© Mad Dog/Shutterstock

CHAPTER 1

Introduction to Healthcare Statistics

Far and away the best prize that life has to offer is the chance to work hard at work worth doing.

—Theodore Roosevelt

CHAPTER OUTLINE

LEARNING OUTCOMES

After completing this chapter, you should be able to do the following:

1. Define and describe the history of statistics.
2. Describe how statistics are used in health care.
3. Identify common users of healthcare statistics.
4. Define data mining.
5. Describe the history of data mining and how it is used in health care.
6. Explain what a dataset is and how it is used.
7. Identify the four types of data.
8. Discuss five basic elements of data mining.

All Microsoft screenshots used with permission from Microsoft.

KEY TERMS

Aspect
Data
Data mining
Data-driven
Dataset
Decision model
Dependent variable
Descriptive model
Evidence-based medicine (EBM)
Flowchart
Independent variable
Interval data
Machine learning
Nominal data
Ordinal data
Parameters
Population
Predictive model
Primary data
Ratio data
Sample
Secondary data
Statistics
Telehealth
Variable
Viewpoint

How Does Your Hospital Rate?

Paul, a 54-year-old man who has been diagnosed with congestive heart failure, is facing heart valve replacement surgery. He and his family would like the best surgeon and hospital available in his area for this major surgery. They are considering three different hospitals. One site for statistical information that Paul and his family might review would be the Health section of the *US News and World Report* website (https://health.usnews.com), specifically the information relating to cardiology. This website ranks the specialist, survival, patient safety, patient volume, and nursing staff.

Consider the following:

1. What statistics should Paul be considering as he makes this decision?
2. How would the hospitals in your area rate?
3. Which one would you pick, and why?
4. Why is it important for a hospital to keep healthcare statistics?

Introduction

Healthcare statistics allow a hospital to assess, improve, and communicate its quality of patient care, develop better policies and procedures for infection control, and achieve many other goals we will be discussing throughout the following chapters. In this chapter, we will introduce to you the history and definition of healthcare statistics and their importance to a hospital. We will also discuss organizations that keep statistics, not just here in the United States, but around the globe, such as the Centers for Disease Control and Prevention (CDC), the World Health Organization (WHO), and other groups. Keeping statistical data allows us to watch for trends in health care and make proactive rather than reactive choices.

Because the word data will be used so often throughout this text and can be used so many different ways in everyday life, it is worth taking a moment to define this term for our purposes. **Data** (singular, *datum*) are units of information, such as measurements, that can be collected and interpreted. They are the commodity in which we deal as healthcare statisticians.

Data mining is another important concept that we will discuss. We will explain its use in the healthcare arena and show you how you can use data mining and benchmarks with your local hospital. We will also cover how to use Excel and R-Project, both powerful data analysis tools, to examine statistical methods.

Let's begin our exploration of healthcare statistics in the United States and around the globe.

History and Rationale of Healthcare Statistics

The history and study of statistics is as much an examination of historical events as it is a study of the probability, logic, and mathematics behind statistical analysis. As the famous mathematics educator Freudenthal and others have noted, learning the history of a topic often aids in the overall understanding of the focus of a student's study (Leen, 1994). It is in this light that this text will discuss statistics in general and statistical analysis in health care.

There are differing views as to the first uses of statistical analysis, and in fact even the word *statistics*, but certainly most would agree that a large majority of the first uses were descriptive in nature, before the term statistics was formally used.

An early example can be found in Mackenzie's book of ancient stories from India. He notes that Rituparna estimated the number of leaves and berries on two branches of a fruit tree and estimated probabilities of dice rolls. With regard to the fruit tree, he estimated the number of leaves and berries on the basis of a twig, which he multiplied by the estimated number of twigs on the two branches. After a night of counting, he found that his estimate was very close to the real number. Most likely the use of two branches provided a way to take the count from each and determine an average, to be used in estimation for the entire tree (Mackenzie, 1913).

In 431 BCE, the author of *The History of the Peloponnesian War*, Thucydides, notes that the origins of probability can be found in the Athenians' evaluation of the height of the wall of Platae. This estimation was done by determining an average size for a brick, counting the number of bricks in an area, and multiplying by the area they were trying to attack to determine the height they would need for ladders to scale the walls. To determine the height, they multiplied the mode (most frequently occurring value of several sampled bricks) by the count of the number of bricks (Thucydides, n.d). As you will later learn, *mode* is the most frequently occurring value in a set of values. Since bricks were not necessarily uniform in size, determining the mode of a standard brick would be important if you were estimating the size of a wall of bricks you needed to scale.

Early accounts of the use of the undefined term *statistics* vary but include the ninth century work "A Manuscript on Deciphering Cryptographic Messages" written by Al-Kindi. It included a thorough account of how to use statistics and frequency analysis to decipher secret encrypted messages. As a formal term to describe the subject of this text, a version of the word first appears as *statistik* by German author Gottfried Achenwall in 1749 in describing data about the state or arithmetic of the state; however, earlier civilizations, such as the Romans, collected state demographic information and other related data earlier than 1749, though not necessarily using the term *statistic* (Johnson & Kotz, 2012). Unrelated to state data, other early recordings of statistical data concerned sailing, temperature, astrology, and related data that were used for predictions. If you are familiar with the *Farmer's Almanac* or similar texts, then you are aware of the direct impact that predictive statistics can have on people and the state.

Today, statistics are used to express everything from temperature and demographic data averages to mortality rates for cardiac surgery in health care and pattern analysis in massive datasets. In a data-driven society, almost every aspect of life in some way has statistical factors associated with it, from your interest rate at the bank to the target heart rate you are trying to meet for a fitness plan. In short, we use statistics to make a great many important decisions, including those relating to finances, work, and, most specifically for the purposes of this text, health care.

DID YOU KNOW? Florence Nightingale was a member of the Royal Statistical Society. As one of the first women to collect statistics on health policy, she led the way for other female statisticians to work in the field. She was also credited with using graphs to present her findings to Queen Victoria in an effort to reform the sanitary conditions in military hospitals, so she was also an early proponent of data visualization methods (Lancaster, 2013).

Now you know!

Definition of Statistics

Before we discuss in detail how statistics are used in health care, identify some of the users of statistics, and examine some of the statistical methods, we should formally define the term statistics as it will apply to our uses. According to Batten (1986), "Statistics is a series of methods to collect, analyze, and interpret masses of numerical data." However, statistics are not typically just an end unto themselves but rather are used to solve real-world problems. Thus, a practical definition of **statistics** might be the collection of data for the purpose of making predictions (inferences) or considerations to answer a question. Along with this definition are many strategies, rules, and procedures that define and formalize our collection and use of data and resultant findings. We will focus on healthcare statistics, but the underlying methods and processes are applicable to other areas, such as research statistics, which is integral to the field of health care.

The Use of Statistics in Health Care

Statistics are frequently used by many healthcare organizations, including hospitals and insurance companies. Such organizations use statistics to aid in making beneficial business decisions based on data they have collected over time. Hospitals collect and summarize data to improve quality of care, analyze cost of patient care, measure utilization of services provided to patients, examine target marketing decisions, and improve potential offerings of services to patients.

For example, a healthcare professional might examine average length of stay at several area hospitals. On finding that one hospital had a considerably longer average length of stay for patients, the hospital administrators might look for some underlying cause. By finding the issue or issues and resolving them, the hospital's quality of service (QOS) would be enhanced.

Hospitals who are accredited are required to retain data from certain areas to maintain their accreditation standards. The main hospital accrediting body is The Joint Commission. As an example of what is required, a hospital might have to keep fetal monitor strips or complete patient records for 7 years on-site, with older data being held off-site in some archived form for long-term storage (perhaps converted to a different medium, such as microfiche, tape, or optical storage).

The federal government is pushing legislation to move forward with the electronic health record. The *Federal Register* contains all the rules and regulations regarding the implementation of the electronic health record. Included in this legislation are antikickback rules for physicians who refer a patient to a lab or other type of facility in which they have ownership. Rules and regulations are also set forth by the Department of Health and Human Services (DHHS) and the Centers for Medicare and Medicaid Services (CMS).

Third-party payers may require hospitals to collect and maintain performance data. Data can be collected and abstracted for these purposes using many different tools and methods. Of critical importance is how the data are analyzed and used. There are some very specific statistical methods that are used to analyze these data for reporting to external and internal consumers of the data, such as third-party payers and marketing data consumers, respectively.

Key Producers and Users of Healthcare Statistics

The Bureau of Labor Statistics, the National Center for Health Statistics (NCHS), the CDC, and CMS are just some of the agencies who produce and maintain healthcare statistics. Vital statistics are also kept in each state and are another source of statistical information, along with census information. These organizations provide and use statistics to improve health care, summarize findings, and examine trends in the United States and around the globe.

The Bureau of Labor Statistics is a unit of the US Department of Labor and serves as the principal agent for the US Federal Statistical System. The primary mission of the Bureau of Labor Statistics is to collect and analyze essential statistical data for use by the public and the US Congress. The most common statistics that are kept by the Bureau of Labor Statistics relate to prices, employment, unemployment, compensation for injuries, and work injuries. For example, according to the US Department of Labor agency, Occupational Safety and Health Administration (OSHA), in 2018 there were 1008 fatalities to workers who performed construction.

The NCHS is the principal agent that delivers statistical information and directs policies and actions that will improve the health of the public. The NCHS is housed within the CDC. The CDC and the NCHS compile statistics for all types of disease for the United States and worldwide. For example, in the United States, rates of the sexually transmitted disease gonorrhea increased by 8% from the year 2010 to 2011, totaling 321,849. By 2017, the number of new gonorrhea cases had climbed all the way to 555,608 (CDC, 2018).

The primary responsibility of CMS is to administer Medicare and Medicaid, the Children's Health Insurance Program, the Health Insurance Portability and Accountability Act (HIPAA), and Clinical Laboratory Improvement Amendments. CMS keeps statistics for each state as to how many people are on Part A or Part B or both for Medicare. As of July 2016, 56.5 million were on only Part A of Medicare, and 52.1 million were on Part B (National Committee to Preserve Social Security and Medicare, 2020). Keeping this type of statistic is vital to health care because people are living longer and will eventually require more healthcare services.

In addition to the organizations already discussed, users of statistical data include the following:

- Federal government agencies gather information that references public health issues such as HIV/AIDS, cancer, births, and deaths.

- Accreditation agencies use statistics to show the most common diagnoses and procedures and the amount of resources used to treat those patients.
- Managed care organizations use statistics to review costs for the level of care that is being provided to their patients.
- Healthcare researchers use the data from health law and regulations, physician practices, and **telehealth** (services that use electronic information and telecommunications technologies to support long-distance clinical health care). Technologies can include videoconferencing, the internet, store-and-forward imaging, streaming media, terrestrial communication, and wireless communication. There are many other types of information used for research.
- Mental health facilities and drug and alcohol facilities use this information to measure the success of the services being provided and success rates of patients.

Note that these parties are external to the healthcare provider (e.g., the hospital). In fact, the hospital would be a primary consumer of this valuable data. Activities such as QOS, as mentioned earlier, are but one reason for this information. Another is governmental oversight via the certificate of need (CON), which is a statement issued by a government agency for projected construction or modification of a healthcare facility. The facility must meet the requirements statistically to meet the CON criteria. It ensures that the new facility will be needed at the time of completion for those additional services. Basically, you might consider that a CON is an assurance that facilities are not built or expanded beyond the requirements of the community.

Hospitals and health information management organizations use statistical information as well. Both of these organizations typically use similar types of information in their statistical analysis. Let's examine the departments that would perform and use the various data and findings.

- Healthcare administrators use health statistics to make data-driven decisions. Think of **data-driven** as making decisions based on statistical information instead of by guessing. For example, if a hospital had data for 3 years on nursing needs of the emergency department on New Year's, you might be able to expand staffing on that day, based on previous years' data. Or if data showed that the number of patients in labor and delivery increased by 10% each year over a span of 4 years, the facility might consider adding more beds and staff to that area.
- Healthcare department managers use statistics to set goals for the department, such as annual budgets.
- Cancer registries use statistical information regarding the different types of cancer, stages, and treatment. They also maintain survival rates of cancer patients. Cancer registries can receive accreditation by the American College of Surgeons (ACS). The cancer registry must meet the standards set by the ACS to be accredited.
- Nursing facilities use statistical information to review the different types of payers of insurance their patients have.
- Home healthcare organizations keep statistical data to track patients and their outcomes. The information includes the following: the number of visits, dressing changes, oxygen machine use, and many other services that home health organizations make. Further data include how many patients are taking their medication as directed, how many are improving, and how many are being readmitted.
- Hospice provides services either in the home or in a healthcare facility. The services they provide are linked to the patient's diagnosis.

Data Mining

Now that we have learned about some consumers of statistical data, we should examine sources for these data and how they are collected.

Definition

The process of extracting information from a large set of data is known as **data mining**. Tan, Steinbach, and Kumar (2003) note that data mining is a "confluence of many disciplines" and show an overlap of statistics, data mining, and artificial intelligence, machine learning, and pattern recognition. As a process, data mining involves the steps of defining, finding, and extracting data or knowledge that is buried in large sets of raw data. In this process, "data is retrieved, consolidated, managed and prepared for analysis" (Valova & Noirhomme, 2009). Following data mining, the resultant data are analyzed and then organized into a usable format.

History

To better understand this concept, let's cover some of the history of data mining while setting some boundaries for how we will obtain data, review some processes and strategies to make sense of mined data, and integrate these data and methods into software analysis tools.

Data mining to make healthcare medical treatment decisions is not new, although the use of formal computerized data mining tools is. In fact, data mining and **evidence-based medicine (EBM)**—medical practice that is based on the best available current methods of diagnosis and treatment, as revealed in research—have existed since the time of Hippocrates (460 BCE) in ancient Greece. Other notable "data miners" in history include Aulus Cornelius Celsus of ancient Rome, and John Snow, the father of modern epidemiology. Celsus wrote that wound cleansing and hygiene were important in health care, though these practices did not take hold until the late 1800s. John Snow tracked via maps the source of cholera in 1854. Thus, we see that statistics has a role in not only medical administration, but treatment as well.

In the l960s, when the computer age was just starting to become commercialized and heavily adopted by the business community, data were collected on magnetic tapes, punch cards, and disks—media that offered considerably less storage capacity than today's options. Data mining actually took a major step in revolution in the l980s when relational databases and structured query languages (SQLs) were developed. By using the data sorting and retrieval power afforded by structured query language, hospitals were better able to analyze and make sense of large sets of data.

Data warehousing, the centralized storage of large sets of data, was introduced in the l990s. It supported online analytic processing and multidimensional databases, which helped it to grow rapidly. In today's world of big business, hospitals and other large corporations use collected information to make large-scale assessments, such as predicted growth over the next 5 years or total revenue over the last 3 years.

Three different areas have provided the growth to make data mining what it is today: statistics, artificial intelligence, and machine learning. Statistics has enabled organizations like the CDC to provide better health care and services to patients, enabling its top 10 achievements in the 20th century: improved vaccinations, improved motor vehicle safety, safer and healthier foods, better control of infectious diseases, improved workplace safety, reduced deaths from heart disease and stroke, better family planning, increased awareness of tobacco as a health hazard, fluoridation in drinking water, and healthier mothers and babies. Statistics play a vital role in keeping the general population healthy. With the advent of computers and the internet, we can now harvest and refine medical decisions to an even finer degree, to the benefit of patients and medical establishments.

Current Applications

Today, data mining or predictive analysis in health care is a growing field. In fact, it is being used more and more to not only predict trends and analyze findings, but as an important tool for medical establishments to improve patient care, improve service offerings, and decrease losses, not the least of which is lost revenue in patient billing. Data mining has long been used in other fields but is still an emerging area within the healthcare industry. Canlas notes that data mining tools are being used not only for e-business and marketing but also by healthcare providers for "analysis of health care centers for better health policy-making, detection of disease outbreaks and preventable hospital deaths, and detection of fraudulent insurance claims" (Canlas, 2009). Moreover, the recent changes in health care, billing, and administration in the United States offer great opportunities for data mining.

Another current trend in data mining is **machine learning**. In addition to predictive analysis for the future, this process can be used to analyze past historical data. Therefore, data mining is both a process and an analysis method. It is a process, as there are procedures for collecting and preparing the raw data. The appropriate analysis method is chosen based on the type of data needed and the desired results. You will examine data mining in more detail in this chapter as a formal (and modern) method. Lastly you will examine some statistical terms and processes and examine how to handle some common statistical formulas using computer applications.

How Does Your Hospital Rate?

Now that Paul has found a website, he can compare facilities. The hospital Paul should review is Hospital A because it has a score of 100 out of 100, but what if that hospital is located in Ohio and Paul lives in Maryland, which is quite a distance to travel? In that case, Paul should look at Hospital B, which is nearby and has a score of 76.6 out of 100.

After choosing to review Hospital B, he finds that the reputation with the specialist is only 23.3%, survival rate is better than expected, the safety rating for patients is moderate, there is a high volume of cardiac patients, it has the highest rating for magnet nursing recognition, and, most important, it is rated seven out of seven for advanced technologies and key patient services, including an advanced trauma center and intensivist staff, meaning they have a staff physician in the intensive care unit at all times. This information has hopefully answered some of Paul's questions about the physicians and quality of care for cardiac patients.

Consider the following:

1. What other information should Paul consider before making his decision?
2. Which facility would you choose if you were Paul?

Basic Statistical Concepts

Although a detailed discussion of statistics goes beyond the scope of this text, the next sections will introduce some of the major statistical concepts relevant to healthcare professionals. Let's start with some basic items applicable to most statistical measures.

Dataset

A **dataset** is the data collected on a subject under examination. When your dataset includes information on every member of the group being investigated, you are examining a **population**. Or you may have a **sample** of data, which is a subset of data that statistically represents the entire population, ideally. A capital letter *N* is used to represent a population size and a lowercase *n* refers to a sample size. These concepts are important and ones we will refer to throughout the text, so it is important to note them.

As an example of the previous definitions, consider the Nielsen media rating group. Suppose they survey a sample of 1 out of 1,000 households, asking them which television shows they like or dislike. Considering how many households there are in the United States, even using this seemingly small sample size will yield thousands of survey results. Similarly, in the healthcare setting, collecting data from 50 randomly selected hospitals from across the country might be a fair and manageable representation of the greater population, which consists of every hospital in the United States. Imagine trying to survey all the hospitals in the United States! Obviously, it would be far easier to sort through and make inferences or summaries about the data when working with a sample rather than an entire population.

Typically, data mining involves very large samples or sets of data, or "big data," as you may hear it referred to. There are significant qualitative differences between the conclusions you can draw when examining a population versus a sample. When examining an entire population, you are not dealing with estimates or probability of an outcome but actual facts about an outcome. For example, consider a case in which out of 100 people surveyed, only 10 had a terminal disease. You know in this case that 10 of the 100 have a terminal disease—not just 10% but 10 real people. In this case, you are dealing with a population and not statistical or inference data. You have all the data and know all of the variables, so there is no need to estimate or infer. On the other hand, if the 100 people surveyed are only a random sample of a larger population rather than the entire population itself, you could generalize that the 10% who reported having a terminal disease represent the entire population. This, however, would be an inference only, not the statement of a fact.

Moving along, **parameters** are descriptive measures of a population and are sometimes referred to as fixed references (Batten, 1986). The use of parameters is different from that of statistical data, in which you are making assumptions about a larger population based on a random sampling of the population. An example of a parameter could be that we have 10,000 people under study in a population. We know we are examining 10,000 people. For example, imagine that you have 10,000 people to make a generalization about. In contrast, using statistical data methods (strictly speaking), you would randomly gather information from, say, 500 of the 10,000 and assume that the other 9,500 would answer the same way to the questions asked. As you can infer, having all the data is typically better than relying only on a sample of data.

However, in statistics, you typically do not have all the data, so you must collect data from many similar sources, which hopefully can lead to a valid finding for the unique situation at hand. In other words, we have data on the basis of which we can generalize findings that *should* be applicable to other groups.

For example, of 10,000 people surveyed, we found that 99% believed that hospital stays following major surgery, as covered by third-party payers, were not long enough. That being determined, we could infer that most others in the population would be closely aligned to the 99% mark, given that we could survey them as well. This of course would still be an assumption as to the outcome, but because the percentage of people who responded in this manner is so high (99%), it is probably true. For example, the data from five hospitals in our state of North Carolina seem to show that if patients are given information on a certain lifestyle change (such as hand washing) on their first visit, their rate of secondary infection is lowered by over 50%.

Variables

A **variable** is a characteristic or property of something that may take on different values. For example, the number of patients admitted to a hospital between the hours of 12 am (midnight) and 7 am might be tabulated and examined over a week's time frame. Each day's data would be a variable. An **independent variable** (also known as an experimental or predictor variable) is a factor we can measure, manipulate, or control for to produce a change in another variable, which is known as the **dependent variable** or outcome variable. For example, imagine that you would like to determine whether using twice the amount of a certain drug for a disease will help patients recover more quickly. In this example, the amount of the drug administered is the independent variable, and the recovery rate of the patients is the dependent variable.

Data Distribution

Data distribution refers to the characteristic pattern that data assume when represented in graphical form. A normal distribution of data is one that is characterized by data that average around a central value with no real tendency to skew left or right. In this case, 50% of data falls to the left of the peak of the curve, and 50% falls to the right, forming a symmetric curve that possesses a single peak when graphed. Because the curve is shaped like a bell, you will hear it referred to as a "bell curve" (**Figure 1.1**). This data distribution pattern is one of the most important statistical concepts to understand.

However, if the distribution is not normal, data could be spread out to the left or to the right, or it could be randomly distributed.

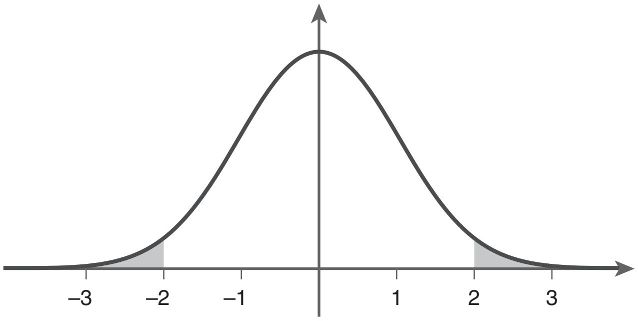

Figure 1.1 Bell curve.

Types of Data

The four types of data you might encounter are nominal (categorical), ordinal, interval, and ratio.

1. **Nominal data**: This type of data indicates categories and cannot be ordered. Examples include the following: techniques (technique A, technique B), gender (male or female), or occupation (e.g., students, professional programmers).
2. **Ordinal data**: This type of data can be ordered (e.g., in terms of size), but the difference between any of the two values may not always be equal. Examples include responses to a Likert-scale question, such as the following: *use it every day*, *use it once a week*, *use it once a month*, *use it once a year*, and *have never used it*. Likert-scale questions will be discussed in more detail in a later chapter.
3. **Interval data**: This type of data can be ordered, and the difference between any two consecutive values is the same, but there is no absolute zero, which allows you to have meaningful negative values. The most famous example of this type of data is temperature in degrees Celsius (°C) or Fahrenheit (°F). The zero values (i.e., 0°C and 0°F) are artificially defined, and negative temperature values are possible. But the difference between any two consecutive values at any point on the scale (i.e., between 0°C and 1°C and between 100°C and 101°C) is the same.

   Note that when working with data from Likert-scale questions, if you can assume that the differences between any two options are equal, you can treat them as interval data. For instance, if your options are *strongly agree, agree, neutral, disagree*, and *strongly disagree*, you may be able to treat them as interval data, which is handy for computer tabulation. You simply assign a numeric value for each selection (e.g., strongly agree = 1, agree = 2) then count the occurrences of each from the dataset, such as "out of 50 people sampled, 5 chose that they strongly agreed with the first question, 15 reported that they only agreed," and so on.

4. **Ratio data**: This type of data can be ordered, the distance between any two consecutive values is the same (interval data), and there is an absolute zero. This means that a meaningful negative value of interval data does not exist (in statistics). Consider as examples the weight of a llama, the height of a tree, the length of something, or a recorded time or speed. A count could be considered as a ratio, as well.

Different statistical methods are designed only for certain types of data. As you work through this text, you should pay particular attention to this fact, as well as the size of the sample you are working with. Both choice of method and sample size are critical to choosing an appropriate statistical measure.

Types of Data Mining Models

Because this text offers the opportunity to mine real data for use in statistics, it will be helpful to discuss some key data mining models in detail here, including predictive, descriptive, and decision models. Subsequent chapters will give you hands-on experience with data mining methods.

Predictive Models

A **predictive model** explicitly predicts future behavior based on past trends. In each case, a model is created or chosen to try to best predict the probability of an outcome. When a predictive model is used in data mining, its main purpose is to forecast probabilities and trends in the future. Data are collected for pertinent predictors. The model is formulated, and finally it is validated and revised as new data become available.

For example, predictive models can be used by insurance companies and government programs such as Medicaid to assist in the prediction of future medical needs. A predictive model can also identify those who may be at high risk for developing a chronic disease or having poor health outcomes. However, initial medical information (data) on these patients is needed to make such predictions. Without their personal health information, identification of their risks would be significantly delayed. This could potentially lead to unfavorable outcomes or a delayed improvement in the patient services offered.

Descriptive Models

A **descriptive model** describes patterns in existing data, including the main features or a summary of the data under examination by the researcher. It provides a hypothetical basis for the system under study. Descriptive models can recognize relationships while being used with other models to make predictions. In this approach, maps, charts, and graphics portray and promote understanding of real-world, complex, and sometimes redundant systems or services. Think of it as a way to describe visually the data in question. There are two general dimensions that are used for this description of mined data, viewpoints and aspects.

A **viewpoint** is a specific context or approach that you intentionally adopt when examining the data that allows you to focus on relevant details in the study and ignore irrelevant data.

An **aspect** is a specific category of data that is used in conjunction with your viewpoint. You might have one or more aspects associated with each viewpoint. Think of a viewpoint as a book chapter, with aspects being headings within the chapter. For example, a viewpoint might be data on patients being treated for lung cancer, such as survival rates and length of survival times. After patients who are cured are removed from consideration, aspects would include data on patients who survived 1 to 6 months, 7 to 12 months, and 1 to 2 years.

Decision Models

A **decision model**, also known as a business rule or business logic, is a logic system to determine desired actions for the business based on thresholds, conditions, or events. When using decision models, all elements in a relationship are examined to forecast or predict results. Units are arranged into groups according to the relationships between the units. All information would be organized into a logic **flowchart**, which is a graphical way to depict a series of actions, given a certain series of events. For example, a "Get up and go to school" flowchart would include events in the order in which they should occur, such as "wake up and turn off the alarm," "take a shower," "dry off," "eat breakfast," etc. Note the importance of the sequence of events in a flowchart. If you took a shower after you dried off, you might be a bit wet at the breakfast table!

As another example, consider diagramming a process to detail patient admittance to a medical facility. Each step would be graphically presented, with details and substeps, as a part of the flowchart. The first step might be to determine whether an arriving patient is an emergency case. This would be a logical true/false decision, with one of two paths being taken, depending on the case. The flowchart then would present the alternate steps for true and for false responses in a graphical fashion.

With decision models, the relationships among all the known data of a situation and the result of the logical decision process can be used to forecast future patterns and thus better prepare the organization with regard to planning. Thus, this model can help optimize and streamline the business and help the business to provide better end service and maximize cost savings. In later chapters we will examine some specific formulas and tools used for flowcharting and decision logic, such as for forecasting. Although there are many ways to examine data, this text will generally be limited to models most appropriate for the healthcare facilities.

Obtaining Data

If you are conducting an exhaustive search for information from many sources, use of data mining and statistical inference might be the best way to proceed. Regarding how to refer to your data, if a dataset was collected by other people, it is considered a **secondary data** source; if you collected the dataset yourself, it is a **primary data** source.

Related to obtaining data, there are five basic steps in data mining:

1. Extract and transform the data and load the data into the data warehouse system.
2. Store and manage data using a relational database system.
3. Provide data access to users.
4. Use application software to analyze the data.
5. Present the gathered information in a meaningful manner.

Keep in mind that if you are examining all of the data available, say from a local source such as the hospital where you work, then you are examining a population (capital *N*) and are using local data, otherwise known as a primary data source. Regardless of where you get your data, you should cite your sources, along with what measures and tools you used. Your results might be suspect if you do not validate where your information came from, so give as many details as you can.

Some may assume that secondary data are inferior to primary data, but make no mistake, using mined secondary data in health care will improve patient care and increase benefits in the future. As paper medical records are replaced with electronic ones, mined secondary data will increase significantly, allowing for quality improvement in patient care and increases in cost savings, patient satisfaction, and revenue. Thus, a leading goal for healthcare facilities at this time is increased use of mined secondary data internally in their facilities, especially on identified areas that are in need of quality improvement.

How Does Your Hospital Rate?

Paul and all Americans want high-quality health care by the best professionals. However, we certainly cannot afford to travel to other states to receive treatment from the top-rated hospitals unless we are rich. When shopping for a medical care facility, consider that you might want the following: timely service, a safe environment, current medical technologies and treatments, and reliable patient-centered care. As Paul found out, the Hospital B has a high score for the use of technology. However, their moderate safety score concerns him; no matter how great the care may be in other areas, receiving a staph infection while in the hospital would hinder his journey back to full health.

To facilitate the sort of rating process that Paul undertook, the US federal government, via Medicare.gov, has created Hospital Compare, a website that allows people to compare area healthcare facilities based on zip code or by facility name.

Consider the following:

1. Use the following website: https://www.medicare.gov/hospitalcompare/search.html. How does your hospital rate for cardiac procedures?
2. Would you have a cardiac procedure performed in your area facility, or would you travel to another facility?
3. What do you think Paul should do?

Global Perspective

Taking our healthcare statistics globally allows us to compare diseases of other countries to those in the United States. In this Global Perspective section, we will examine rheumatic heart disease. This disease is caused by rheumatic fever, which affects the mitral valve in the heart. If a child has strep throat or scarlet fever that is not treated properly with antibiotics, he or she will have a higher risk of acquiring rheumatic heart disease. The mitral valve between the left atrium and left ventricle is affected by this condition. This condition affects children mainly between the ages of 5 and 15 years. Surgery to repair the mitral valve is one way to treat this condition.

The data below were retrieved from WHO. **Table 1.1** shows data for rheumatic heart disease in the United States and India, comparing males and females between the ages of 0 and 14 years in 2008. The population of males followed in India was 195,436, and the population of females was 179,323. The population of males in the United States was 32,646, and the population of females was 31,056.

On an interesting note, further research could be done to understand why India has a 2.3% rate of rheumatic heart disease in males and 3.0% rate in females. Questions that could be asked might include the following:

- Why is the female rate of rheumatic heart disease higher than that of males?
- Do children in India receive the antibiotic at all?
- Are incorrect dosages of the medication given?
- Are children not able to get the surgery necessary to repair the valve?
- Why do children in India acquire strep throat or rheumatic fever?

Table 1.1 Incidence of Rheumatic Heart Disease in India and the United States in 2008

Country	Sex	Population with Rheumatic Heart Disease
India	Male	2.3%
	Female	3.0%
United States	Male	0%
	Female	0%

Many other questions could be asked, and further research could be done with this information to follow up with types of treatment given to children in India compared with those in the United States.

This comparison of a disease's prevalence in the United States versus India is just a small sample of some of the interesting global trends and statistics that we will cover throughout subsequent chapters.

Hands-on Statistics 1.1: Examine Other Data from WHO

1. Using the same data source (WHO) as referenced in the Global Perspective section, compare India with a different country and discuss the results of your comparison. See http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html.
2. What are some strengths and weaknesses of this data source?

Chapter Summary

This chapter established a foundation for learning more about healthcare statistics and examining sources of data. It also examined some parts of the process involved in data analysis. The study of statistics is complex, involving extensive mathematical algorithms and research considerations, but rest assured that subsequent chapters will further explain and provide real-world examples of how to use the statistical methods involved in healthcare statistics. When you finish this text, you will be confident and well versed in the subject and will be able to put your knowledge to immediate practical use.

Also covered in this chapter is an introduction to data analysis and presentation of your findings. Hopefully, this chapter has whet your appetite for these subjects. In coming chapters, you will learn more topics related to statistics, data harvesting, data analysis, and presentation of your findings, all emphasizing real-world data. This chapter is only a small step toward an exciting journey to find out how powerful statistics can be.

Apply Your Knowledge

1. Consider users of healthcare statistics other than those mentioned in the chapter. List at least three users of healthcare statistics, and describe the statistics they keep.
2. Other than Florence Nightingale, who were some of the first users of statistics, and what statistics did they keep?
3. Name at least two countries that use telehealth.
4. Write a pro and con for some health issue, such as "smoking increases your risk of lung disease" or "running on pavement is bad for your knees." Be creative!
5. Discussion: When researching *data mining*, you may come across the terms *data dredging* or *data snooping*. What do they mean? What implications should a researcher know about these terms? How might they be avoided?
6. Go to the following website: http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html. Review data for males and females in India and in the United States, comparing the percentages between the different age groups of those with rheumatic heart disease: 15 to 50 years and 60+ years. Write a brief explanation of your findings.
7. Go to http://www.cdc.gov/DataStatistics/. Click on a topic of your choice and compare health statistics for your state with those of another state. Did your state have a higher statistic than your comparison?

References

Batten, J. (1986). *Research in education* (rev. ed.). Greenville, NC: Morgan Printers.

Canlas, R., Jr. (2009). *Data mining in healthcare: Current applications and issues* (Master's thesis). Adelaide, South Australia: Carnegie Mellon University in Australia.

Centers for Disease Control and Prevention. (2018). Table 10. Selected nationally notifiable disease rates and number of new cases: United States, selected years 1950–2017. Retrieved from https://www.cdc.gov/nchs/data/hus/2018/010.pdf

Johnson, N., & Kotz, S. (2012). *Leading personalities in statistical sciences: From the seventeenth century to the present*. Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118150719.ch2/summary.

Lancaster, L. (2013). Celebrating statisticians: Florence Nightingale. JMP Blog. Retrieved from https://community.jmp.com/t5/JMP-Blog/Celebrating-statisticians-Florence-Nightingale/ba-p/30247

Leen, S. (1994). The legacy of Hans Freudenthal. *Educational Studies in Mathematics, 25*(1-2), 164.

Mackenzie, D. (1913). Indian myth and legend. Sacred Texts. Retrieved from http://www.sacred-texts.com/hin/iml/index.htm

National Committee to Preserve Social Security and Medicare. (n.d.). Medicare. Retrieved February 21, 2020, from https://www.ncpssm.org/our-issues/medicare/medicare-fast-facts/

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Pearson Education.

Thucydides. (n.d.). *The history of the Peloponnesian War* (R. Crawley, Trans.). Project Gutenberg. Retrieved from http://www.gutenberg.org/files/7142/7142-h/7142-h.htm

Valova, I., & Noirhomme, M. (2009). Comparative analysis of advanced technologies for processing of large data sets. *Information Technologies and Control*, 1-13.

Web Links

Using Excel to do Basic Statistical Analysis: https://www.statisticshowto.datasciencecentral.com/mode/#excel

Excel Tutorials for Statistical Data Analysis: http://www.stattutorials.com/EXCEL/EXCEL_TTEST2.html

Introductory Statistics: Concepts, Models, and Applications: http://www.psychstat.missouristate.edu/introbook/sbk25m.htm

MS Excel: How to Use the QUARTILE Function (WS): http://www.techonthenet.com/excel/formulas/quartile.php

Excel and Quartiles: http://www.meadinkent.co.uk/excel-quartiles.htm

Drawing a Normal Curve: http://www.tushar-mehta.com/excel/charts/normal_distribution/

Sexually Transmissible Infections: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4102.0Main+Features10Jun+2012

Sexually Transmitted Diseases (STDs): Data and Statistics: http://www.cdc.gov/std/stats11/trends-2011.pdf

Medicare Enrollment—Aged Beneficiaries: As of July 2010: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareEnrpts/Downloads/10Aged.pdf

Has Statistics Made Us Healthier? The Role of Statistics in Public Health: http://www.statisticsviews.com/details/feature/5025891/Has-statistics-made-us-healthier-The-role-of-statistics-in-public-health.html

Descriptive Statistics: http://www.businessdictionary.com/definition/descriptive-statistics.html

Discrete/Continuous: http://www.chegg.com/homework-help/definitions/discrete-continuous-31

Hospital Compare: https://www.medicare.gov/hospitalcompare/search.html