Healthcare Data Analytics

WILLIAM R. HERSH

Learning Objectives

After reading this chapter the reader should be able to:

• Discuss the difference between descriptive, predictive and prescriptive analytics
• Outline the characteristics of “Big Data”
• Enumerate the necessary skills for a worker in the data analytics field
• List several limitations of healthcare data analytics
• Discuss the critical role electronic health records play in healthcare data analytics

Introduction

One of the promises of the growing critical mass of clinical data accumulating in electronic health record (EHR) systems is secondary use (or re-use) of the data for other purposes, such as quality improvement and clinical research. The growth of such data has increased dramatically in recent years due to incentives for EHR adoption in the US funded by the Health Information Technology for Economic and Clinical Health (HITECH) Act. In the meantime, there has also seen substantial growth in other kinds of health-related data, most notably through efforts to sequence genomes and other biological structures and functions. The analysis of this data is usually called analytics (or data analytics). This chapter will define the terminology of this field, provide an overview of its promise, describe what work has been accomplished, and list the challenges and opportunities going forward.

Terminology of Analytics

The terminology surrounding the use of large and varied types of data in healthcare is evolving, but the term analytics is achieving wide use both in and out of healthcare. A long-time leader in the field defines analytics as “the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions.” IBM defines analytics as “the systematic use of data and related business insights developed through applied analytical disciplines (e.g. statistical, contextual, quantitative, predictive, cognitive, other [including emerging] models) to drive fact-based decision making for planning, management, measurement and learning. Analytics may be descriptive, predictive or prescriptive.”

Adams and Klein have authored a primer on analytics in healthcare that defined different levels and their attributes of the application of analytics. They noted three levels of analytics, each with increasing functionality and value:

• Descriptive – standard types of reporting that describe current situations and problems
• Predictive – simulation and modeling techniques that identify trends and portend outcomes of actions taken
• Prescriptive – optimizing clinical, financial, and other outcomes
Much work is focusing now on predictive analytics, especially in clinical settings attempting to optimize health and financial outcomes.

There are a number of terms related to data analytics. A core methodology in data analytics is machine learning, which is the area of computer science that aims to build systems and algorithms that learn from data.\textsuperscript{8} One of the major techniques of machine learning is data mining, which is defined as the processing and modeling of large amounts of data to discover previously unknown patterns or relationships.\textsuperscript{9} A subarea of data mining is text mining, which applies data mining techniques to mostly unstructured textual data.\textsuperscript{10} Another close but more recent term in the vernacular is big data, which describes large and ever-increasing volumes of data that adhere to the following attributes:\textsuperscript{11}

- Volume – ever-increasing amounts
- Velocity – quickly generated
- Variety – many different types
- Veracity – from trustable sources

With the digitization of clinical data, hospitals and other healthcare organizations are generating an ever-increasing amount of data. In all healthcare organizations, clinical data takes a variety of forms, from structured (e.g., images, lab results, etc.) to unstructured (e.g., textual notes including clinical narratives, reports, and other types of documents). For example, it is estimated by Kaiser-Permanente that its current data store for its 9+ million members exceeds 30 petabytes of data.\textsuperscript{12} Other organizations are planning for a data-intensive future. Another example is the American Society for Clinical Oncology (ASCO) that is developing its Cancer Learning Intelligence Network for Quality (CancerLinQ).\textsuperscript{13} CancerLinQ will provide a comprehensive system for clinicians and researchers consisting of EHR data collection, application of clinical decision support, data mining and visualization, and quality feedback.

Another source of large amounts of data is the world’s growing base of scientific literature and its underlying data that is increasingly published in journals and other articles (see Chapter on online medical resources). One approach to this problem that has generated attention is the IBM Watson project, which started as a generic question-answering system that was made famous by winning at the TV game show Jeopardy!\textsuperscript{14} IBM has since focused Watson in the healthcare domain.\textsuperscript{15}

Kumar et al. have noted that the process of big data analytics resembles a pipeline, and have developed an approach that specifies four major steps in this pipeline, to which we can place data sources and actions on it pertinent to healthcare and biomedicine.\textsuperscript{16} The pipeline begins with input data sources, which in healthcare and biomedicine may include clinical records, financial records, genomics and related data, and other types, even those from outside the healthcare setting (e.g., census data). The next step is feature extraction, where various computational techniques are used to organize and extract elements of the data, such as linking records across sources, using natural language processing (NLP) to extract and normalize concepts, and matching of other patterns. This is followed by statistical processing, where machine learning and related statistical inference techniques are used to make conclusions from the data. The final step is the output of predictions, often with probabilistic measures of confidence in the results. (figure 3.1)

The growing quantity of data requires that its users have a good understanding of its provenance, which is where the data originated and how trustworthy it is for large-scale processing and analysis.\textsuperscript{17} A number of researchers and thought leaders have started to specify the path that will be required for big data to be applied in healthcare and biomedicine.\textsuperscript{18–20} An edited volume was recently published about analytics applied in various aspects healthcare and life sciences.\textsuperscript{21}

A more peripheral but related term is business intelligence, which in healthcare refers to the “processes and technologies used to obtain timely, valuable insights into business and clinical data”.\textsuperscript{7} Another relevant term is the notion promoted by the Institute of Medicine of the learning health system.\textsuperscript{22–23} Advocates of this approach note that routinely collected data can be used for continuous learning to allow the healthcare system to better carry out disease surveillance and response, targeting of healthcare services, improving decision-making, managing misinformation, reducing harm, avoiding costly errors, and advancing clinical research.\textsuperscript{24}
Another set of related terms come from the call for new and much more data-intensive approaches to diagnosis and treatment of disease variably called personalized medicine, precision medicine, or computational medicine. Advocates for these approaches note the inherent complexity of nonlinear systems in biomedicine, with large amounts and varied types of data that will need models to enable their predictive value. Technology thought leader O’Reilly notes that data science is transforming medicine, striving to solve its equivalent of the “Wanamaker Dilemma” for advertisers, named after the problem of knowing that half of advertising by merchants does not work, but that the half that does not work is not known.

One of the major motivators for data analytics comes from new models of healthcare delivery, such as accountable care organizations (ACOs), where reimbursement for conditions and episodes is bundled in a variety of ways, providing incentive move to deliver high-quality care in cost-efficient ways. ACOs require a focused IT infrastructure that provides data that can be used to predict and quickly act on excess costs. One of the challenges for healthcare data is that patients often get their care and testing in different settings (e.g., a patient seen in a physician office, sent to a free-standing laboratory or radiology center, and also seen in the offices of specialists or being hospitalized. This has increased the need for development of health information exchange (HIE), where data is shared among entities caring for a patient across business boundaries. A well-known informatics blogger has succinctly noted that “ACO = HIE + analytics”.

**Challenges to Data Analytics**

There are, of course, challenges to data analytics. One concern is that data generated in the routine care of patients may be limited in its use for analytical purposes. For example, such data may be inaccurate or incomplete. It may be transformed in ways that undermine its meaning (e.g., coding for billing priorities). It may exhibit the well-known statistical phenomenon of censoring, i.e., the first instance of disease in record may not be when it was first manifested (left censoring) or the data source may not cover a sufficiently long time interval (right censoring). Data may also incompletely adhere to well-known standards, which makes combining it from different sources more difficult. Finally, clinical data mostly
only allows observational and not experimental studies, thus raising issues of cause-and-effect of findings discovered.

Others have noted larger challenges around analytics and big data. Boyd and Crawford have expressed some “provocations” for the growing use of data-driven research. They note that research questions asked of the data tend to be driven by what can be answered, as opposed to prospective hypotheses. They also note that data are not always as objective as we might like, and that “bigger” is not necessarily better. Finally, they raise ethical concerns over how the data of individuals is used, the means by which it is collected, and the possible divide between those who have access to data and those who do not. Similar concerns focused specifically on healthcare data by Neff, who describes a myriad of technical, financial, and ethical issues that must be addressed before we will be able to make use of big data routinely for clinical practice and other health-related purposes. These challenges also create ethical issues, such as who owns data and who has privileges to use it.

**Research and Application of Analytics**

The research base around applying analytics to improve healthcare delivery is still in its early stages. There is an emerging base of research that demonstrates how data from operational clinical systems can be used to identify critical situations or patients whose costs are outliers. There is less research, however, demonstrating how this data can be put to use to actually improve clinical outcomes or reduce costs.

Studies using EHR data for clinical prediction have been proliferating. One common area of focus has been the use of data analytics to identify patients at risk for hospital readmission within 30 days of discharge. The importance of this factor comes from the US Centers for Medicare and Medicaid Services (CMS) Readmissions Reduction Program that penalizes hospitals for excessive numbers of readmissions. This has led several researchers to assess EHR data in its value to predict patients at risk for readmission.

A number of other critical clinical situations have been amenable to detection by analytics applied to EHR and other clinical data:

- Predicting 30-day risk of readmission and death among HIV-infected inpatients
- Identification of children with asthma
- Risk-adjusting hospital mortality rates
- Detecting postoperative complications
- Measuring processes of care
- Determining five-year life expectancy
- Detecting potential delays in cancer diagnosis
- Identifying patients with cirrhosis at high risk for readmission
- Predicting out of intensive care unit cardiopulmonary arrest or death

Additional efforts have focused on helping to identify patients for participation in research protocols or improve diagnosis of disease:

- Identifying patients who might be eligible for participation in clinical studies
- Determining eligibility for clinical trials
- Identifying patients with diabetes and the earliest date of diagnosis
- Predicting diagnosis in new patients

Other researchers have also been able to use EHR data to replicate the results of randomized controlled trials (RCTs). One large-scale effort has come from the Health Maintenance Organization Research Network’s Virtual Data Warehouse (VDW) Project. Using the VDW, for example, researchers were able to demonstrate a link between childhood obesity and hyperglycemia in pregnancy. Another demonstration of this ability has come from United Kingdom General Practice Research Database (UKGPRD), a repository of longitudinal records of general practitioners. Using this data, Tannen et al. were able to demonstrate the ability to replicate the findings of the Women’s Health Initiative and RCTs of other cardiovascular diseases. Likewise, Danaei et al. were able to combine subject-matter
expertise, complete data, and statistical methods emulating clinical trials to replicate RCTs demonstrating the value of statin drugs in primary prevention of coronary heart disease.\textsuperscript{60}

These large repositories have been used for other research purposes. For example, the UKGPRD has been used for determining risk factors for pancreatic cancer\textsuperscript{61} and gastroesophageal cancer.\textsuperscript{62} Another large data repository in the US allowed replication of prospective cohort studies for risks of venous thromboembolic events in a manner much more efficient that historical retrospective analyses.\textsuperscript{63} In addition, the Observational Medical Outcomes Partnership (OMOP) was to apply risk-identification methods to records from ten different large healthcare institutions in the US, although with a moderately high sensitivity vs. specificity tradeoff.\textsuperscript{64} Finally, a case report demonstrated a situation where a clinical research database was queried to help make a decision whether to anticoagulate a child with systematic lupus erythematosus (SLE), a question for which no scientific literature existed to answer.\textsuperscript{65} For an example of data analytics at a large healthcare system, see the Info box.

<table>
<thead>
<tr>
<th>Case Study: Veterans Health Administration (VHA)</th>
</tr>
</thead>
</table>
| The VHA is a large healthcare system with a long track record of EHR use (VistA). In 2013, the VHA had 30 million unique electronic patient records with 2 billion clinical notes (100,000 notes added daily). They also have had a corporate data warehouse (CDW) of structured data which allows them to analyze clinical and administrative data for patients at risk of hospital admission (from falls, coronary disease, PTSD, etc.). Analytics are run once weekly on all primary care patients looking for “at risk” patients who would likely require more coordinated care using care managers, home health and telehealth. In 2012, VHA researchers reported in the American Journal of Cardiology on the use of predictive analytics on heart failure patients. Specifically, using six categories of risk factors derived from the EHR they could successfully predict which patients were at risk of hospitalization and death.\textsuperscript{66}

According to Dr. Stephen Fihn, Director of Analytics and Business Intelligence for the VHA, the VHA is embarking on a 24-month pilot project to expand the use of healthcare data analytics. They will use natural language processing and machine learning to analyze patient records to aid in diagnosis, identify dangerous drug-drug interactions and optimally design treatment strategies.\textsuperscript{67}

Another approach used more novel methods. Denny and colleagues have developed methods for carrying out genome-wide association studies (GWAS) that associate specific findings from the EHR (the “phenotype”) with the growing amount of genomic and related data (the “genotype”) in the Electronic Medical Records and Genomics (eMERGE) Network.\textsuperscript{68} eMERGE has demonstrated the ability to validate existing research results and generate new findings,\textsuperscript{69} being able to identify genomic variants, among others, associated with atrioventricular conduction abnormalities,\textsuperscript{70} red blood cell traits,\textsuperscript{71} white blood cell count abnormalities,\textsuperscript{72} and thyroid disorders.\textsuperscript{73} More recent work has “inverted” the paradigm to carry out phenome-wide association studies (PheWAS) that associated multiple phenotypes with varying genotypes.\textsuperscript{74,75} Genome-wide and phenome-wide association studies are also discussed in the chapter on bioinformatics.

Clearly a large and growing body of research demonstrates that EHR and other clinical data can be used to predict outcomes, including adverse ones, as well as diagnoses and eligibility for research studies. The next step in research is to find evidence that such methods lead to improved patient outcomes. There are unfortunately a small number of studies, and their results are mixed. One study showed that a readmission tool applied to an existing case management approach helped reduce readmissions,\textsuperscript{76} while another found that use of a Bayesian network model embedded in EHR to predict hospital-acquired pressure ulcers led to a tenfold reduction in such ulcers as well as a reduction by one-third in intensive care unit length of stay for such patients.\textsuperscript{77} Another study found that a readmission risk tool intervention reduced risk of readmission for patients with congestive heart failure but not those with acute myocardial infarction or pneumonia.\textsuperscript{78} Another study found that an automated prediction model integrated into an existing EHR was successful in identifying patients on admission who were at risk for readmission within 30 days of discharge, but its use had no effect on 30-day all-cause and 7-day unplanned readmission rates in the 12-month period after it was implemented.\textsuperscript{79}
Role of Informaticians in Analytics

Although much has been written extolling the virtues of analytics and big data analytics, little of it focuses on the human experts who will carry out the work, to say nothing of those who will support their efforts in building systems to capture data, put it into usable form, and apply the results of analysis. Many of those who collect, analyze, use, and evaluate data will come from the workforce of biomedical and health informatics. To this end, we must ask questions about the job activity as well as the education of those who work in this emerging area that some call data science. Data analytics thought leader Davenport asserts that data science is the “sexiest job of the 21st century,” in that those who perform it have rare qualities in high demand.

In the worlds of healthcare and biomedicine, the field poised to lead in data science is informatics. After all, informatics has led the charge in implementing systems that capture, analyze, and apply data across the biomedical spectrum from genomics to healthcare to public health. From basic biomedical scientists to clinicians and public health workers, those who are researchers and practitioners are drowning in data, needing tools and techniques to allow its use in meaningful and actionable ways.

Data science is more than statistics or computer science applied in a specific subject domain. Dhar notes that a key aspect of data science, in particular what distinguishes it from statistics, is an understanding of data, its varying types, and how to manipulate and leverage it. He points out that skills in machine learning are key, based upon a foundation of statistics (especially Bayesian), computer science (representation and manipulation of data), and knowledge of correlation and causation (modeling). Dhar also notes a challenge to organizational culture that might occur as organizations moved from “intuition-based” to “fact-based” decision-making.

It is also clear that there are two types of individuals working with analytics and big data. A report by the McKinsey consulting firm states that there will soon be a need in the US for 140,000-190,000 individuals who have “deep analytical talent”. Furthermore, the report notes there will be need for an additional 1.5 million “data-savvy managers needed to take full advantage of big data”. Analyses from the UK find similar results. An analysis by SAS estimated that by 2018, there will be over 6400 organizations that will hire 100 or more analytics staff. Another report found that data scientists currently comprise less than 1% of all big data positions, with more common job roles consisting of developers (42% of advertised positions), architects (10%), analysts (8%) and administrators (6%). It was also found that the technical skills most commonly required for big data positions as a whole were NoSQL, Oracle, Java and SQL. While these estimates are not limited to healthcare, they also do not include other countries that will have comparable needs to the US and the UK for such talent.

A report from IBM Global Services noted healthcare organizations are lagging behind in hiring individuals who are proficient in both “numerate” and business-oriented skills. An additional report from IBM Global Services list “expertise” among the critical attributes in organizations that are needed to complement technology. This expertise includes the supplementation of business knowledge with analytics knowledge, establishing formal career paths for analytics professionals, and tapping partners to supplement skills gaps that may exist. Another US-based report by PriceWaterhouseCoopers on health IT talent shortages noted that healthcare organizations wanting to keep ahead needed to acquire talent in Systems and data integration, data statistics and analytics, technology and architecture support, and clinical informatics.

The US National Institutes of Health (NIH) also recognizes that big data skills will be important for conducting biomedical research. In 2013, NIH convened a workshop on enhancing training in big data among researchers. Similar to the healthcare domain, participants called for skills in quantitative sciences, domain expertise, and ability to work in diverse teams. The workshop also noted a need for those working in big data to understand concepts of managing and sharing data. Trainees should also have access to real-world data problems and real-sized data sets to solve them. Longer-term training would be required for those becoming experts and leaders in data science.
What do biomedical and health informaticians working in analytics and big data need to know? An emerging consensus can be drawn from the reports above indicates that a combination of skills will be required:

1. Programming - especially with data-oriented tools, such as SQL and statistical programming languages
2. Statistics - working knowledge to apply tools and techniques
3. Domain knowledge - depending on one's area of work, bioscience or health care
4. Communication - being able to understand needs of people and organizations and articulate results back to them

Thus to be relevant, informatics educational programs will need to introduce concepts of analytics, big data, and the underlying skills to use and apply them into their curricula. There will be a need for appropriate coursework for those who will become the “deep analytical talent” as well as higher breadth, perhaps with lesser depth, for the order of magnitude more individuals who will apply the results of big data analytics in healthcare and biomedical research.

**Recommended Reading**

The following are interesting references to expand your healthcare data analytics knowledge:

- *Mining Electronic Health Records in the Genomics Era*. A book chapter providing an overview of techniques for extracting structured and narrative text from EHRs, with a focus on genotype-phenotype correlations.
- *Caveats and recommendations for use of operational clinical data in research*. A pair of papers noting challenges and overcoming them for use of EHR data in clinical research

**Future Trends**

As the volume of clinical data and the need for analytics continues to accelerate, systematic approaches will be required for sustained success. One recent analysis laid out recommendations for operational use of clinical data. Although focused on comparative effectiveness research, the recommendations can be applied for almost any data analytics task. The authors called for:

- Adherence to best practices for use of data standards and interoperability
- Processes to evaluate availability, completeness, quality, and transformability of data
- Toolkits and pipelines to manage data and its attributes
- Challenges and metrics for assessing “research grade” of operational data
- Standardized reporting methods for operational data and its attributes
- Adaptation of “best evidence” approaches to use of operational data
- Appropriate use of informatics expertise to assist with optimal use of operational data and to develop published guidelines for doing so
- Research agenda to determine biases inherent in operational data and to assess informatics approaches to improve data

The “best evidence” approach is modeled on the framework of evidence-based medicine (EBM), applying the four basic steps of EBM to clinical data instead of scientific studies:

- Ask an answerable question – can question be answered by the data we have?
- Find the best evidence – in this case, the best evidence is the EHR data needed to answer the question
- Critically appraise the evidence – does the data answer the question? Are there confounders?
- Apply it to the patient situation – can the data be applied to this setting?
Key Points

- Healthcare data has proliferated greatly, in large part due to the accelerated adoption of EHRs
- Analytic platforms will examine data from multiple sources, such as clinical records, genomic data, financial systems, and administrative systems
- Analytics is necessary to transform data to information and knowledge
- Accountable care organizations and other new models of healthcare delivery will rely heavily on analytics to analyze financial and clinical data
- There is a great demand for skilled data analysts in healthcare; expertise in informatics will be important for such individuals

Conclusion

Clearly there is great promise ahead for healthcare driven by data analytics. The growing quantity of clinical and research data, along with methods to analyze and put it to use, can lead to improve personal health, healthcare delivery, and biomedical research. However, there is also a continued need to improve the completeness and quality of data as well as conduct research to demonstrate how to best apply it to solve real-world problems. In addition, human expertise, including in informatics, will be required to optimally carry out such work.

References

29. Longworth DL. Accountable care organizations, the patient-centered medical home, and health care reform: what does it all mean? Cleveland Clinic Journal of Medicine, 2011. 78: 571-589.


82. Wang L, Porter B, Maynard C et al. Predicting Risk of Hospitalization or Death Among Patients with Heart Failure in the Veterans Health Administration. AJC. 2012;110(9):1342-1349


