Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

A review of big data in health care: challenges and opportunities

This article was published in the following Dove Press journal:
Open Access Bioinformatics
31 October 2014
Number of times this article has been viewed

Susan E White

Health Information Management and Systems Division, School of Health and Rehabilitation Sciences, The Ohio State University, Columbus, OH, USA

Abstract: Health care is a data-rich industry. Administrative databases hold a tremendous number of transactions for each patient treated. The expansion of the adoption of electronic health records due to the Health Information Technology for Economical and Clinical Health (HITECH) provision of the American Recovery and Reinvestment Act is increasing the amount of data available exponentially. Still, the health care industry has been slow to leverage the vast data to improve care and health care operations. The adoption of value-based purchasing programs by Medicare and commercial payers along with increased demand for accountable care organizations motivated by the Affordable Care Act are moving both providers and payers to use data to improve operations. Health care's big data has the potential to revamp the process of health care delivery in the US and inform providers about the most efficient and effective treatment pathways. Value-based purchasing programs are incenting both health care providers and insurers to investigate new ways to leverage health care data to measure the quality and efficiency of care. The use of analytics in health care data presents of a number of daunting challenges, but also rich opportunities.

Keywords: big data, predictive analytics, patient privacy, information governance, data governance

Introduction

Big data is defined in a number of ways. Gartner defines big data as "high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."[1] Their "3V" definition is used by many other organizations. Health care data does appear to fit the "3V" portion of the Gartner definition. According to HealthCatalyst,[2] Health care firms with more than 1,000 employees store over 400 TB of data per firm. This places health care fourth after securities and investment services, communications and media, and manufacturing. This certainly qualifies health care as a high-data volume industry. The transactional data in the health care industry changes rapidly. Claims are paid on a daily basis; patient data is abstracted into electronic health records (EHRs) multiple times a day; and the results of diagnostic tests are recorded electronically in real time. All of these attributes support the assertion that health care data meet the high-velocity criteria. Finally, health care data vary from discrete coded data elements to images of diagnostics tests to unstructured clinical notes. Although health care data meet the volume, velocity, and variety criteria, historically that data has not been used to enhance insight or decision making to its fullest extent.

Correspondence: Susan E White
Health Information Management and Systems Division, School of Health and Rehabilitation Sciences, The Ohio State University, 453 West 10th Ave – Atwell Hall #543, Columbus, OH 43210, USA
Tel +1 614 247 2495
Email white.2@osu.edu

Methods

A literature review was conducted to identify recent articles about the use of big data in health care. The following search terms were used: "big data in healthcare," "big data in health care," "big data medicine," and "big data clinical." The search terms were used with PubMed, Google Scholar, Science Direct, and Web of Knowledge as well as Google to identify both peer-reviewed literature and professional journal articles addressing the use and application of big data in health care settings. The primary search included only articles published within the last 5 years. Secondary references from older articles were used to allow full discussion of issues identified in primary, more recent articles.

Challenges

There are a number of challenges that make it difficult to use health care data to its fullest extent. First, the data in many health care providers, specifically hospitals, are often segmented or siloed. Administrative data such as claims, reimbursement, and cost information are stored and used by the financial and operational management teams. This data is used to carry out the business side of health care, but generally not used to inform patient care or treatment protocols. Clinical data such as patient history, vital signs, progress notes, and the results of diagnostic tests are stored in the EHR. Clinical data is accessed and maintained by the physicians, nurses, and other frontline clinical staff and is used to track patient care and communicate treatment plans throughout the team of clinicians providing care to the patient. Quality and outcomes data such as surgical site infections, rates of return to surgery, patient falls, and Centers for Medicare and Medicaid Services' (CMS) value-based purchasing measures[3] are in the domain of the quality or risk management departments. This data is collected and typically used to make retrospective measurement of the performance of the provider. The Health Research Institute's 2011 Clinical Informatics Survey[4] found that 43% of respondents listed "data being kept in silos throughout the organization" as an organizational barrier to analyzing clinical data. This survey included the provider, health insurer, and pharmaceutical industry professionals. Therefore, this issue of siloed or segmented data sources expands beyond providers and throughout the health care industry.

Data analytics involving the optimal use of the hospital's resources and improving patient outcomes can only be achieved by combining these datasets to meet the second portion of the Gartner definition, "cost-effective, innovative forms of information processing for enhanced insight and decision making." Many providers are working hard to overcome this issue and use tools such as data warehouses and decision support databases to allow researchers and analysts to combine data from traditionally segmented sources. Researchers at the University of Michigan[5] point out the value of the Learning Health System, which was defined by the Institute of Medicine to be "in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and health care."[6] Rubin and Friedman provide a number of examples of the power of combined health care data sources, including avoidance of duplicate diagnostic tests, accelerating the speed that information travels to frontline clinicians, and improved communications of epidemics or significant drug adverse events.[5]

A second significant challenge in leveraging health care's big data to its fullest extent is protecting the patient's privacy. The sharing of health care data between organizations is often stated as a goal and organizations such as regional health information organizations were specifically formed to bring together health care data from stakeholders including providers, payers, and public health organizations. The Health Insurance Portability and Accountability Act requires covered entities to protect patient information.[7] Patient data may be shared after de-identification, but protecting the patient from either direct or indirect identification while still maintaining the usefulness of the data is challenging. Covered entities, including health care providers and health insurance companies among others, often err on the conservative side and release only aggregate data or data with all potential identifiers removed. The removal of all protected health information requires the removal of the data elements found in Table 1.[8] Removing these data elements and meeting the Health Insurance Portability and Accountability Act de-identification criteria of "safe harbor" makes the use of data for trending or longitudinal care studies nearly impossible.

The removal of date elements (C in Table 1) is problematic when studies involve a time component such as those examining readmission rates or morality rates. Often researchers are left with the option to base all of the data on some baseline date and use a de-identified measure of lag or number of days until the event of interest. This practice can compromise the generalizability of the study if clinical practice or treatment options change over time during the study period.

Even if the privacy of the patient can be protected, many health care providers are reluctant to share data because of market competition. A physician many not want their competitors to know exactly how many procedures they performed and where. The patient insurance mix or demographics may provide one hospital a financial advantage over another.

Table 1 Restricted data elements

A. Names
B. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:
   1. The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
   2. The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
C. All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
D. Telephone numbers
E. Fax numbers
F. Email addresses
G. Social security numbers
H. Medical record numbers
I. Health plan beneficiary numbers
J. Account numbers
K. Certificate/license numbers
L. Vehicle identifiers and serial numbers, including license plate numbers
M. Device identifiers and serial numbers
N. Web universal resource locators (URLs)
O. Internet protocol (IP) addresses
P. Biometric identifiers, including finger and voice prints
Q. Full-face photographs and any comparable images
R. Any other unique identifying number, characteristic, or code, except as permitted by paragraph (C) of the guidance document

Notes: Data from Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule [webpage on the Internet]. Washington, DC: Department of Health and Human Services; 2013 [cited June 14, 2013]. Available from: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#guidancedetermination. Accessed June 14, 2014.[8]

Although most hospitals are operated as not-for-profit entities, they are still a business and all of the rules of operating a competitive business apply. There are a number of publicly available datasets that may allow competitors to glean similar information, but those sources are typically historical data or limited to government payers.

The patients themselves are increasingly becoming a source of data. The collection of this data and the impact of its inclusion into the health care record are critical in the formulation of a robust data governance plan. This data may be collected through monitoring systems that are connected to an offsite database via wireless technology or periodically uploaded from the device during an office visit. In either scenario, the data must be validated to ensure that the patient was actually using the monitoring device and not transferring it to another person in the household. The risk of compromised data integrity is much higher with these patient collected data than sources that are under the direct control of the clinician.

Possibly the most significant challenge in aggregating and analyzing big health care data is the amount of unstructured data. Structured or discrete data includes data that can be stored and retrieved in a relational database. Unstructured data in health care include: test results, scanned documents, images, and progress notes in the patients' EHR. Although standards such as the Clinical Documentation Architecture[9] allow interoperability and sharing of EHR data, the contents of the defined fields are often free text and therefore unstructured data. As free-text search software tools become more mature and natural language processing software is integrated into those tools, unstructured data will likely be one of the most valuable portions of the health care big data picture.

Certainly the federal government is encouraging the use of data to improve care with their EHR incentives through Stages 1 and 2 of Meaningful Use.[10] In order to qualify for incentive payments, eligible health care providers must meet a set of objectives that include patient safety and quality measurement. These financial incentives are intended to accelerate the use of data available in EHRs, but many providers struggle to meet the criteria and the value of the underlying data is questionable. The Government Accountability Office[11] recently released a report titled "Electronic Health Record Programs: Participation Has Increased, But Action is Needed to Achieve Goals, Including Improved Quality of Care" that recommends that a comprehensive strategy should be developed to ensure the reliability of data being collected and submitted to meet the meaningful use criteria.

One important challenge that must be acknowledged in health care data analytics is that the analysis is often a secondary use of the data. For instance, administrative data is collected primarily for the accounting of services rendered and the collection of payment. EHR data is primarily collected to track patient progress, treatment, and clinical status. When these data are then used to measure quality and outcomes, the original use of the data must be acknowledged as a potential limitation and may compromise the reliability and validity of any resulting models.

Comprehensive data and information governance programs may be used to address many of these challenges within and across providers.[12] A data governance program includes rules regarding data format and the appropriate use of data sources and data fields. Rigorous data governance policies ensure that the content and format of data is consistent and supports the technical aspects of mapping and combining data from various sources. An information governance program addresses the processing, analysis, and protection of the data. Information governance policies will

guide data users in determining whether or not a secondary use of the data is appropriate and also the level of detail that may be released while still protecting the patient's identity. In order to be most effective, data and information governance activities must be cross-departmental for data sets internal to one entity and cross-organizational for data sets that draw from multiple organizations. This type of structure will help break down both internal and external data silos.

Opportunities

Once these technological, legal, and philosophical challenges are solved, the large question is: how can the analysis of health care's big data be used to improve the delivery and efficiency of care delivery? Health care's big data is currently used to solve a number of operational and clinical issues, even in this imperfect state. Data analytics applications such as predictive modeling, population health, and quality measurement are all moving forward quickly.

Predictive modeling is currently used by the CMS and other health care payers for fraud prevention.[13] Predictive modeling uses statistical techniques and historical data to estimate the probability of future results. CMS contractors are using these techniques to determine which claims are likely to be fraudulent prior to the payment for the service. Traditionally, CMS detected fraud in Medicare claims by performing post-payment reviews via contractors such as recovery audit contractors. This pre-payment auditing is similar to the review activities used by credit card companies. Just as a bank may use a customer's spending profile to determine that it is unlikely they purchased a computer in Mexico when they reside in Ohio, a data-driven model will help CMS determine that it is unlikely that a podiatrist performed an angioplasty procedure. This proactive approach to fraud prevention is more effective than the previous method of pay and chase.

Predictive modeling may also be used to determine which patients are most likely to benefit from a care management plan. Care management plans are used to prevent hospitalizations for patients with chronic conditions such as diabetes, asthma, or chronic obstructive pulmonary disorder. Often the plans include contact with a health care professional to ensure that the patient is compliant with their medication plan and does not require further services that might prevent an expensive event such as an emergency department visit or inpatient admission. Predictive modeling identifies the high cost risk drivers and allows early intervention and patient management.[14]

The identification and tracking of patients with type 2 diabetes was discussed in a recent article in *Big Data*.[15] The author suggests using a two-step process to identify subsets of patients that have similar clinical indications and care patterns. First patients are divided into groups based on the primary diagnosis and then a statistical clustering method is applied to further divide the subsets. This method uses readily available administrative datasets, but patients must be tracked longitudinally to determine the treatment patterns. Therefore, the method is applicable in settings where patient level data is available over time and across providers.

Big data presents a tremendous opportunity for the measurement and reporting of quality in health care. The Health Quality Alliance[16] lists 70 regional quality measurement initiatives throughout the country. The projects are underway in 26 states and include both statewide and local efforts. CMS is committed to rewarding high quality care with their Medicare Hospital Value-Based Purchasing Program.[3] According to an article in the *National Law Review*,[17] all major commercial insurance companies in the US had started some level of value-based contracting with providers. Some are implementing these as pay-for-performance programs that seek to control unnecessary services as well as reward high quality care. All of these programs are heavily reliant on both administrative and abstracted health care data.

The CMS' Hospital Value-Based Purchasing Program includes a pay-for-performance component. In the inpatient setting, hospitals are rewarded when they achieve a high total performance score (TPS) compared to other hospitals in the country. They are also rewarded for improvement in their TPS over time. The TPS included three components in the US federal fiscal year 2014: clinical process of care measures (45%), patient experience of care (30%), and an outcome of mortality (25%).[3] The clinical process of care measures include indicators for using best practice guidelines for the treatment of acute myocardial infarction, heart failure, pneumonia, cardiology patients, and venous thromboembolism. This component also includes five measures regarding postoperative infection prevention. The TPS will add an efficiency measure and a number of patient safety measures in future years. As the number of measures grows, hospitals that do not have adequate EHRs and data integration will find it difficult to meet the measurement requirements.

Electronic medical record (EMR) data may also be used to study drug efficacy. Researchers at the University of Pennsylvania School of Medicine[18] compared the results of randomized controlled trials versus using an EMR to compare cardiovascular outcomes. Although observational studies using retrospective EMR data may not control for some covariates, they propose innovative methods to adjust for

some of these issues. They include the use of strict inclusion and exclusion criteria for the data as well as new statistical techniques. In the end, the authors found that the results of the EMR analysis matched the results of the randomized controlled trials for nine of the 17 outcomes. This is a promising result given the fact that the cost of randomized controlled trials is much higher than the cost of using readily available EMR data to compare treatment modalities.

The use of data-driven discoveries in therapeutics shows great promise. Personalized medical treatment protocols may be identified through the mining of large clinical databases.[19] Such analyses may result in the identification of patterns of side effects or even a potential new application of a current therapeutic agent.

Future directions

Health care data certainly meets the definition of big data. The challenges surrounding the full aggregation and use of health care data are not insurmountable. Meeting those challenges will require a culture shift in health care both internal to providers and between providers and other portions of the industry. The biggest challenge is determining the proper balance between protecting the patient's information and maintaining the integrity and usability of the data. Robust information and data governance programs will address a number of these challenges.

The sharing of data between organizations must be addressed before the full potential of big data in health care may be unlocked. The concept of the Learning Health System core values[5] may serve as a guiding concept for advancing the efforts to create a collection of health care data that may be used to realize the many opportunities outlined in this manuscript.

Recall that Gartner defines big data as "high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."[1] Health care data currently meets the "3Vs" of the big data definition. Health care spending represents 17% of the gross domestic product in the US.[20] Therefore, realizing the second portion of the Gartner definition of big data, namely "innovative forms of information processing for enhanced insight and decision making," will make a significant impact on not only the health care delivery system but the US as a whole.

Disclosure

The author reports no conflicts of interest in this work.

References

1. IT glossary: big data [webpage on the Internet]. Stamford, CT: Gartner; 2012 [cited May 25, 2012]. Available from: http://www.gartner.com/it-glossary/big-data/. Accessed June 17, 2014.
2. Crapo J. Big data in healthcare: separating the hype from the reality [webpage on the Internet]. Salt Lake City, UT: HealthCatalyst; 2014 [cited January 9, 2014]. Available from: http://www.healthcatalyst.com/healthcare-big-data-realities. Accessed June 17, 2014.
3. Hospital value-based purchasing [webpage on the Internet]. Baltimore, MD: Centers for Medicare and Medicaid Services; 2014 [updated May 18, 2014]. Available from: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/index.html. Accessed June 18, 2014.
4. Heatlh Research Institute. *Needles in a Haystack: Seeking Knowledge With Clinical Informatics*. New York, NY: Health Research Institue; 2012. Available from: http://www.pwc.com/mx/es/servicios-asesoria-negocios/archivo/2012-05-needles-in-a-haystack.pdf. Accessed September 4, 2014.
5. Rubin JC, Friedman CP. Weaving together a healthcare improvement tapestry. Learning health system brings together health data stakeholders to share knowledge and improve health. *J AHIMA*. 2014;85(5):38–43.
6. Olsen LA, Aisner D, McGinnis JM. *The Learning Healthcare System*. Washington, DC: National Academies Press; 2007.
7. Department of Health and Human Services. Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule. 45 CFR 160 and 164; 2013.
8. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule [webpage on the Internet]. Washington, DC: Department of Health and Human Services; 2013 [cited June 14, 2013]. Available from: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#guidancedetermination. Accessed June 14, 2014.
9. CDA® Release 2 [webpage on the Internet]. Ann Arbor, MI: Health Level Seven International; 2011 [cited December 12, 2011]. Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7. Accessed June 17, 2014.
10. Definition stage 1 of meaningful use [webpage on the Internet]. Baltimore, MD: Centers for Medicare and Medicaid Services; 2014 [updated July 18, 2014]. Available at: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Meaningful_Use.html. Accessed June 17, 2014.
11. US Government Accountability Office. *Electronic Health Record Programs: Participation Has Increased, but Action is Needed to Achieve Goals, Including Improved Quality of Care*. Washington, DC: US Government Accountability Office; 2014. Available from: http://www.gao.gov/assets/670/661399.pdf. Accessed September 4, 2014.
12. Knight K, Stainbrook C. *Benchmarking White Paper: 2014 Information Governance in Healthcare*. Minneapolis, MN: Cohasset Associates; 2014. Available from: http://www.ahima.org/~/media/AHIMA/Files/HIM-Trends/IG_Benchmarking.ashx. Accessed September 4, 2014.
13. White SE. Predictive modeling 101. How CMS's newest fraud prevention tool works and what it means for providers. *J AHIMA*. 2011;82(9):46–47.
14. Cousins MS, Shickle LM, Bander JA. An introduction to predictive modeling for disease management risk stratification. *Dis Manag*. 2002;5(3):157–167.
15. Bradley P. Implications of big data analytics on population health management. *Big Data*. 2013;1(3):152–159.
16. Intiatives to measure and report on quality [webpage on the Internet]. Washington, DC: Quality Alliance Steering Committee; 2009 [cited March 20, 2009]. Available from: http://www.healthqualityalliance.org/initiatives. Accessed June 18, 2014.

17. Hintz JE, O'Connor MC. First steps in the era of value-based health care purchasing. Western Springs: National Law Review; 2012 [cited June 11, 2012]. Available at: http://www.natlawreview.com/article/first-steps-era-value-based-health-care-purchasing. Accessed June 18, 2014.
18. Tannen RL, Weiner MG, 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. *BMJ*. 2009;338:b81.
19. Issa NT, Byers SW, Dakshanamurthy S. Big data: the next frontier for innovation in therapeutics and healthcare. *Expert Rev Clin Pharmacol*. 2014;7(3):293–298.
20. White S. *Principles of Finance for Health Information and Informatics Professionals*. Chicago, IL: AHIMA Press; 2012.