Bigdata in Healthcare

Fatahiyya Ali Lawal
Department of Computer Science
SRM University Kattankulathur Chennai

***Abstract*— The healthcare sector historically has generated large amounts of data over the years, driven by record keeping, compliance & regulatory requirements, and patient care. While most of the data is stored in hard copy form, the current trend is toward making these large amounts of data in digitised form. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery meanwhile reducing the costs, these massive quantities of data (known as 'big data') supports a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management . This paper provides an overview of big data analytics for researchers and practitioners in the healthcare sector.**

Keywords*— *Big Data, Electronic HealthRecord,The 4v's of Big Data, Data Science, Data Science Framework.

I. INTRODUCTION

What Is Big Data? By Edd Dumbill

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it. The hot IT buzzword of 2012, big data has become viable as cost-effective approaches have emerged to tame the volume, velocity, and variability of massive data. Within this data lie valuable patterns and information, previously hidden because of the amount of work required to extract them. To leading corporations, such as Wal-Mart or Google, this power has been in reach for some time, but at fantastic cost. Today's commodity hardware, cloud architectures and open source software bring big data processing into the reach of the less well-resourced. Big data processing is eminently feasible for even the small garage startups, who can cheaply rent server time in the cloud. The value of big data to an organization falls into two categories: analytical use and enabling new products. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analyzing shoppers' transactions and social and geographical data. Being able to process every item of data in reasonable time removes the troublesome need for sampling and promotes an investigative approach to data, in contrast to the somewhat static nature of running predetermined reports.

Fig1.1 Big Data

II. EVOLUTION OF THE ELECTRONIC HEALTH RECORD

EHRs have typically contained "hard" objective data in digital form from the analysis of lab tests, including numerical (both discrete and continuous) and categorical valuations (e.g., normal, above normal, below normal). More recently, providers are collecting many more structured attributes, such as vital signs, demographic information, and prescription histories, in-patient EHRs and are tracking these data over time. These structured data are often supplemented with unstructured (freeform) text, such as notes from clinicians, hospital physicians and nurses, and patients' primary care physicians and specialists. EHRs also routinely include unstructured images, such as X-rays and MRI scans, and genomic data, all of which are examined by specialists, who append summary notes to the EHR. We anticipate that EHRs will further expand in scope to include content from a larger team of providers, including in-home caregivers, community social workers, family members, and the patients themselves.

A. On the Nature of Raw Data

Opportunities to derive value from data depend to some extent on the nature of the raw data themselves. A dataset and the individual records within it can be characterized in several ways, including volume, velocity, and variety — the "3Vs"7 model commonly used to describe big data.

B. Types

The data in a record or dataset may be text strings or numeric values of various kinds, such as real or imaginary numbers, expressed separately or in arrays or matrices. Data

arrays may be indexed or timesequenced to form ordered data such as image, video, and audio records.
Structure Any record or data element in a record may be unrelated to others (unstructured data), related in a fixed and formally defined way such as a hierarchy or relational linkage (structured data), or related in ways specific to the individual record or dataset (semistructured data).

- ***Volume***

Volume refers both to the number of records in a dataset and the size of an individual record. Roughly speaking, small datasets are those that can be managed, accessed, and analyzed with traditional file-based and commodity applications. Medium datasets require well-understood technologies, such as relational database management systems.
Large datasets need specialized data management tools for analysis, storage and maintenance that must coordinate across processors and storage platforms. Note that the actual sizes of "small", "medium", and "large" datasets continue to change as data storage and retrieval technology advances.

- ***Velocity***

Velocity refers to how often new data arrive for collection, storage, and analysis, as well as how much data arrive at any one instant. High-velocity applications such as retail purchase transactions may require specialized tools to adequately handle incoming floods of data.

- **Variety**

Variety describes both the number of datasets and the number of data types employed in analytics. Aggregating, aligning, linking, and correlating information across highly diverse datasets and data types is an ongoing challenge in data science for all applications, including emerging technologies to address the alignment of patient data across multiple EHRs.

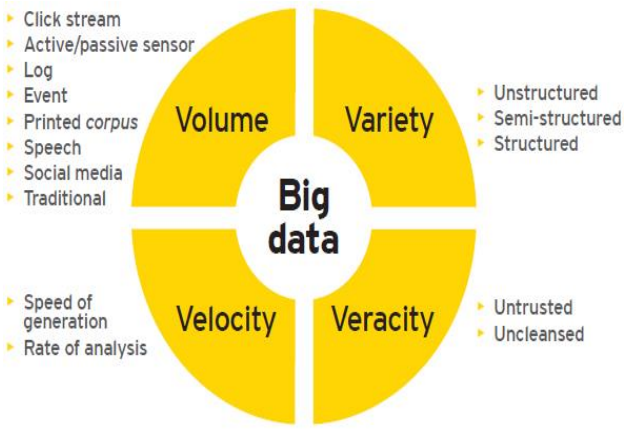

Fig 2.1 the four V's of Big data.

- ***Completeness Richness, and Balance***

These factors indicate how fully the dataset represents a desired population, activity, or other variable of interest. Completeness is the fraction of a given population represented within the dataset. Richness refers to the amount of data available for each entity (e.g., an individual or organization). Balance is a measure of how prevalent a given phenomenon is (or is estimated to be) in the dataset; for example, in a healthcare analytics context, a dataset including all in-hospital patients would have a different readmission rate than a dataset of patients with chronic diseases, since co morbidities are presumably more prevalent in the latter dataset.

- ***Labeling***

Labeling indicates to what extent objects, events, and other phenomena of interest are explicitly marked, e.g., an X-ray image with notes attached from a radiologist indicating whether a tumor was malignant or benign.

III. ADVANTAGES OF DATA SCIENCE IN HEALTH CARE

How is data science transforming health care? There are many ways in which health care is changing, and needs to change. We have reached a point at which our need to understand treatment effectiveness has become vital — to the health care system and to the health and sustainability of the economy overall. Why do we believe that data science has the potential to revolutionize health care? After all, the medical industry has had data for generations: clinical studies, insurance data, hospital records. But the health care industry is now awash in data in a way that it has never been before: from biological data such as gene expression, next-generation DNA sequence data, proteomics, and metabolomics, to clinical data and health outcomes data contained in ever more prevalent electronic health records (EHRs) and longitudinal drug and medical claims. We have entered a new era in which we can work on massive datasets effectively, combining data from clinical trials and direct observation by practicing physicians.

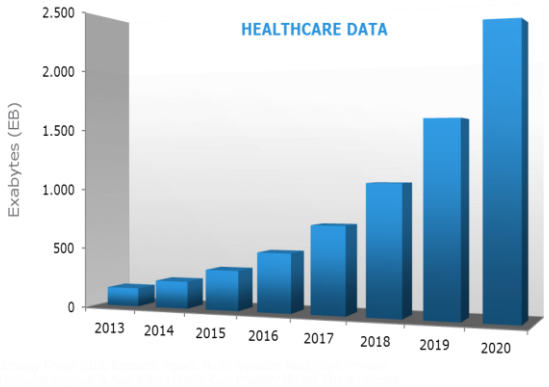

Fig 3.1 Health care Data Chart

When we combine data with the resources needed to work on the data, we can start asking the important questions about what treatments work and for whom. The opportunities are huge: for entrepreneurs and data scientists looking to put their skills to work disrupting a large market, for researchers trying to make sense out of the flood of data they are now generating, and for existing companies (including health insurance companies, biotech, pharmaceutical, and medical device companies, hospitals and other care providers) that are looking to remake their businesses for the coming world of

outcome-based payment models. Making Health Care More Effective What, specifically, does data allow us to do that we couldn't do before? For the past 60 or so years of medical history, we've treated patients as some sort of an average. A doctor would diagnose a condition and recommend a treatment based on what worked for most people, as reflected in large clinical studies. Over the years, we've become more sophisticated about what that average patient means, but that same statistical approach didn't allow for differences between patients. A treatment was deemed effective or ineffective, safe or unsafe, based on double-blind studies that rarely took into account the differences between patients. With the data that's now available, we can go much further.

Enabling Data science is not optional in health care reform; it is the linchpin of the whole process. Data doesn't help if it can't be moved, if data sources can't be combined. There are two big issues here. First, a surprising number of medical records are still either hand-written, or in digital formats that are scarcely better than hand-written (for example, scanned images of hand-written records). Getting medical records into a format that's computable is a prerequisite for almost any kind of progress. Second, we need to break down those silos. Anyone who has worked with data knows that, in any problem, 90% of the work is getting the data in a form in which it can be used; the analysis itself is often simple. We need electronic health records: patient data in a more-or-less standard form that can be shared efficiently, data that can be moved from one location to another at the speed of the Internet. Not all data formats are created equal, and some are certainly better than others: but at this point, any machine-readable format, even simple text files, is better than nothing. While there are currently hundreds of different formats for electronic health records, the fact that they're electronic means that they can be converted from one form into another. Standardizing on a single format would make things much easier, but just getting the data into some electronic form, any, is the first step. Once we have electronic health records, we can link doctor's offices, labs, hospitals, and insurers into a data network, so that all patient data is immediately stored in a data center: every prescription, every procedure, and whether that treatment was effective or not. This isn't some futuristic dream; it's technology we have now. Building this network would be substantially simpler and cheaper than building the networks and data centers Electronic health records enable us to go far beyond the current mechanism of clinical trials. In the past, once a drug has been approved in trials, that's effectively the end of the story: running more tests to determine whether it's effective in practice would be a huge expense. A physician might get a sense for whether any treatment worked, but that evidence is essentially anecdotal: it's easy to believe that something is effective because that's what you want to see. And if it's shared with other doctors, it's shared while chatting at a medical convention. But with electronic health records, it's possible (and not even terribly expensive) to collect documentation from thousands of physicians treating millions of patients. We can find out when and where a drug was prescribed, why, and whether there was a good outcome. We can ask questions that are never part of clinical trials: is the medication used in combination with anything else? What other conditions is the patient being treated for? We can use machine learning techniques to discover unexpected combinations of drugs that work well together, or to predict adverse reactions. We're no longer limited by clinical trials; every patient can be part of an ongoing evaluation of whether his treatment is effective, and under what conditions.

Technically, this isn't hard. The only difficult part is getting the data to move, getting data in a form where it's easily transferred from the doctor's office to analytics centers. To solve problems of hot-spotting (individual patients or groups of patients consuming inordinate medical resources) requires a different combination of information. You can't locate hot spots if you don't have physical addresses. Physical addresses can be geocoded (converted from addresses to longitude and latitude, which is more useful for mapping problems) easily enough, once you have them, but you need access to patient records from all the hospitals operating in the area under study. And you need access to insurance records to determine how much health care patients are requiring, and to evaluate whether special interventions for these patients are effective. Not only does this require electronic records, it requires cooperation across different organizations (breaking down silos), and assurance that the data won't be misused (patient privacy). Again, the enabling factor is our ability to combine data from different sources; once you have the data, the solutions come easily. Breaking down silos has a lot to do with aligning incentives. Currently, hospitals are trying to optimize their income from medical treatments, while insurance companies are trying to optimize their income by minimizing payments, and doctors are just trying to keep their heads above water.

IV. THE 4 "VS" OF BIG DATA ANALYTICS IN HEALTHCARE

Like big data in healthcare, the analytics associated with big data is described by three primary characteristics: volume, velocity and variety (http://www-01.ibm.com/software/data/bigdata/). Over time, health-related data will be created and accumulated continuously, resulting in an incredible volume of data. The already daunting volume of existing healthcare data includes personal medical records, radiology images, clinical trial data FDA submissions, human genetics and population data genomic sequences, etc. Newer forms of big data, such as 3D imaging, genomics and biometric sensor readings, are also fueling this exponential growth.

Fortunately, advances in data management, particularly virtualization and cloud computing, are facilitating the development of platforms for more effective capture, storage and manipulation of large volumes of data . Data is accumulated in real-time and at a rapid pace, or velocity. The constant flow of new data accumulating at unprecedented rates presents new challenges. Just as the volume and variety of data that is collected and stored has changed, so too has the velocity at which it is generated and that is necessary for retrieving, analyzing, comparing and making decisions based on the output.

Most healthcare data has been traditionally static—paper files, x-ray films, and scripts. Velocity of mounting data

increases with data that represents regular monitoring, such as multiple daily diabetic glucose measurements (or more continuous control by insulin pumps), blood pressure readings, and EKGs. Meanwhile, in many medical situations, constant real-time data (trauma monitoring for blood pressure, operating room monitors for anesthesia, bedside heart monitors, etc.) can mean the difference between life and death.

Future applications of real-time data, such as detecting infections as early as possible, identifying them swiftly and applying the right treatments (not just broad-spectrum antibiotics) could reduce patient morbidity and mortality and even prevent hospital outbreaks. Already, real-time streaming data monitors neonates in the ICU, catching life-threatening infections sooner. The ability to perform real-time analytics against such high-volume data in motion and across all specialties would revolutionize healthcare. Therein lies variety. As the nature of health data has evolved, so too have analytics techniques scaled up to the complex and sophisticated analytics necessary to accommodate volume, velocity and variety. Gone are the days of data collected exclusively in electronic health records and other structured formats. Increasingly, the data is in multimedia format and unstructured. The enormous variety of data— structured, unstructured and semi-structured—is a dimension that makes healthcare data both interesting and challenging.

Structured data is data that can be easily stored, queried, recalled, analyzed and manipulated by machine. Historically, in healthcare, structured and semi-structured data includes instrument readings and data generated by the ongoing conversion of paper records to electronic health and medical records. Historically, the point of care generated unstructured data: office medical records, handwritten nurse and doctor notes, hospital admission and discharge records, paper prescriptions, radiograph films, MRI, CT and other images.

Already, new data streams—structured and unstructured—are cascading into the healthcare realm from fitness devices, genetics and genomics, social media research and other sources. But relatively little of this data can presently be captured, stored and organized so that it can be manipulated by computers and analyzed for useful information. Healthcare applications in particular need more efficient ways to combine and convert varieties of data including automating conversion from structured to unstructured data.

The structured data in EMRs and EHRs include familiar input record fields such as patient name, data of birth, address, physician's name, hospital name and address, treatment reimbursement codes, and other information easily coded into and handled by automated databases. The need to field-code data at the point of care for electronic handling is a major barrier to acceptance of EMRs by physicians and nurses, who lose the natural language ease of entry and understanding that handwritten notes provide. On the other hand, most providers agree that an easy way to reduce prescription errors is to use digital entries rather than handwritten scripts. The potential of big data in healthcare lies in combining traditional data with new forms of data, both individually and on a population level. We are already seeing data sets from a multitude of sources support faster and more reliable research and discovery. If, for example, pharmaceutical developers could integrate population clinical data sets with genomics data, this development could facilitate those developers gaining approvals on more and better drug therapies more quickly than in the past and, more importantly, expedite distribution to the right patients . The prospects for all areas of healthcare are infinite.

Veracity assumes the simultaneous scaling up in granulityity and performance of the architectures and platforms, algorithms, methodologies and tools to match the demands of big data. The analytics architectures and tools for structured and unstructured big data are very different from traditional business intelligence (BI) tools. They are necessarily of industrial strength. For example, big data analytics in healthcare would be executed in distributed processing across several servers ("nodes"), utilizing the paradigm of parallel computing and 'divide and process' approach. Likewise, models and techniques—such as data mining and statistical approaches, algorithms, visualization techniques—need to take into account the characteristics of big data analytics. Traditional data management assumes that the warehoused data is certain, clean, and precise.

V. DATA SCIENCE FRAMEWORK AND PROCESS

This framework comprises four fundamental stages, each of which is underpinned with a firm understanding of the business or clinical problem and healthcare delivery environment, though they may not proceed entirely linearly. It is not uncommon to proceed through each stage to establish a basic data science workflow, then revisit earlier stages to accommodate shifts in the problem definition or environment resources, new data sources, additional analyses, and explore new information.

- ***Data Collection***

Data collection encompasses the activities from the moment raw data are gathered from their originating sources to the time of their placement (possibly in a transformed state) in repositories appropriate for storage, retrieval, and analysis. It also includes the planning and systems required for ongoing management and security of the collected data.

- ***Data Curation***

Data curation, where data are prepared for meaningful analysis, is perhaps the most important step in determining whether a data science project has the potential to produce measurable business value. This stage includes, but is much broader than, the traditional extraction, transformation, and load (ETL) methods of both commodity and custom-developed big data management systems. Specifically, the first phase of data curation involves standardizing the raw data to uniform specifications. This typically involves data cleansing, in which potential errors (e.g., data corrupted in collection, data values outside meaningful or permissible ranges) are repaired and/or removed from the dataset, as well as reformatting and conversion of data values to standard units and formats. It may also include transformations such as conversion of audio files to text transcripts.

The second phase of data curation prepares standardized data for analysis. Data of different types and from different sources may be merged, aligned and linked, or aggregated into new records specific to an activity, process, or organization. Large datasets may be sampled to create smaller datasets that are more amenable to automated analysis and/ or to correct statistical imbalances in the full dataset. Sampling may be inappropriate for some analysis activities, such as rare event or anomaly detection.

- ***Data Analysis***

This stage includes identification, selection, and execution of algorithmic and computational models and methods. It typically involves extension and novel integration of existing analytical approaches rather than the creation of new analytical methods. Successful data analysis solutions do not take a "kitchen sink" approach of applying analysis methods indiscriminately and hoping for good outcomes. Instead, they are driven by the organization's need for insightful and actionable intelligence and focus on producing relevant and accurate evidence informing such action.

VI. CONCLUSIONS

Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. In the future we'll see the rapid, widespread implementation and use of big data analytics across the healthcare organization and the healthcare industry. To that end, the several challenges highlighted above, must be addressed. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. Big data analytics and applications in healthcare are at a nascent stage of development, but rapid advances in platforms and tools can accelerate their maturing process.

REFRENCES

[1] A community white paper developed by leading researchers across the United States.Challanges and opportunities with Big Data white paper 2012. https://www.purdue.edu/discoverypark/cyber/assets/pdfs/BigDataWhitePaper.pdf.
[2] Hitachi Data Systems. Build the future of Bug Data today white paper November 2013.
[3] Michael Driscoll Big Data Now, Ed. O'Reilly, USA 2012.
[4] What is Big Data .http://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data.
[5] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In SODA, 2013.
[6] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data, Commun. ACM 39 (11) (1996) 27–34.
[7] Big data analytics in healthcare: promise and potential http://www.fordhamcdt.org/pdf/2047-2501-2-3.pdf
[8] Data Science in Health care https://www.leidos.com/sites/default/files/Leidos%20WP%20-%20Data%20Science%20Solutions%20for%20Health%201.15.pdf
[9] O'Reilly Media "Big Data Now" 2012 Edition. . Chapter 6 Page 83-90
[10] O'Reilly Media "Big Data Now"2014 Edition. Page 47-55 Current Perspectives from O'Reilly Media
[11] http://www.goodreads.com/book/show/12491550-big-data-now
[12] Canadian Medical Protective Association, The impact of Big Data on HealthCare and Medical Practice https://www.cmpa-acpm.ca/documents/10179/301372750/com_14_big_data_design-e.pdf