From: Hoyt, RE, Yoshihashi, A, Eds. (2014). *Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, Sixth Edition*. Pensacola, FL, Lulu.com.

3

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.[1] 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.[2-3] 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.[4] 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".[5] 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." [6]

Adams and Klein have authored a primer on analytics in healthcare that defined different levels and their attributes of the application of analytics.[7] 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.[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.[9] A subarea of data mining is *text mining*, which applies data mining techniques to mostly unstructured textual data.[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:[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.[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).[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.

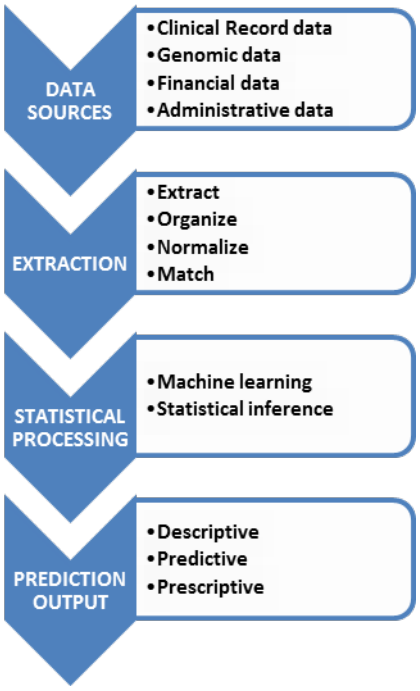
Another source of large amounts of data is the world's growing base of scientific literature and its underlying data that is increasingly published in journals and other articles (see Chapter on online medical resources). One approach to this problem that has generated attention is the IBM Watson project, which started as a generic question-answering system that was made famous by winning at the TV game show Jeopardy![14] IBM has since focused Watson in the healthcare domain.[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.[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.[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.[18-20] An edited volume was recently published about analytics applied in various aspects healthcare and life sciences.[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".[7] Another relevant term is the notion promoted by the Institute of Medicine of the *learning health system*.[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.[24]

Figure 3.1 The Analytics Pipeline (Adapted from Kumar)[16]



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,*[25] *precision medicine,*[26] or *computational medicine*.[27] 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.[28]

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.[29] ACOs require a focused IT infrastructure that provides data that can be used to predict and quickly act on excess costs.[30] 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.[31] A well-known informatics blogger has succinctly noted that "ACO = HIE + analytics".[32]

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.[33] 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.[34] 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.[35] These challenges also create ethical issues, such as who owns data and who has privileges to use it.[36]

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.[37] This has led several researchers to assess EHR data in its value to predict patients at risk for readmission.[38-40]

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 [41]
- Identification of children with asthma[42]
- Risk-adjusting hospital mortality rates [43]
- Detecting postoperative complications [44]
- Measuring processes of care [45]
- Determining five-year life expectancy [46]
- Detecting potential delays in cancer diagnosis [47]
- Identifying patients with cirrhosis at high risk for readmission [48]
- Predicting out of intensive care unit cardiopulmonary arrest or death [49]

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 [50]
- Determining eligibility for clinical trials [51]
- Identifying patients with diabetes and the earliest date of diagnosis [52]
- Predicting diagnosis in new patients [53]

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.[54] Using the VDW, for example, researchers were able to demonstrate a link between childhood obesity and hyperglycemia in pregnancy.[55] 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 [56-57] and RCTs of other cardiovascular diseases. [58-59] 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.[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 [61] and gastroesophogeal cancer.[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.[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.[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.[65] For an example of data analytics at a large healthcare system, see the Info box.

Case Study: Veterans Health Administration (VHA)

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.[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.[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.[68] eMERGE has demonstrated the ability to validate existing research results and generate new findings,[69] being able to identify genomic variants, among others, associated with atrioventricular conduction abnormalities,[70] red blood cell traits,[71] white blood cell count abnormalities,[72] and thyroid disorders.[73] More recent work has "inverted" the paradigm to carry out phenome-wide association studies (PheWAS) that associated multiple phenotypes with varying genotypes.[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,[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.[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.[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.[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*. [80] 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. [81]

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 health care to public health. [82] 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.[80] 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".[83] 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. [84] 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%). [85] 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. [86] 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.[87] 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.[88]

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. [89] 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.[66]
- *Caveats[33] and recommendations[90] 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
- *Analytics in Healthcare and the Life Sciences: Strategies, Implementation Methods, and Best Practices*. A book describing tools and best practices for use of analytics for clinical care, pharmaceutical research, and patient engagement.[21]

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.[90] 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: [90]

- 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

1. Safran C, Bloomrosen M, Hammond WE, Labkoff SE, Markel-Fox S, Tang P, et al., Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J AmMed Infor Assoc. 2007; 14: 1-9.
2. Blumenthal D. Wiring the health system--origins and provisions of a new federal program. New England Journal of Medicine.2011;365: 2323-2329.
3. Blumenthal D. Implementation of the federal health information technology initiative. New England Journal of Medicine.2011;365: 2426-2431.
4. Miller K, *Big Data Analytics in Biomedical Research*, Biomedical Computation Review. (Winter 2011/2012): 14-21. http://biomedicalcomputationreview.org/sites/default/files/w12_f1-big_data.pdf. (Accessed December 4, 2013)
5. Davenport TH and Harris JG, Competing on Analytics: The New Science of Winning. 2007, Cambridge, MA: Harvard Business School Press.
6. Anonymous, The value of analytics in healthcare - From insights to outcomes. 2012, IBM Global Services: Somers, NY, http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-healthcare-analytics.html.(Accessed December 4, 2013)
7. Adams J and Klein J. Business Intelligence and Analytics in Health Care - A Primer. 2011, The Advisory Board Company: Washington, DC, http://www.advisory.com/Research/IT-Strategy-Council/Research-Notes/2011/Business-Intelligence-and-Analytics-in-Health-Care. (Accessed December 5, 2013)
8. Mohri M, Rostamizadeh A, and Talwalkar A. Foundations of Machine Learning. 2012, Cambirdge, MA: MIT Press.
9. Bellazzi R and Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics, 2008. 77: 81-97.
10. Cohen AM and Hersh WR. A survey of current work in biomedical text mining. Briefings in Bioinformatics, 2005. 6: 57-71.

11. Zikopoulos P, Eaton C, deRoos D, Deutsch T, and Lapis G. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. 2011, New York, NY: McGraw-Hill.
12. Gardner E. The HIT Approach to Big Data. Health Data Management. March 1, 2013. http://www.healthdatamanagement.com/issues/21_3/The-HIT-Approach-to-Big-Data-Anayltics-45735-1.html. (Accessed November 30, 2013)
13. Sledge GW, Miller RS, and Hauser R, CancerLinQ and the future of cancer care. 2013 ASCO Educatonal Book, 2013: 430-434. http://meetinglibrary.asco.org/content/58-132
14. Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, et al. *B*uilding Watson: an overview of the DeepQA Project. AI Magazine, 2010. 31(3): 59-79. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303. (Accessed December 1,2013)
15. Ferrucci D, Levas A, Bagchi S, Gondek D, and Mueller E. Watson: beyond Jeopardy! Artificial Intelligence.2012;199-200: 93-105.
16. Kumar A, Niu F, and Ré C. Hazy: making it easier to build and maintain big-data analytics. Communications of the ACM, 2013. 56(3): 40-49.
17. Buneman P and Davidson SB. Data provenance – the foundation of data quality. 2010, Carnegie Mellon University Software Engineering Institute: Pittsburgh, PA, http://www.sei.cmu.edu/measurement/research/upload/Davidson.pdf.
18. Minelli M, Chambers M, and Dhiraj A. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. 2013, Hoboken, NJ: Wiley.
19. Murdoch TB and Detsky AS. The inevitable application of big data to health care. Journal of the American Medical Association.2013;309: 1351-1352.
20. Groves, Kayyali B, Knott D, and VanKuiken S. The big-data revolution in US health care: Accelerating value and innovation. 2013, McKinsey Global Institute, http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care. (Accessed December 1, 2013)
21. McNeill D, ed. Analytics in Healthcare and the Life Sciences: Strategies, Implementation Methods, and Best Practices. 2013, Pearson Education: Upper Saddle River, New Jersey.
22. Friedman CP, Wong AK, and Blumenthal D, Achieving a nationwide learning health system. Science Translational Medicine, 2010. 2(57): 57cm29. http://stm.sciencemag.org/content/2/57/57cm29.full. (Accessed December 2, 2013)
23. Smith M, Saunders R, Stuckhardt L, and McGinnis JM. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. 2012, Washington, DC: National Academies Press.
24. Okun S, McGraw D, Stang P, Larson E, Goldmann D, Kupersmith J, et al., Making the Case for Continuous Learning from Routinely Collected Data. 2013, Institute of Medicine: Washington, DC, http://www.iom.edu/Global/Perspectives/2013/MakingtheCaseforContinuousLearning.aspx.
25. Hamburg MA and Collins FS. The path to personalized medicine. New England Journal of Medicine, 2010. 363: 301-304.
26. Anonymous. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. 2011, Washington, DC: National Academies Press.
27. Winslow RL, Trayanova N, Geman D, and Miller MI. Computational medicine: translating models to clinical care. Science Translational Medicine, 2012. 4: 158rv11. http://stm.sciencemag.org/content/4/158/158rv11.short.
28. O'Reilly T, Loukides M, Steele J, and Hill C. How Data Science Is Transforming Health Care. 2012, Sebastapol, CA: O'Reilly Media.
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.
30. Anonymous. A Health IT Framework for Accountable Care. 2013, Certification Commission for Health Information Technology: Chicago, IL, https://http://www.cchit.org/hitframework.(Accessed December 2, 2013)
31. Kuperman GJ. Health-information exchange: why are we doing it, and what are we doing? Journal of the American Medical Informatics Association, 2011. 18: 678-682.
32. Halamka J. The "Post EHR" Era. Life as a Healthcare CIO. February 12, 2013. http://geekdoctor.blogspot.com/2013/02/the-post-ehr-era.html.(Accessed December 3, 2013)
33. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al., Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical Care, 2013. 51(Suppl 3): S30-S37.

34. Boyd D and Crawford K. Six Provocations for Big Data. 2011, Microsoft Research: Cambridge, MA, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431.
35. Neff G. Why big data won't cure us. Big Data, 2013. 1: 117-123.
36. Trotter F. Who Owns Patient Data? The Health Care Blog. August 20, 2012. http://thehealthcareblog.com/blog/2012/08/20/who-owns-patient-data/.(Accessed December 4, 2013)
37. Anonymous. Readmissions Reduction Program. 2013, Center for Medicare and Medicaid Services: Washington, DC, http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html. (Accessed December 5, 2013)
38. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al., An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Medical Care, 2010. 48: 981-988.
39. Donzé J, Aujesky D, Williams D, and Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA Internal Medicine, 2013. 173: 632-638.
40. Gildersleeve R and Cooper P. Development of an automated, real time surveillance tool for predicting readmissions at a community hospital. Applied Clinical Informatics, 2013. 4: 153-169.
41. Nijhawan AE, Clark C, Kaplan R, Moore B, Halm EA, and Amarasingham R. An electronic medical record-based model to predict 30-day risk of readmission and death among HIV-infected inpatients. Journal of Acquired Immunodeficiency Syndrome, 2012. 61: 349-358.
42. Afzal Z, Engelkes M, Verhamme KM, Janssens HM, Sturkenboom MC, Kors JA, et al. Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases. Pharmacoepidemiology and Drug Safety, 2013. 22: 826-833.
43. Escobar GJ, Gardner MN, Greene JD, Draper D, and Kipnis P. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. Medical Care, 2013. 51: 446-453.
44. FitzHenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. Medical Care, 2013. 51: 509-516.
45. Tai-Seale M, Wilson CJ, Panattoni L, Kohli N, Stone A, Hung DY, et al. Leveraging electronic health records to develop measurements for processes of care. Medical Care, 2013: Epub ahead of print.
46. Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, and Choudhary A. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. Journal of the American Medical Informatics Association, 2013. 20(e1): e118-e124.
47. Murphy DR, Laxmisan A, Reis BA, Thomas EJ, Esquivel A, Forjuoh SN, et al. Electronic health record-based triggers to detect potential delays in cancer diagnosis. BMJ Quality & Safety, 2013: Epub ahead of print.
48. Singal AG, Rahimi RS, Clark C, Ma Y, Cuthbert JA, Rockey DC, et al. An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission. Clinical Gastroenterology and Hepatology, 2013. 11: 1335-1341.
49. Alvarez CA, Clark CA, Zhang S, Halm EA, Shannon JJ, Girod CE, et al. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. BMC Medical Informatics & Decision Making, 2013. 13: 28. http://www.biomedcentral.com/1472-6947/13/28.(Accessed December 5, 2013)
50. Voorhees E and Hersh W. Overview of the TREC 2012 Medical Records Track. The Twenty-First Text REtrieval Conference Proceedings (TREC 2012). 2012. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf.(Accessed November 11, 2013)
51. Köpcke F, Lubgan D, Fietkau R, Scholler A, Nau C, Stürzl M, et al. Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. BMC Medical Informatics & Decision Making, 2013. 13: 134. http://www.biomedcentral.com/1472-6947/13/134.
52. Makam AN, Nguyen OK, Moore B, Ma Y, and Amarasingham R. Identifying patients with diabetes and the earliest date of diagnosis in real time: an electronic health record case-finding algorithm.

BMC Medical Informatics & Decision Making, 2013. 13: 81. http://www.biomedcentral.com/1472-6947/13/81.
53. Gottlieb A, Stein GY, Ruppin E, Altman RB, and Sharan R. A method for inferring medical diagnoses from patient similarities. BMC Medicine, 2013. 11: 194. http://www.biomedcentral.com/1741-7015/11/194.
54. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, et al. Building a virtual cancer research organization. Journal of the National Cancer Institute Monographs, 2005. 35: 12-25.
55. Hillier TA, Pedula KL, Schmidt MM, Mullen JA, Charles MA, and Pettitt DJ. Childhood obesity and metabolic imprinting: the ongoing effects of maternal hyperglycemia. Diabetes Care, 2007. 30: 2287-2292.
56. Tannen RL, Weiner MG, Xie D, and Barnhart K. A simulation using data from a primary care practice database closely replicated the Women's Health Initiative trial. Journal of Clinical Epidemiology, 2007. 60: 686-695.
57. Weiner MG, Barnhart K, Xie D, and Tannen RL. Hormone therapy and coronary heart disease in young women. Menopause, 2008. 15: 86-93.
58. Tannen RL, Weiner MG, and Xie D. Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. Pharmacoepidemiology and Drug Safety, 2008. 17: 671-685.
59. Tannen RL, Weiner MG, and Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. British Medical Journal, 2009. 338: b81. http://www.bmj.com/cgi/content/full/338/jan27_1/b81.
60. Danaei G, Rodríguez LA, Cantero OF, Logan R, and Hernán MA. Observational data for comparative effectiveness research: An emulation of randomised trials of statins and primary prevention of coronary heart disease. Statistical Methods in Medical Research, 2011. 22: 70-96.
61. Stapley S, Peters TJ, Neal RD, Rose PW, Walter FM, and Hamilton W. The risk of pancreatic cancer in symptomatic patients in primary care: a large case-control study using electronic records. British Journal of Cancer, 2012. 106: 1940-1944.
62. Stapley S, Peters TJ, Neal RD, Rose PW, Walter FM, and Hamilton W. The risk of oesophago-gastric cancer in symptomatic patients in primary care: a large case-control study using electronic records. British Journal of Cancer, 2013. 108: 25-31.
63. Kaelber DC, Foster W, Gilder J, Love TE, and Jain AK. Patient characteristics associated with venous thromboembolic events: a cohort study using pooled electronic health record data. Journal of the American Medical Informatics Association, 2012. 19: 965-972.
64. Ryan PB, Madigan D, Stang SE, Overhage JM, Racoosin JA, and Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Statistics in Medicine, 2012. 31: 4401-4415.
65. Frankovich J, Longhurst CA, and Sutherland SM. Evidence-based medicine in the EMR era. New England Journal of Medicine, 2011. 365: 1758–1759.
66. 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
67. Pedulli L. Veteran's Affairs Drives Data Mining. Clinical Innovation and Technology. May 23 2013. www.clinical-innovation.com (Accessed December 22 2013)
68. Denny JC. Mining Electronic Health Records in the Genomics Era, in PLOS Computational Biology: Translational Bioinformatics, Kann M and Lewitter F, Editors. 2012.
69. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Genomics, 2010. 4(1): 13. http://www.biomedcentral.com/1755-8794/4/13.
70. Denny JC, Ritchie MD, Crawford DC, Schildcrout JS, Ramirez AH, Pulley JM, et al., Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. Circulation, 2010. 122: 2016-2021.
71. Kullo LJ, Ding K, Jouni H, Smith CY, and Chute CG. A genome-wide association study of red blood cell traits using the electronic medical record. PLoS ONE, 2010. 5(9): e13011.

72. Crosslin DR, McDavid A, Weston N, Nelson SC, Zheng X, Hart E, et al. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. Human Genetics, 2012. 131: 639-652.
73. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. American Journal of Human Genetics, 2011. 89: 529-542.
74. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature Biotechnology, 2013. 31: 1102-1111.
75. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association, 2013. 20(e1): e147-154.
76. Gilbert P, Rutland MD, and Brockopp D. Redesigning the work of case management: testing a predictive model for readmission. American Journal of Managed Care, 2013. 19(11 Spec No. 10): eS19-eSP25. http://www.ajmc.com/publications/issue/2013/2013-11-vol19-sp/redesigning-the-work-of-case-management-testing-a-predictive-model-for-readmission.(Accessed December4, 2013)
77. Cho I, Park I, Kim E, Lee E, and Bates DW. Using EHR data to predict hospital-acquired pressure ulcers: A prospective study of a Bayesian Network model. International Journal of Medical Informatics, 2013. 82: 1059-1067.
78. Amarasingham R, Patel PC, Toto K, Nelson LL, Swanson TS, Moore BJ, et al. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. BMJ Quality & Safety, 2013. 22: 998-1005.
79. Baillie CA, VanZandbergen C, Tait G, Hanish A, Leas B, French B, et al. The readmission risk flag: Using the electronic health record to automatically identify patients at risk for 30-day readmission. Journal of Hospital Medicine, 2013. 8: 689-695.
80. Dhar V.Data science and prediction. Communications of the ACM, 2013. 56(12): 64-73.
81. Davenport TH and Patil DJ. Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review. October, 2012. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/.
82. Hersh W. A stimulus to define informatics and health information technology. BMC Medical Informatics & Decision Making, 2009. 9: 24. http://www.biomedcentral.com/1472-6947/9/24/.
83. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: The next frontier for innovation, competition, and productivity. 2011, McKinsey Global Institute, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
84. Anonymous. An assessment of demand for labour and skills, 2012-2017. 2013, SAS: London, England. http://www.e-skills.com/Documents/Research/General/BigDataAnalytics_Report_Jan2013.pdf.(Accessed December 3, 2103)
85. Anonymous. Adoption and employment trends, 2012-1017. 2013, SAS: London, England, http://www.e-skills.com/Documents/Research/General/BigDataAnalytics_Report_Nov2013.pdf.(Accessed November 30,2013)
86. Fraser H, Jayadewa C, Mooiweer P, Gordon D, and Piccone J. Analytics across the ecosystem - A prescription for optimizing healthcare outcomes. 2013, IBM Global Services: Somers, NY, http://www-935.ibm.com/services/us/gbs/thoughtleadership/healthcare-ecosystem/.(Accessed December 2, 2013)
87. Balboni F, Finch G, Rodenbeck-Reese C, and Shockley R. Analytics: A blueprint for value. 2013, IBM Global Services: Somers, NY, http://www-935.ibm.com/services/us/gbs/thoughtleadership/ninelevers/.(Accessed December 1, 2103)
88. Anonymous. Solving the talent equation for health IT. 2013, PriceWaterhouseCoopers, http://www.pwc.com/us/HITtalent.(Accessed December 5, 2013)

89. Anonymous. Report of Workshop on Enhancing Training for Biomedical Big Data. 2013, National Institutes of Health, http://bd2k.nih.gov/pdf/bd2k_training_workshop_report.pdf.(Accessed December 6, 2013)
90. Hersh WR, Cimino JJ, Payne PR, Embi PJ, Logan JR, Weiner MG, et al. Recommendations for the use of operational electronic health record data in comparative effectiveness research. eGEMs (Generating Evidence & Methods to improve patient outcomes), 2013. 1: 14. http://repository.academyhealth.org/egems/vol1/iss1/14/.(Accessed December 1, 2013)