REVIEW

Big Data Bigger Changes: The State of Big Data in Medicine

Eitan Fleischman
Sackler School of Medicine, Tel Aviv University, Tel Aviv

Abstract

As the costs of healthcare increase exponentially and access to healthcare declines, big data plays a pivotal role in remedying both situations. Big data is defined as extremely large data sets analyzed for trends and patterns. It provides the tools to both reduce costs in the healthcare system by reducing waste and hospital stays, while also providing the means to devise more effective, direct, and optimized quality treatment. This article outlines the state of big data in medicine and discusses current and potential uses of this tool.

Micah Belzberg: *Big Data*

Introduction

Currently, healthcare in the United States represents 17.6% of the United States gross domestic product. This equates to nearly $600 billion dollars over the expected spending in countries of similar size and economic distribution, according to analysts at McKinsey (1). Increasing healthcare costs can be attributed to both healthcare suppliers and patients. Hospitals and other medical suppliers have contributed through wasteful spending, rising prescription drug costs, high administrative costs, consolidation of healthcare systems, and premature implementation of new technologies (2). The patient contribution consists mainly of increased service utilization due to an aging population coupled with increasingly unhealthy lifestyle choices.

Due to increasing healthcare costs, the US government and other payers have begun transitioning away from a fee-for-fee system to an outcomes and value based payment system (3). In the old model, physicians and hospitals were compensated for tests and procedures leading to a positive incentive to increase medical services with a lesser emphasis on patient outcomes. Physicians were reimbursed for the cost of the specific services provided to the patient at the time of visit and were not rewarded for improving patient health. In a value-based payment system, insurance payers reimburse providers for preventative measures and reduced readmission rates. Primary care has become a cornerstone of the new model with an emphasis on the physicians' role in keeping patients out of the hospital and promoting healthier lifestyles for their patients on a proactive basis. By focusing on preventive measures, healthcare providers may be able to avoid the need for more expensive and time consuming procedures through utilization of lower cost preventive visits with prescreening and early intervention.

Healthcare providers are looking to utilize big data to augment current preventive care techniques and aid in reducing wasteful spending (4). Big data is defined as data that are exceedingly complex and large, and cannot be processed and managed by traditional processing means (5-6). Large computer frameworks are required to process big data sets with the goal of identifying trends and optimization parameters in a given field. Big data has been described using the 3V's of data: variety, velocity, and volume. Variety denotes big data's vast composites of different data types. Velocity represents the speed of data acquisition. Big

data can be gathered and compiled retroactively, but the majority of data is garnered in real time. Lastly, volume speaks to big data's immense size. Analysts believe that the US could save more than $300 billion each year by integrating big data product into the healthcare system (5-6). In the field of medicine, there are three types of big data: "omics", healthcare, and social.

Data Types

"Omics"

"Omics" data, short for genomics data, is large data sets composed from the results of different genomic modalities such as high-throughput sequencing, NGS (next generation sequencing), and mass spectrometry (7). One of the main uses of "omics" data is to gather information regarding gene biomarkers and other genetic data. SNPs, deletions, as well as epigenetic information present with serious phenotypical differences in the patient population. Genomes consist of 30,000-35,000 unique genes without any forms of variation (5). This presents the potential for an immense multivariable data set requiring big data processing techniques to analyze and understand their results.

The first major challenge associated with gathering and analyzing "omics" data is the difference in the velocity of data collection. As presented in Wu P-Y et al., the frequency of acquisition changes with modality type (7). For instance, a genome requires only one sampling, while other tissue typing may vary with environment, pathology, or patient compliance and may require multiple acquisition periods. This sampling inequality may result in potential data contamination and quality deterioration. The second challenge inherent in big data across the statistics gamut is how to analyze such large quantities of data. "Omics" data presents a good example of this issue as some data sets include more than 104 distinct variables.

Ultimately, researchers and biostatisticians are bypassing the major challenges with "omics" data through the use of distributed computer platforms, such as Apache Hadoop for storage, and Cloud Based Computing outlets, such as Amazon EC2 (7). These powerful computer programs allow researchers to sort through and categorize big data sets for further analyzation. Researchers are also integrating data sets to create larger networks for analysis (8). For instance, Brown et al., utilizes large integrative networks to better understand tumor microenvironments (TME). Determining and understanding the TME will allow physicians to create focused and personalized tumor treatments.

Healthcare

Healthcare data consist mainly of electronic health records (EHR) (7). This data is traditionally viewed as the "patient file", albeit an electronic version of what was once paper. An EHR is composed of physician notes, patient history, diagnosis codes, administrative data, and pharmaceutical history that are stored electronically. The EHRs are stored in specific EHR system programs such as Athenaclinicals, EpicCare, or Allscripts for security and accessibility.

Healthcare data shares the same constraints with "omics" data—its speed and collection methodology are irregular (7). Electronic healthcare data also contains user errors, missing data, and incongruities in terminology. Lastly, healthcare data contains the most patient specific information of all three of the data types. It is fairly straightforward to view a patient's medical history. However, due to its patient sensitive nature, healthcare data must conform to strict privacy legislation such as HIPAA.

Healthcare data can be used for patient profiling analytics (9). A patient's medical health record can be run through advanced analytics to determine if the patient is at high risk for future medical complications and, along with the physician, create proactive and preventative lifestyle changes and preventive care protocols. EHRs can also be used to create composite medical profiles, where a physician can use other patients' profiles to develop treatment plans for similar status patients. Patient record de-identification may be a potential solution to maintain patient anonymity and privacy (10). During a de-identification, the analyst replaces all major identifiers (which are dictated by the local review board) with non-specific numbers or characters. This way data may be shared, but the key patient identifiers are masked.

Key Point: Precision Oncology and Genomic Data Commons (GDC):

In oncology field, various organizations, including the National Cancer Institute, have developed an information system called the NCI GDC to collect raw genomic data as well as diagnostic, histologic, and clinical outcome data from NCI-funded projects such as the Cancer Genome Atlas (TCGA) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program. One of the goals is that the GDC could recognize patients with rare molecular subtypes of cancer who could be contacted for potential participation in clinical trials appropriate for their cancer.

Reference: Grossman RL, Heath AP, Ferretti V, et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med 2016; 375:1109-1112.

Social

Social data is the newest form of big data being incorporated into the healthcare field. It harnesses data from large web-based applications and personal user devices to gather medical trends and data (11). Data is congregated from online applications such as Google Trends, Twitter, and Facebook, as well as wearable devices such as fitness trackers, step counters and other self-reporting devices and sensors.

Similar to "omics data", social data is subject to complications of data pollution and size. It has been estimated that 2.5 quintillion bytes of social data are created each day (11). It is a lower quality data because it is unstructured and the data producers are not vetted data analysts. Tweets, Facebook posts, and chatrooms express opinion but aren't necessarily factually accurate. Compared to data collected by researchers in individual studies and questionnaires, social big data may be viewed as unscientific and unreliable. Additionally, the data isn't inherently medically related and must be combed through and parsed to find relevant medical data.

Despite its drawbacks, social data represents the most attainable source of data in medicine. It is not restricted by healthcare privacy parameters like EMR and "omics" data, and, as the largest data of the three subcategories, is utilized in machine learning algorithms (11). Machine-learning algorithms find data cluster correlations and devise hypotheses. With larger data sets, machine algorithms generate more hypotheses and have higher confidence than with smaller data sets. Social data isn't inherently healthcare oriented, but it can be used to detect and identify medical trends. In 2004 early social data indexed from Chinese press reports predicted the acute respiratory syndrome epidemic (12). While the algorithm was unrefined, it showed that social data may potentially inform the medical world.

Applications of Big Data

Big data has the ability to improve quality of care, healthcare access, more efficiently distributing finite medical services while reducing costs in an overly expensive field. Big data achieves this by augmenting current prescreening protocols, reducing readmissions, optimizing spending, and reducing fraud. The emphasis on preventive medicine and the subsequent redistribution of other services results in improved patient satisfaction and increased patient access to medical services (3). We will discuss archetypal examples of big data implementations at different stages of a patient's time course and finish the discussion with the economic implications of big data integration into healthcare practice.

Patient prescreening can exceed its current utilization boundaries through incorporation of big data trend analysis and modeling. Using social data, a physician could analyze current medical web searches, stratifying for location, population, and age group, to begin forming an idea of possible questions or problems related to an incoming patient. A careful analysis of social data products led researchers to determine a correlation between a silicosis outbreak and the public's silicosis related searches and social media (12). They proposed using Google Trends, among other social media outlets, to aid physicians in proactively approaching patients during physical exams and meetings with questions or concerns they might have, thus improving quality and accuracy of

care. Familiarity with population clusters' concerns and risk factors will allow physicians to more quickly and accurately address a patient's concerns.

"Omic" data in combination with patient prescreening can also be used to increase accuracy and speed of patient diagnosis. Columbia University Medical Center utilizes big data advanced analytics to diagnose brain aneurysm injuries faster than current protocols allow (5). Utilizing physiological data linked to a patient's brain injuries allows physicians to diagnoses serious brain complications 48 hours faster than previous diagnoses. This allows physicians to avoid serious complications and death in many of their aneurysm patients. Blue Shield of California is developing an integrated big data system collaboration between doctors, hospitals, and healthcare plans which delivers evidence-based diagnostic recommendations to physicians, improving diagnostic precision (1). The evidence based diagnostic recommendations are built from "omic" and healthcare big data sets analyzed for trends in diagnoses which can be applied to new patients. With this analysis information in hand, Blue Shield doctors will be able to more accurately and efficiently diagnose and treat patients in a shorter period of time, resulting in shorter hospital stays.

Eliminating monetary inefficiencies also contributes to optimized patient care. Earlier release, faster diagnoses, and elimination of service redundancies allow for an optimized redistribution of finite medical resources. In 2012, the Minnesota Department of Health conducted a study to analyze the state's hospital admissions, readmissions, and emergency room visits to determine if there were any "preventable events" (14). They utilized existing claims data in combination with big data analytics to determine that 1.3 million patient encounters costing approximately $2 billion qualified as "preventable" hospital visits. Approximately two-thirds of the hospital visits may have been preventable if patients had been treated via primary care, been given more medical and hospital information, and/or if there had been better coordination between physicians, social services, and the patients' families. Beyond reducing costs, reduction of readmissions and hospital stays leads to a direct reduction in postoperative complications and comorbidities (15).

Eliminating fraud in the reimbursement system is an integral part of reducing healthcare costs. The National Health Care Anti-Fraud Associated estimates that 3% of healthcare spending is lost to healthcare fraud (16). These numbers constitute a massive hemorrhaging of healthcare funds, increasing costs and lowering reimbursement pools. The Centers for Medicare and Medicaid Services (CMS), representing the largest healthcare provider in the United States, has over the past five years implemented big data fraud prevention services (FPS) to combat fraud, waste, and abuse (17). The big data FPS predictive algorithms have already saved the US government over $1 billion and represent a return on investment of $11.60 per federal tax dollar invested in the big data asset.

Challenges

Currently big data's main challenge is the lack of a user friendly architectural framework (9). All the promise and potential discussed above is hindered by lack of accessibility for the lay physician, nurse, or hospital administrator. Current programs require a deep understanding of machine learning and computer sciences. In the future, developers need to create large, accessible databanks that both retain patient data security while simultaneously creating an intuitive and standardized program that contains a full suite of analytical capabilities.

Another major challenge is working with strict privacy restrictions on medical data. HIPAA regulations include mainly "omic" and health data, although recently social data has been questioned as well (11, 18). HIPAA regulations currently make it difficult for hospitals to share data with each other. Hospitals need to focus on the transmission and sharing of de-identified data with other healthcare providers. Without shared data, smaller healthcare networks will have a harder time developing the care improving, cost saving big data tools. The US Government has realized big data's current restrictions, and have begun implementing programs like the NIH Big Data to Knowledge Program (BD2K) to support big data sharing and ease of access (11).

Conclusions

Big data represents a significantly impactful technology in the world of medicine. This article discussed the background of big data, several relevant applications, challenges, and opportunities for future use. Although in its infancy, big data has the capabilities to change

the healthcare arena through increased efficiency and decreased costs. If analysts are correct, big data represents a $300 billion a year infusion of funds into the healthcare system. Implementing big data translates into direct healthcare solutions for various populations by reallocating limited medical services and improving healthcare outcomes for all.

References

1. Kayyali B, Knott D, Kuiken SV. The big-data revolution in US health care: Accelerating value and innovation | McKinsey & Company. http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care. Accessed November 6, 2016.
2. Mack M. What Drives Rising Healthcare Costs? http://www.gfoa.org/sites/default/files/GFR081626.pdf. Accessed March 4, 2017.
3. Shrank WH. Primary Care Practice Transformation and the Rise of Consumerism. J Gen Intern Med. February 2017. doi:10.1007/s11606-016-3946-1.
4. Dewdney SB, Lachance J. Electronic Records, Registries, and the Development of "Big Data": Crowd-Sourcing Quality toward Knowledge. Front Oncol. 2017;6. doi:10.3389/fonc.2016.00268.
5. Belle A, Thiagarajan R, Soroushmehr SMR, et al. Big Data Analytics in Healthcare. BioMed Res Int. 2015;2015:370194. doi:10.1155/2015/370194.
6. Tan SS-L, Gao G, Koch S. Big Data and Analytics in Healthcare. Methods Inf Med. 2015;54(6):546-547. doi:10.3414/ME15-06-1001.
7. Wu P-Y, Cheng C-W, Kaddi C, et al. Advanced Big Data Analytics for -Omic Data and Electronic Health Records: Toward Precision Medicine. IEEE Trans Biomed Eng. October 2016. doi:10.1109/TBME.2016.2573285.
8. Brown JAL, Ni Chonghaile T, Matchett KB, et al. Big Data-Led Cancer Research, Application, and Insights. Cancer Res. 2016;76(21):6167-6170. doi:10.1158/0008-5472.CAN-16-0860.
9. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;2:3. doi:10.1186/2047-2501-2-3.
10. Ross MK, Wei W, Ohno-Machado L. "Big Data" and the Electronic Health Record. Yearb Med Inform. 2014;9(1):97-104. doi:10.15265/IY-2014-0003.
11. Hansen MM, Miron-Shatz T, Lau AYS, Paton C. Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives. Contribution of the IMIA Social Media Working Group. Yearb Med Inform. 2014;9:21-26. doi:10.15265/IY-2014-0004.
12. Sommer A. The Utility of "Big Data" and Social Media for Anticipating, Preventing, and Treating Disease. JAMA Ophthalmol. 2016;134(9):1030-1031. doi:10.1001/jamaophthalmol.2016.2287.
13. Bragazzi NL, Dini G, Toletone A, et al. Leveraging Big Data for Exploring Occupational Diseases-Related Interest at the Level of Scientific Community, Media Coverage and Novel Data Streams: The Example of Silicosis as a Pilot Study. PloS One. 2016;11(11):e0166051. doi:10.1371/journal.pone.0166051.
14. News release: Novel MDH study yields first statewide estimate of potentially preventable health care events. http://www.health.state.mn.us/news/pressrel/2015/hcevents.html. Accessed December 26, 2016.
15. Sweeney JF. Postoperative Complications and Hospital Readmissions in Surgical Patients. Ann Surg. 2013;258(1):19-20. doi:10.1097/SLA.0b013e318297a37e.
16. Simborg DW. Healthcare Fraud: Whose Problem is it Anyway? J Am Med Inform Assoc JAMIA. 2008;15(3):278-280. doi:10.1197/jamia.M2672.
17. Medicare's "Big Data" Tools Fight & Prevent Fraud to Yield Over $1.5 Billion in Savings. CMS Blog. May 2016. https://blog.cms.gov/2016/05/27/medicares-big-data-tools-fight-prevent-fraud-to-yield-over-1-5-billion-in-savings/. Accessed December 25, 2016.
18. Health Big Data Recommendations. Privacy and Security Workgroup (PSWG) of the Health Information Technology Policy Committee (HITPC) (Aug. 2015), at p. 3. Available online at: http://www.healthit.gov/sites/faca/HITPC_Health_Big_Data_Report_FINAL.pdf. Accessed January 3, 2016.