CHAPTER

1

INTRODUCTION

Chapter at a Glance

This book introduces health administrators, nurses, physician assistants, medical students, and data scientists to statistical analysis of electronic health records (EHRs). The future of medicine depends on understanding patterns in EHRs. This book shows how to use EHRs for precision and predictive medicine. This chapter introduces why a new book on statistical analysis is needed and how healthcare managers, analysts, and practitioners can benefit from fresh educational tools in this area.

Why a New Book on Statistics?

This textbook provides a radically different alternative to books on statistical analysis. It de-emphasizes hypothesis testing. Its focuses primarily on removing confounding in EHRs. It emphasizes data obtained from EHRs and thus, by necessity, involves a great deal of structured query language (SQL).

The management and practice of healthcare are undergoing revolutionary changes (McAfee and Brynjolfsson 2012). More information is available than ever before, both inside and outside of organizations. Massive databases, often referred to as *big data*, are available and accessible. These data can inform management and practitioners' decisions. The growing use of EHRs has enabled healthcare organizations, especially hospitals and insurance companies, to access large data sets. Inside organizations, EHRs can measure countless operational and clinical metrics that enhance the organization's productivity.

All sorts of data points are available for scrutiny. Analysts can track who is doing what and who is achieving which outcomes. Providers can be benchmarked; front desk staff efficiency can be monitored. Data are available on the true cost of operations, as nearly every activity is tracked. Contracts with health maintenance organizations can be negotiated with real data on cost of services. Data are available on profitability of different operations, so unprofitable care can be discontinued. Managers can detect unusual patterns in the data. For example, they can see that hospital occupancy affects emergency department backup.

In the healthcare field, data are available on pharmaceutical costs and their relationship to various outcomes. Many organizations have lists of medications on their formulary, and now such lists can be based on both cost and outcome data. Medications can be prescribed with more precision and less waste. Data can be used to predict future illnesses; diseases can be prevented before they occur. The wide availability of massive amounts of data has made managing with numbers easier and more insightful. The following are some examples of how healthcare organizations are gathering massive databases to enable insights into best practices (Jaret 2013):

1. The Personalized Medicine Institute at Moffitt Cancer Center tracks more than 90,000 patients at 18 different sites around the country.
2. In any given year, the Veterans Affairs Informatics and Computing Infrastructure (VINCI) collects data on more than 6 million veterans across 153 medical centers.
3. Kaiser Permanente has a database of 9 million patients.
4. Aurora Health Care system has 1.2 million patients in its data systems.
5. The University of California's medical centers and hospitals have a database with more than 11 million patients.
6. The US Food and Drug Administration agency has the combined medical records of more than 100 million individuals to track the postlaunch effectiveness of medications.
7. The Agency for Healthcare Research and Quality has compiled claims data across 50 states.
8. The Centers for Medicare & Medicaid Services releases 5 percent samples of its massive data.

In addition to planned efforts to collect information, data gather on their own on the web. Patients' preferences, organization market share, and competitive advantages can all be determined from analysis of internet comments (Alemi et al. 2012). The internet of things collects massive data on consumers' behavior. Most web data are in text format. Analysis of these data requires text processing, a growing analytical field.

Big data is influencing which managers succeed and which will not. "As the tools and philosophies of big data spread, they will change the long-lasting ideas about the practice of management" (Eshkenazi 2012). Companies that get insights through analysis of big data are expected to do better than those that do not, and therefore these managers will succeed more often. There are many examples of how data-driven companies succeed over counterparts that ignore data analysis. At Mercy Hospital in Iowa City, Iowa,

managers who benchmark their clinicians and pay them for performance report 6.6 percent improvements in the quality of care (Izakovic 2007).

Many investigators point out that the Veterans Health Administration (VHA) was able to reinvent itself because it focused on measurement of performance (Longman 2010). The VHA healthcare system had poor quality of care—until the VHA became data driven. Then, over a short interval, VHA managers and clinicians were able to not only change the culture but change patient outcomes. According to Longman (2010), the VHA system now reports some of the best outcomes for patients anywhere in United States.

A recent study of 330 North American companies showed widespread positive attitudes toward data evaluation. The more companies characterized themselves as data driven, the more they were likely to outperform their competitors financially and operationally. Data-driven companies were 5 percent more productive and 6 percent more profitable than less data-driven companies (Brynjolfsson, Hitt, and Kim 2011).

In healthcare, companies that rely heavily on Lean (a process improvement tool) and other similar tools can be classified as data driven, even if they rely on small data sets. These companies use statistical process control to verify that changes have led to improvements. Many studies show that when organizations fully implement statistical process control tools, including an emphasis on measurement (Nelson et al. 2000), they deliver better care at lower cost (Shortell, Bennett, and Byck 1998). The use of these techniques is widespread, making it an essential capability of modern managers (Vest and Gamm 2009).

In healthcare, the use of EHRs has been associated with reductions in medication errors (Stürzlinger et al. 2009). Managers have used EHRs to maximize reimbursement in ways that have surprised insurers (Abelson, Creswell, and Palmer 2012). Other managers report analyzing data in EHRs to reduce "never events" (unreimbursable accidents) in their facilities and to measure quality of care (Glaser and Hess 2011). These efforts show that analysts are finding ways to use the data in EHRs to improve their organizations. Such efforts are expected to continue, creating an unprecedented shift toward the heavy use of data.

Big data has changed and continues to change health insurance. Insurance companies are trimming their networks using data on the performance of their doctors. New start-up insurance companies are competing more effectively with well-established insurance companies by situating their secondary providers near their target market. Insurance companies are deciding what to cover and what to discourage through data analysis. Risk assessment is changing, and more accurate models are reducing the risk of insurance. In risk rating, chronological age may not be as important as history of illness.

Value-based payment systems have transformed who assumes risk. Value-based reimbursement has changed how hospitals and clinics are paid. With this paradigm shift, insurers hold hospital managers accountable for quality of care inside and outside of hospitals. For example, a hospital that does a hip replacement is paid a fixed amount of money for expenses, including the cost of surgery and out-of-hospital costs 90 days after surgery. The hospital manager needs to make sure not only that the healthcare organization's surgeons are effective and that its operation does not lead to unnecessary long stays, but also that patients are discharged to nursing homes or other institutions that actively work on the patients' recovery. Affiliation with a home health care organization or nursing home could help decrease readmission and could easily reduce the hospital's payments. For 90 days, no matter where the patient is cared for, the hospital manager is at risk for cost overruns. Value-based reimbursements have increased the need to analyze data and affiliate with providers and institutions that are cost-effective.

Big data is changing clinical practice as well. The availability of data has enabled managers and insurers to go beyond traditional roles and address clinical questions. For the first time, analysts can measure the comparative effectiveness of different healthcare interventions. They can talk to physicians, nurse practitioners, and physician assistants about their clinical practices. They can discourage patients from undergoing unnecessary operations. For years, clinical decisions were made by clinicians, but the availability of data is beginning to change this. For example, the Centers for Disease Control and Prevention uses Data to Care (D2C) procedures to identify HIV patients who have stopped taking their medications. Careful communication with these patients can bring them back to care. In addition, payers such as Amazon are organizing population-level interventions to improve delivery of care. Analysts are alerting primary care providers about potential substance abuse and alerting patients about the need for flu shots. These efforts are giving extended clinical roles to data analysts.

Data are changing the healthcare equation. Today, managers have data on what is best for patients, and they can work with their clinicians to change practices. For example, analysts have been able to examine pairs of drugs that cause a side effect not associated with the use of either drug on its own. They found that Paxil, a widely used antidepressant, and Pravastatin, a cholesterol-lowering drug, raise patients' blood sugar level when used together (Tatonetti et al. 2012). In this example, and other comparative effectiveness studies, we see an emerging new role for data scientists.

Content of Chapters

Statistical Analysis of Electronic Health Records differs from existing introductory statistics books in many ways. Exhibit 1.1 lists how this textbook's emphasis differs from that of other managerial statistics books. First, it exclusively focuses on the application of statistics to EHRs. All examples in this book come from healthcare. They include use of statistics for healthcare marketing, cost accounting, strategic management, personnel selection, pay-for-performance, value-based payment systems, insurance contracting, and clinician benchmarking. These examples are given to illustrate the importance of quantitative analysis to management of healthcare.

Second, the book de-emphasizes traditional hypothesis testing and emphasizes statistical process control. For healthcare managers, hypothesis testing is of little use; such testing requires the use of static populations and context-free tests that simply do not exist in the real world. In contrast, healthcare managers have to examine their hypotheses over time and thus need to rely on statistical process control. Alternately, they need to test a hypothesis while controlling for other conditions and must therefore rely on multivariate analysis as opposed to univariate hypothesis tests.

Most existing books focus on hypothesis testing through confidence intervals and standardized normal distributions. *Statistical Analysis of Electronic Health Records* introduces these concepts through statistical process control. Confidence intervals are discussed in terms of 95 percent upper and lower control limits in control charts. The use of geometric distributions in time-between control charts is discussed. This book covers the use of Bernoulli and binomial distributions in creating probability control charts. It discusses the use of normal distributions in creating X-bar control charts and provides students with knowledge of hypothesis testing in the context of observational data collected over time.

Third, this book differs from most other introductory statistics textbooks in that it mostly relies on EHR-based data. Healthcare is swimming in data. Data analysts need to structure and delete large amounts of data before they can address a specific problem. EHR data are observational, not experimental. Managers rarely have the option to run randomized experiments. Because data come from operational EHRs, where data are collected from patients who voluntarily participate in various treatments, a number of steps must be taken to remove confounding in data. In jest, analysts call these steps "torturing data until they confess."

In EHRs, data are available in numerous small tables, and not in one large matrix, as most statistical books require. This book gives considerable attention to how data from different tables should be merged. Throughout the book, I have relied on SQL to make the manipulation of data easier.

Because the data are inside EHRs, SQL is required to manage the data—other statistical packages are just not available for EHRs. Statistical analysis is really just the tip of the iceberg; much more work and time go into preparing the data than into analyzing them. *Statistical Analysis of Electronic Health Records* pays special attention to preparation of the data using SQL.

In comprehensive EHRs, data are available on patients from birth until death. To use these data, we need to understand their time frame. Several statistical methods have been designed based on the sequenced order of events. EHR data enable new methods of analysis not otherwise available.

Data are collected passively as events occur. Over time, more data are available, and one major task of the manager is to decide which data are relevant. The data themselves never stop flowing, and the manager must decide which period he would like to examine and why. EHRs are also full of surprises, and some data must be discarded because they are erroneous (e.g., male pregnancy, visits after death).

Perhaps most important, this book focuses on causal interpretation of statistics. In the past, statisticians have focused on association among variables. They have worked under the slogan that "correlation is not causation." While that statement is valid, policymakers, managers, and other decision makers act on the statistical findings as if correlation was causal. Any action assumes that the statistical findings are causal—that is, that changing one variable will lead to the desired impact. Statisticians who insist on avoiding causal interpretation of their findings are naive and are ignoring the obvious: their findings might be used differently than their planned precautions might have indicated. At the same time, they are also right to assert that causes are more than correlations. To interpret a variable as causing a change in another variable, we need to establish four principles:

1. *Association.* Causes have a statistically significant impact on effects.
2. *Sequence.* Causes occur before effects.
3. *Mechanism.* A third variable mediates the impact of the cause on the effect.
4. *Counterfactual.* In the absence of causes, effects are less likely to occur.

These four criteria allow us to discuss and vet causes rather than simply evaluating associations. In recent decades, statisticians have revisited their approach of avoiding causal interpretation and have introduced new techniques and methods that allow for evaluation of causality. For example, causal network models are an alternative to regression analysis. Network models allow the verification of the four assumptions of causality; regression models do not. Another example, propensity scoring, allows statisticians to

remove confounding in multivariate analysis and provides a causal estimate of the impact of a variable. This book starts with associations and conditional probabilities, but it uses these concepts to move on to propensity-matched regression analysis or causal networks. Even in early chapters, where we

EXHIBIT 1.1 Comparison of Emphasis of Managerial Statistics Books

Topic	Emphasis of Other Books	Emphasis of This Book
Distributions	• Normal, uniform, and other continuous distributions with little coverage of discrete probability theory	• Probability distribution in discrete events, including Bernoulli, binomial, geometric, and Poisson distributions • Normal distribution as an approximation
Data	• Measures collected from independent samples • Prospective data collection	• Longitudinal, time-based, repeated measures • Text as data • Observational and retrospective data
Study design	• Experimental design • Close-ended surveys	• Matched case control using observational data • Surveys of existing text
Confidence interval estimation	• Normal distribution estimation of confidence interval	• Estimation of upper and lower control limits in process control charts • Bootstrapped estimates of variability
Univariate methods of inference	• Comparison of mean to population • Comparison of two means • Paired *t*-test and comparison of dependent means • Analysis of variance	• Statistical process control tools such as XmR charts, p-charts, time-between charts, Tukey charts • Risk-adjusted process control tools
Multivariate analysis	• Correlation analysis • Multiple linear regression analysis • Logistic regression	• K-nearest neighbor • Propensity scoring • Sentiment analysis • Causal analysis • Multilevel intercept regression

discuss stratification and distributions, we lay the foundation for causal interpretations. In openly discussing causality, this book differs from many other introductory books on statistics.

Digital Aids and Multimedia

The book is accompanied by (1) slides to teach the course content, (2) video lectures, (3) video examples to illustrate the points made in the lecture, (3) extensive end-of-chapter exercises, (4) solutions to odd-numbered examples, and (5) a sample test set for midterm and finals. Topics in these supplements may be broader than the book, so take a look at them.

Relationship to Existing Courses

Students often do not understand the relationship between an introductory statistics course and other material they cover in health administration. *Statistical Analysis of Electronic Health Records* makes these linkages explicit. At the end of each chapter, the book directs you to the course website for problems to solve. Each problem is tied to a specific health administration or health informatics course. For example, problems in statistical process control are linked to courses in quality improvement. A problem in fraud detection is tied to the course in accounting. For still another example, comparative effectiveness analysis is linked to courses in strategy, informatics, and program evaluation. The expectation is that students will not only learn statistical concepts but also understand the connections between this course and various other courses in health administration programs.

Audience

The primary audience of this book is health administration and informatics students. In addition, nursing, physician assistant, and medical students may benefit. This book is not intended for a nonhealthcare audience.

Five Courses in One Book

This book can be used to teach many different courses:

1. The chapter on data preparation (chapter 2) and the chapter on risk assessment (chapter 5) can be used to teach an introductory course about SQL. These chapters present basic SQL commands and their use in constructing predictive models. Throughout the book, numerous examples of SQL code are provided that can further help students learning database design and analysis. The supplemental material of this chapter provides a syllabus for how to use this book to teach a course on SQL.
2. Chapters 3 through 7 can be used to replace an introductory course in statistics that focuses on hypothesis testing. These chapters introduce the concept of hypothesis testing and distributions. A syllabus is provided for courses that are exclusively focused on traditional hypothesis testing. The syllabus lists specific chapters and parts of chapters that may be helpful.
3. Chapters that focus on process control (chapters 5 through 10) can be used in a course on quality improvement. Many quality improvement courses discuss the general concepts but not the statistical tools, which is unfortunate. This book can improve the content of courses on quality improvement. A syllabus is provided for this type of course.
4. Chapters 11 and 12 can be used to teach a course on multivariate regression analysis. Chapters 13 (on propensity scoring), 14 (on hierarchical modeling), and 18 (on stratified regression) further show the value of ordinary regression. Again, a syllabus is provided for how to use this book to teach regression.
5. Chapters 13 through 20 can also be used to teach a course on causal analysis, especially in the context of comparative effectiveness analysis. These chapters enable students to remove confounding in EHR data. The supplemental material includes a syllabus for how to use this book to teach causal and comparative effectiveness courses.

Supplemental Resources

See tools for course design and syllabuses for various types of courses on the web.

References

Abelson, R., J. Creswell, and G. Palmer. 2012. "Medicare Bills Rise as Records Turn Electronic." *New York Times,* October 26. www.nytimes.com/2012/09/22/business/medicare-billing-rises-at-hospitals-with-electronic-records.html.

Alemi, F., M. Torii, L. Clementz, and D. C. Aron. 2012. "Feasibility of Real-Time Satisfaction Surveys Through Automated Analysis of Patients' Unstructured Comments and Sentiments." *Quality Management Health Care* 21 (1): 9–19.

Brynjolfsson, E. L. Hitt, and H. Kim. 2011. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" Accessed October 15, 2018. www.a51.nl/storage/pdf/SSRN_id1819486.pdf.

Eshkenazi A. 2012. "Joining the Big Data Revolution." *SCM NOW Magazine*. Accessed April 10, 2019. www.apics.org/apics-for-individuals/apics-magazine-home/magazine-detail-page/2012/10/26/joining-the-big-data-revolution.

Glaser, J., and R. Hess. 2011. "Leveraging Healthcare IT to Improve Operational Performance." *Healthcare Financial Management* 65 (2): 82–85.

Izakovic, M. 2007. "New Trends in the Management of Inpatients in U.S. Hospitals—Quality Measurements and Evidence-Based Medicine in Practice." *Bratislavské Lekárske Listy* 108 (3): 117–21.

Jaret, P. 2013. "Mining Electronic Records for Revealing Health Data." *New York Times*, January 14. www.nytimes.com/2013/01/15/health/mining-electronic-records-for-revealing-health-data.html.

Longman, P. 2010. *Best Care Anywhere: Why VA Health Care Is Better Than Yours*, 2nd ed. San Francisco: Berrett-Koehler Publishers.

McAfee, A., and E. Brynjolfsson. 2012. "Big Data: The Management Revolution." *Harvard Business Review* 90 (10): 60–66.

Nelson, E. C., M. E. Splaine, M. M. Godfrey, V. Kahn, A. Hess, P. Batalden, and S. K. Plume. 2000. "Using Data to Improve Medical Practice by Measuring Processes and Outcomes of Care." *Joint Commission Journal on Quality Improvement* 26 (12): 667–85.

Shortell S. M., C. L. Bennett, and G. R. Byck. 1998. "Assessing the Impact of Continuous Quality Improvement on Clinical Practice: What It Will Take to Accelerate Progress." *Milbank Quarterly* 76 (4): 593–624.

Stürzlinger, H., C. Hiebinger, D. Pertl, and P. Traurig. 2009. "Computerized Physician Order Entry: Effectiveness and Efficiency of Electronic Medication Ordering with Decision Support Systems." *GMS Health Technology Assessment* 19 (5): Doc07.

Tatonetti N. P., P. P. Ye, R. Daneshjou, and R. B. Altman. 2012. "Data-driven Prediction of Drug Effects and Interactions." *Science Translational Medicine* 4 (125): 125ra31.

Vest, J.R., and L. D. Gamm. 2009. "A Critical Review of the Research Literature on Six Sigma, Lean and StuderGroup's Hardwiring Excellence in the United States: The Need to Demonstrate and Communicate the Effectiveness of Transformation Strategies in Healthcare." *Implementation Science* 1 (4): 35.