azerbaijan	0	0	0	0	0	0	8	5	0	0	0	0	0	0
Bangladesh	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Cambodia	0	0	0	0	4	4	2	2	1	1	1	0	1	0
China	1	1	0	0	8	5	13	8	5	3	4	4	7	4
Djibouti	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Egypt	0	0	0	0	0	0	18	10	25	9	8	4	39	4
Indonesia	0	0	0	0	20	13	55	45	42	37	24	20	21	19
Iraq	0	0	0	0	0	0	3	2	0	0	0	0	0	0
Lao People's Democratic Republic	0	0	0	0	0	0	0	0	2	2	0	0	0	0
Myanmar	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Nigeria	0	0	0	0	0	0	0	0	1	1	0	0	0	0
Pakistan	0	0	0	0	0	0	0	0	3	1	0	0	0	0
Thailand	0	0	17	12	5	2	3	3	0	0	0	0	0	0
Turkey	0	0	0	0	0	0	12	4	0	0	0	0	0	0
Viet Nam	3	3	29	20	61	19	0	0	8	5	6	5	5	5
Total	4	4	46	32	98	43	115	79	88	59	44	33	73	32

* *Total* number of cases includes number of deaths. WHO reports only laboratory-confirmed cases.

Reproduced, with the permisison of the publisher, from World Health Organization. Cumulative number of confirmed human cases of avian influenza A (H5N1) reported to WHO. Available at www.who.int/influenza/human_animal_interface/H5N1_cumulative_table_archives/en/

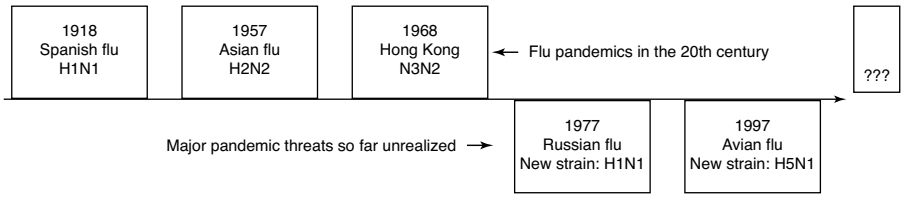


Figure 2-11 Major Flu Pandemics since 1900.

Data from Centers for Disease Control and Prevention. Avian influenza infection in humans (bird flu). 2006. Available at: www.pandemicflu.gov.; Centers for Disease Control and Prevention. Pandemics and pandemic threats since 1900. 2006. Available at: www.pandemicflu.gov/general/historicaloverview.html.; National Institute of Allergy and Infectious Diseases, National Institutes of Health. Focus on the flu: timeline of human flu pandemics. 2006. Available at: www3.niaid.nih.gov/news/focuson/flu/illustrations/timeline/timeline.htm.; Kilbourne ED. Influenza pandemics of the 20th century. Emerg Infect Dis. 2006;12(1):9–14.; Krause R. The swine flu episode and the fog of epidemics.Emerg Infect Dis. 2006;12(1):40–43.; World Health Organization.Cumulative number of confirmed human cases of avian influenza A(H5N1) reported to WHO. Available at: www.who.int/csr/disease/avian_influenza/country/cases_table_2006_12_27/en.; and Taubenberger J. 1918 influenza: the mother of all pandemics. Emerg Infect Dis. 2006;12(1):15–22.

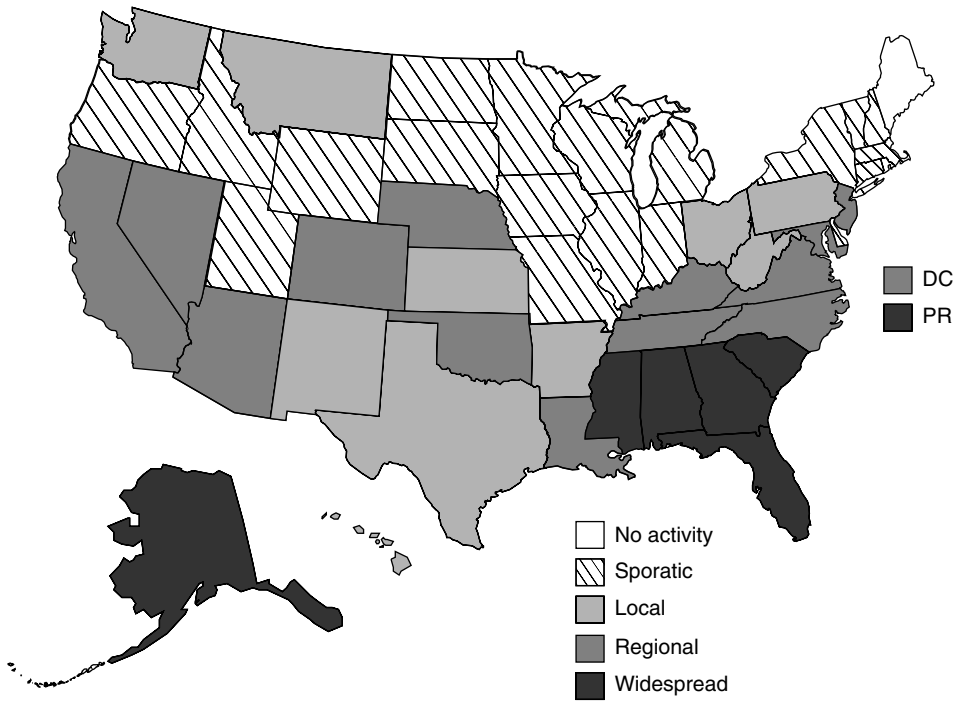


Figure 2-12 Timeline of Global H1N1 Cases, 2009.

Reproduced from Centers for Disease Control and Prevention. Update: influenza activity--United States, April -- August 2009. *MMWR*. 2009; 58(36): 1009-1012.

Attack Rates

Attack rates (ARs) parallel CFRs, providing us with the number of people at risk who developed a disease or had an outcome over a specified period of time. For example, if 50 children with asthma were outside on a hot day with high smog levels and 20 had asthma attacks, the AR would be 40%.

Table 2-6 Summary Table—Rates and Measures

Incidence	New cases per unit time
Prevalence	New and existing cases per unit time
Attack rate	Proportion of persons with the given exposure who developed the outcome of interest per unit time
Case fatality rate	Proportion of persons with disease who die from the disease per unit time
Vaccine efficacy	Proportion of persons immunized who retain protection from disease per unit time

Vaccine Efficacy

Vaccine efficacy (VE) tells us how many out of those immunized against a specific disease did not develop that disease during a specified period of time. For example, the series of pertussis immunizations has a VE of 80% for the first 5 years, after which protection begins to wane. If 100 children were immunized against pertussis, given the average VE over the first 5 years, an estimated 20% of the children would be susceptible to pertussis. Note that this does not mean they *would* get the disease but only that they could be susceptible if exposed.

A summary of rates and measures may be found in Table **Table 2-6**. After this refresher on ways to describe disease occurrence, we will move on to specific details of analytic studies, one of the most exciting aspects of epidemiology. In analytic studies, you will hone your skills to compare exposed and unexposed persons and cases and noncases, understand interventional approaches, and answer the research questions you will be able to articulate.

Discussion Questions

1. Review the 2011 MMWR *Surveillance Summaries* paper titled "Out-of-Hospital Cardiac Arrest Surveillance—Cardiac Arrest Registry to Enhance Survival (CARES), United States, October 1, 2005–December 31, 2010 (available at http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6008a1.htm?s_cid=ss6008a1_w). Describe the person, place, and time attributes found in this study with relation to those who experienced and those who survived the cardiac arrest.
2. Go to the data warehouse provided by the World Health Organization (WHO) (http://apps.who.int/gho/data/). You are going to download a simple data set of your choosing (not the range of publicly available data: everything from vehicles registered to immunizations to maternal deaths to infectious disease has data available for public access analysis). Any exportable data set from this site is fine to use. Import the data set into your statistical programming software and simply describe it in *every way you can*—categorically, continuously, any way to describe the data that is appropriate to the data set you have selected. For example, if you looked at the numbers of vehicles by country, you could provide measures of central tendency and dispersion (means, median, mode, IR, standard deviation, etc.) for all of them, or you could recode to look at regions (e.g., by continents,

quadrant, country geography, mountainous, sea level, etc.). You can graph overall and by whatever category interests you and observe outliers and patterns. You can recode to categorize (e.g., look at countries with > or ≤ than the median, divide into tertiles, quartiles, etc.). You can map if you know GIS. You can plot. The point is to get to know this simple data set so that you practice describing data. If you wish to make this more challenging, use several outcomes and develop ecological analyses of them. For example, do locations with increased vaccination coverage also have increased vehicles? Be imaginative.

3. Using http://www.cdc.gov/DiseasesConditions/ or other resources, identify conditions and epidemiological questions that could be answered using the descriptive epidemiologic study designs described. Choose a condition or disease and practice matching each of the descriptive designs referenced in this section to a condition or disease so you can have practice with each of the designs. What are some of the strengths and weaknesses of each design and approach?
4. Referring to **Table 2-7**, calculate the observed death rates in adults for both states. Does one state seem to have a more severe hypertension problem than the other?
5. Referring to **Table 2-8**, calculate the death rates that would be expected for each age group in the overall adult U.S. population given the age-specific death rates of the two states. After adjusting, how do the death rates compare between the two states?
6. Using the expected number of deaths in the standard population given the California and Louisiana rates, calculate the adjusted mortality incidence from hypertensive disease for the two states.

California

$$\frac{\sum \text{exp}}{\sum \text{pop}} =$$

Louisiana

$$\frac{\sum \text{exp}}{\sum \text{pop}} =$$

Table 2-7 Direct Standardization of Deaths from Hypertensive Disease in 2009

	California			Louisiana		
Age	**Obs # deaths**	**Population**	**Observed death rate/ 100,000**	**Obs # deaths**	**Population**	**Observed death rate/ 100,000**
25–34	64	5,385,409	1.2	14	624,512	
35–44	202	5,218,771		45	561,813	
45–54	630	5,199,654		144	642,810	
55–64	916	3,838,067		221	507,753	
65–74	930	2,179,099		197	300,183	
75–84	1,740	1,366,454		287	183,822	
85+	3,419	602,502		382	70,291	543.5

Data from Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2009 on CDC WONDER Online Database, released 2012. Data for year 2009 are compiled from the Multiple Cause of Death File 2009, Series 20 No. 20, 2012, Accessed at http://wonder.cdc.gov/ucd-icd10.html on May 29, 2012 1:03:46 PM

7. Referring to **Table 2-9**, calculate the number of deaths per 100,000 population from hypertensive disease that would be expected in California and Louisiana if the death rate in the overall United States were occurring in that state.

Table 2-8 Standard Population, 2009

Age	Projected standard population (year 2000)	Expected # deaths, CA rate	Expected # deaths, LA rate
25–34	37,233,437	446.8	819.1
35–44	44,659,185		
45–54	37,030,152		
55–64	23,961,506		
65–74	18,135,514		
75–84	12,314,793		
85+	4,259,173		
Sum	177,593,760		

Data from Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2009 on CDC WONDER Online Database, released 2012. Data for year 2009 are compiled from the Multiple Cause of Death File 2009, Series 20 No. 20, 2012, Accessed at http://wonder.cdc.gov/ucd-icd10.html on May 29, 2012 1:03:46 PM

Table 2-9 Indirect Standardization of Mortality from Hypertensive Disease in California and Louisiana, 2009

		California			Louisiana		
Age	U.S. hypertensive death rate/ 100,000 (2009)	Population	Obs # deaths	Expected # deaths	Population	Obs # deaths	Expected # deaths
25–34	1.3	5,385,409	64	70.0	624,512	14	
35–44	4.7	5,218,771	202		561,813	45	
45–54	12.7	5,199,654	630		642,810	144	
55–64	23.3	3,838,067	916		507,753	221	
65–74	39.7	2,179,099	930		300,183	197	
75–84	104.1	1,366,454	1,740		183,822	287	
85+	416.4	602,502	3,419		70,291	382	292.7
Sum	20.1	23,789,956	7,901		2,891,184	1,290	

Data from Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2009 on CDC WONDER Online Database, released 2012. Data for year 2009 are compiled from the Multiple Cause of Death File 2009, Series 20 No. 20, 2012, Accessed at http://wonder.cdc.gov/ucd-icd10.html on May 29, 2012 1:03:46 PM

8. Calculate the standardized mortality ratio from hypertensive disease in 2009 for California and Louisiana. How do the states compare to the overall U.S. mortality rate? To each other?

California

$$\frac{\sum \text{Obs}}{\sum \text{Exp}} =$$

Louisiana

$$\frac{\sum \text{Obs}}{\sum \text{Exp}} =$$

9. Go to the CDC Diabetes Data and Trends web page (currently available at http://apps.nccd.cdc.gov/DDTSTRS/default.aspx), and investigate the prevalence and incidence in a given year for a particular subpopulation (e.g., women 18–44 years of age). How are they different? Can you estimate the duration of the disease from these data?
10. Complete the following Kaplan-Meier table (Table 2-10). The outcome of interest is death related to a new treatment for those with stage III emphysema. Calculate the survival function, S_i.

Table 2-10 Kaplan-Meier Approach to Calculating Incidence

i (years)	n_i	d_i	**Conditional Pr(death)** $q_i = \frac{d_i}{n_i}$	**Conditional Pr(survived)** $p_i = 1 - q_i$	**Cumulative Pr(survival)** S_i
1	500	2	0.004	0.996	0.996
4	492	10			0.976
7	488	13		0.973	
9	470	9			
11	472	5			
14	455	25			
15	438	3			0.865

11. What assumptions are required to use this approach? Can you graph your data?