**Introduction to Medical Big data Analytics**

**Introduction**

It all began seventy years ago when the first attempts to quantify the growth rate in the *volume of data* or what has popularly been known as the “information explosion” was predicted. In this section we will be discussing about the history of big data, challenges faced, its needs in today’s world and the various opportunities it gives us to manage big data in various applications. It also discusses the major milestones in the history of sizing data volumes in the evolution of the idea of “big data” and observations pertaining to data or information explosion.

**Big Data over internet**

**1.1.1 Age of Big Data**

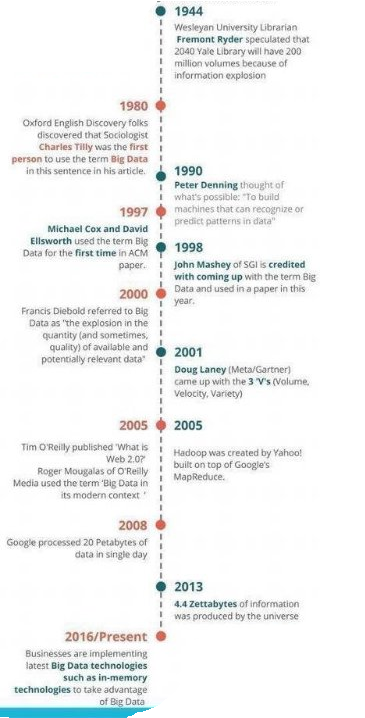
Billions of people use internet everyday .As a result super computers sites and large data centers must provide high–performance computing services to huge number of internet users concurrently. Data centers have to be upgraded using fast servers, storage systems and high-bandwidth networks. The purpose is to facilitate storage for such large data stored into various data centers.

**1.1.1.1 The Platform Evolution (publications)**

Scientists long before, estimated that data would rapidly grow and there was a need to find a way to manage growth of such data .This marked the beginning for an extensive research that was carried on to study Big data.

The year 1944, marked the beginning of the study of Big data by Fremont Rider , Wesleyan University Librarian, who published “ [*The Scholar and the Future of the Research Library*](http://www.amazon.com/Scholar-Research-Library-Problem-Solution/dp/B005ZN4N8C/ref=sr_1_1?s=books&ie=UTF8&qid=1339603976&sr=1-1&keywords=The+Scholar+and+the+Future+of+the+Research+Library)” where he estimated that American university libraries were doubling in size every sixteen years. Given this growth rate, Rider speculated that the Yale Library in 2040 will have “approximately 200,000,000 volumes, which will occupy over 6,000 miles of shelves requiring a cataloging staff of over six thousand persons.” In 1961, Derek Price published”[*Science since Babylon*](http://www.amazon.com/Science-Since-Babylon-Enlarged-Edition/dp/0300017987/ref=sr_1_1?s=books&ie=UTF8&qid=1339603940&sr=1-1&keywords=Science+Since+Babylon)”, in which he charts the growth of scientific knowledge by looking at the growth in the number of scientific journals and papers. He concludes that the number of new journals has grown exponentially rather than linearly, doubling every fifteen years and increasing by a factor of ten during every half-century. In November 1967,  B. A. Marron and P. A. D. de Maine publish “[Automatic data compression](http://dl.acm.org/citation.cfm?id=363790.363813&coll=DL&dl=GUIDE&CFID=105563503&CFTOKEN=46382026)” in the *Communications of the ACM*, stating that ”The ‘information explosion’ noted in recent years makes it essential that storage requirements for all information be kept to a minimum.” In 1971, Arthur Miller writes in “[*The Assault on Privacy*](http://www.amazon.com/THE-ASSAULT-ON-PRIVACY-Computers/dp/B001OMSNV6/ref=sr_1_2?s=books&ie=UTF8&qid=1339604062&sr=1-2&keywords=The+Assault+on+Privacy) “that “too many information handlers seem to measure a man by the number of bits of storage capacity his dossier will occupy.” In 1975, The Ministry of Posts and Telecommunications in Japan starts conducting the Information Flow Census, tracking the volume of information circulating in Japan .The census introduces “amount of words” as the unifying unit of measurement across all media. In April 1980,   I.A. Tjomsland gives a talk titled “Where Do We Go from Here?” at the [Fourth IEEE Symposium on Mass Storage Systems](http://www.amazon.com/Digest-Papers-Requirements-Symposium-80CH1581-8/dp/B002ZCRY46/ref=sr_1_1?s=books&ie=UTF8&qid=1339604145&sr=1-1&keywords=Fourth+IEEE+Symposium+on+Mass+Storage+Systems), in which he says “Data expands to fill the space available”. In 1981, The Hungarian Central Statistics Office starts a research include measuring information volume in bits. From August 1983-1990, Ithiel de Sola Pool and Becker studied flow of information and saving the data in the form of bits into the storage devices. 1996-1998 marked the year for, digital storage where it becomes more cost-effective for storing data than paper and implementing big data in virtualization.

August 1999-2000: Big data was used for scientific Virtualization and they came to a conclusion that 92% of the new information was stored on magnetic media, mostly in hard disks. October 2000 Peter Lyman and Hal R. Varian at UC Berkeley publish “[How Much Information?](http://www2.sims.berkeley.edu/research/projects/how-much-info/)” It is the first comprehensive study to quantify, in computer storage terms, the total amount of new and original information (not counting copies) created in the world annually and stored in four physical media. Figure 1.1 illustrates the evolution of Big Data and its usage in publications by various researchers and scientists.



**Figure 1.1 Stages of evolution how the term big data first used in publications**

**(courtesy of RameshDontha,www.DigitalTransformationsPro.com)**

**1.1.1.2 Big Data – Definition**

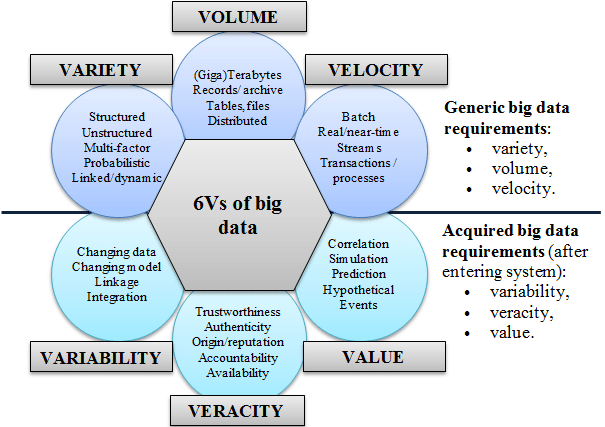
Big data analytics has many perspectives. Arguably, the three most discussed are the business

Perspective, i.e. value and business outcomes, the technology perspective, i.e. computing re-

Sources and IT infrastructure management, and the social perspective, i.e. stakeholders and

Their knowledge, skills and abilities [8]. These perspectives are usually comprised of components, which represent both functional and non-functional requirements [9].

Big data impacts both the strategy development process and  the actual strategies developed in a number of ways, often called the 6V’s"[10] —Volume (vast amounts of data), Variety (significant heterogeneity in the type of data available in the set), and Velocity (speed at which a data scientist or user can access and analyze the data) [1],Variability(changing the data model ),Value(usefulness in applications). This is as shown in figure 1.2.



**Figure 1.2 6Vs of big data**

**[Courtesy of Demchenko, De Laat and Membrey, 2014]**

The V’s of the Health Care Big Data are:

1)**Volume :** According to Health Catalyst [2], healthcare firms with over 1,000 employees store over 400 terabytes of data per firm (reported in the year of 2009), which qualifies healthcare as a high-data volume industry, despite the real-time streams of web and social media data. Contributing to the huge volume of healthcare data are various sources of data, from traditional personal medical records and clinical trial data to new types of data such as various sensor readings and 3D imaging [3].Recently the proliferation of wearable medical devices has significantly added fuel to the healthcare data. Those devices are able to continuously monitor a series of physiological information, such as bio potential, heart rate, blood pressure, and so forth [4].

2) **Variety:** Healthcare data could be characterized by the variety of sources and the complexity of different forms of data. Generally, healthcare data could be classified into unstructured, structured, and semi structured. Historically, most unstructured data usually come from office medical records, handwritten notes, paper prescriptions, MRI, CT, and so on.[5] The structured and semi structured data refers to electronic accounting and billings, actuarial data, laboratory instrument readings, and EMR data converted from paper records [8].Nowadays, more and more data streams add variety to healthcare information, both structured and unstructured, including intelligent wearable devices, fitness devices, social media, and so on.

3) **Velocity:** Compared with relatively static data such as paper files, x-ray films, and scripts, it is gradually becoming more important and challenging to process a real-time stream, such as various monitoring data, accurately and in a timely manner, in order to provide the right treatment to the right patient at the right time [6]. A concrete example can be found in the prevalence of wearable monitoring devices, which provide continuous and ever-accumulated physiological data. Being able to perform real-time analytics on continuous monitoring data could help predict life threatening pathological changes and offer appropriate treatment as early as possible.[5]

 4) **Veracity:** Coming from a variety of sources, the large volume of healthcare data varies in its quality and complexity. It is not uncommon that the healthcare data contains biases, noise, and abnormalities, which poses a potential threat to proper decision-making processes and treatments to patients. The biggest challenge is determining the proper balance between protecting the patient’s information and maintaining the integrity and usability of the data. As Tech crunch [7] points out, “while today we rely on the well-trained eye of the general practitioner and the steady hand of the surgeon, tomorrow’s lifesavers will be the number-crunching data scientists, individuals with only a passing understanding of first aid.”

5) **Variability:** It specifies how varied the data is? Data which is collected from different sources has data outliers to specify the difference of one type of data from another.

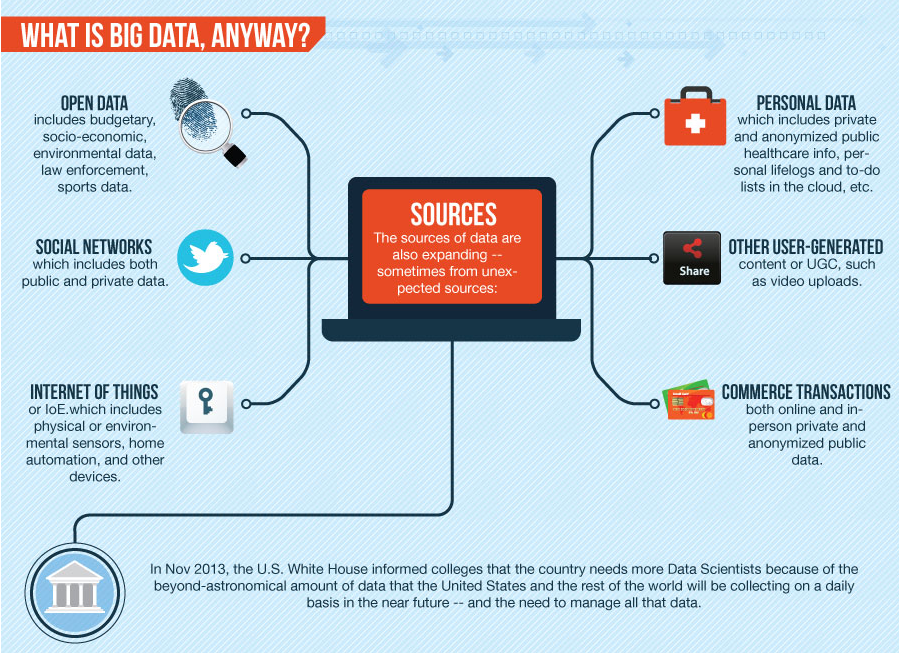
6) **Value:**  It specifies how efficiently the data is used and its utility in different applications .It also shows whether the data was useful for a specific application or not.

**1.1.1.3 Big Data –looking to internet as a model**

Data as we see is coming from different types of sources and is expanding — sometimes from unexpected sources:

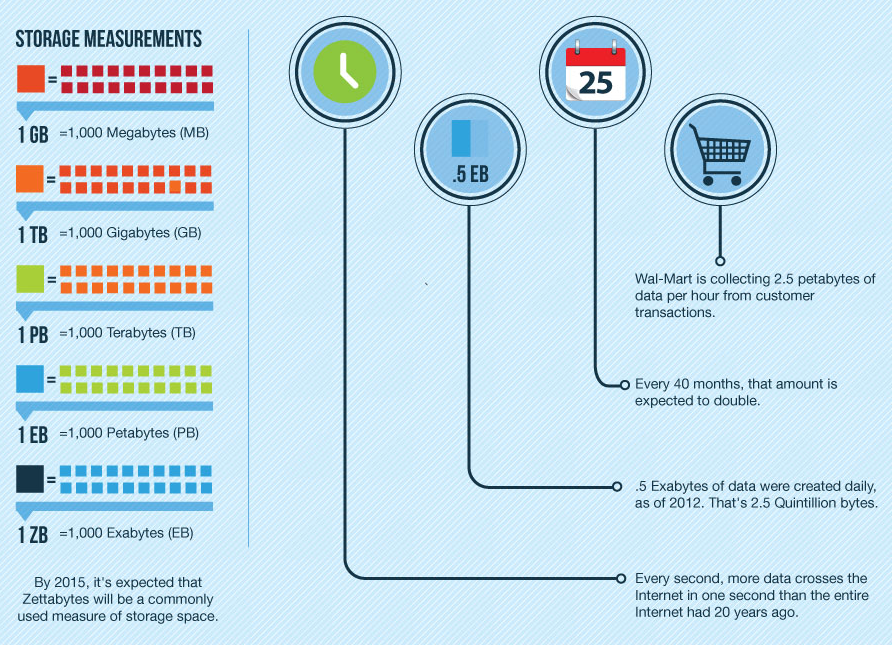
* Open Data —budgetary and socio-economic (cities, states/ provinces, countries), environmental data (land, oceans, weather, astronomy), law enforcement, sports data.
* Social networks — which includes both public and private data.
* Internet of Things (IoE) — which includes physical or environmental sensors (sometimes attached to creatures including bees and cows), home automation, and other devices — along with all other devices that are Internet-connected.
* Personal data — which includes private and anonymized public healthcare info, personal life logs and to-do lists in the cloud, etc.
* Other user-generated content (UGC), such as video uploads
* Commerce transactions — both online and in-person private and anonymized public data.

This is merely a short list of sources from which the world will collect data. Internet is the main source of the massive amounts of data and the question remains .Are we ready for the massive amounts of data in our future? Figure 1.3 shows the sources of data collected and stored.



**Figure 1.3 Data collected from different sources and stored**

The rate at which we are collecting is expanding daily .Big data will come from anywhere and everywhere, and there is a growing need to access and consume data .Even the data that is already collected will contribute to Big data, as this data is saved into any private cloud storage .Figure 1.4 shows how the data is collected and why it is collected? Now the big question lies on how will the data be managed?



**Figure 1.4: How much data is collected and why?**

There have been significant sources that include Internet of Things (IOT) and all the “smart” devices connected to it. For example, devices in IOT collectively produce massive quantities of data in short period of time .For example, Smart sensors collect data at about 65 million units for 2013, and they are expected to be 2.8 trillion by 2019.Similarly some other examples have been listed below.

General electric has gas turbines with 100 sensors producing 1,000 pieces of data per second.4.4 million vehicles generate 1MB of data per day ie.1.6 EB of data per year. Some of these sensors are outside on poles near forests, attached to animals or insects, on wind and gas turbines, others in home automation devices and so on

In healthcare data if in a few years there were 3Billion  people wearing smart watches or smart wristbands that generate even just a half Megabyte per week or about 2MB per month of health data (pulse, temperature, etc.), that would be 6 ZB per month that healthcare providers might utilize for preventive medicine. At least 200 new startups formed between 2010 and Apr 2013 that are focused on healthcare applications. 40% of these startups have further focused on health intervention or predictive features. The estimated potential yearly savings in healthcare costs is $300B-$450B by applying analysis of big data for predictive and preventive care.

Kaiser Permanente has saved around $1B from reduced office visits and lab tests by using aggregated data to improve treatment of cardiovascular disease. It’s estimated that Facebook handles at least 350 GB of data per minute, and that’s from all over the world. In Figure 1.5, The graph of time v/s the data volume from different sources shows how the data is stored at different levels.



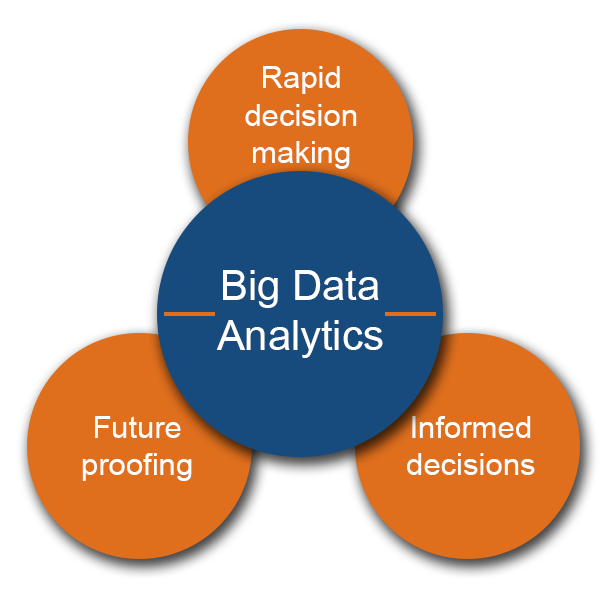
**Figure 1.5.Big data in consumer context**

**(Courtesy IDC and Berkeley Data growth estimate)**

**Big data analytics**

**1.2.1 Big data Analytics –Technologies for business processes**

 Big data analytics is the process of examining large and varied data sets -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions as shown in the figure 1.6.



**Figure 1.6 : Big data Analytics**

Driven by specialized analytics systems and software, big data analytics can point the way to various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals. Big data analytics applications enable [data scientists](http://searchbusinessanalytics.techtarget.com/definition/Data-scientist), predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence [(BI)](http://searchdatamanagement.techtarget.com/definition/business-intelligence) and analytics programs. That encompasses a mix of [semi-structured](http://whatis.techtarget.com/definition/semi-structured-data) and [unstructured data](http://searchbusinessanalytics.techtarget.com/definition/unstructured-data) -- for example, internet [clickstream](http://searchsoa.techtarget.com/definition/click-stream) data, web server logs, social media content, text from customer emails and survey responses, mobile-phone call-detail records and machine data captured by sensors connected to the [internet of things](http://whatis.techtarget.com/definition/Internet-of-Things). On a broad scale, [data analytics](http://searchdatamanagement.techtarget.com/definition/data-analytics) technologies and techniques provide a means of analyzing data sets and drawing conclusions about them to help organizations make informed business decisions. BI queries answer basic questions about business operations and performance. Big data analytics involves complex applications with elements such as [predictive models](http://searchdatamanagement.techtarget.com/definition/predictive-modeling), statistical algorithms and what-if analyses powered by high-performance analytics systems.

**1.2.1.1 Big data analytics - Technologies and tools**

Unstructured and semi-structured data types typically don't fit well in traditional [data warehouses](http://searchsqlserver.techtarget.com/definition/data-warehouse) that are based on [relational databases](http://searchsqlserver.techtarget.com/definition/relational-database) oriented to structured data sets. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually, as in the case of real-time data on stock trading, the online activities of website visitors or the performance of mobile applications. As a result, many organizations that collect process and analyze big data turn to [No-SQL](http://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL) databases as well as Hadoop and its companion tools. These include the following tools:

* [**YARN**](http://searchdatamanagement.techtarget.com/definition/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator): a cluster management technology and one of the key features in second-generation Hadoop.
* [**MapReduce**](http://searchcloudcomputing.techtarget.com/definition/MapReduce): a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.
* [**Spark**](http://searchbusinessanalytics.techtarget.com/definition/Apache-Spark): an open-source parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.
* [**HBase**](http://searchdatamanagement.techtarget.com/definition/Apache-HBase): a column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).
* [**Hive**](http://searchdatamanagement.techtarget.com/definition/Apache-Hive): an open-source data warehouse system for querying and analyzing large datasets stored in Hadoop files.
* [**Kafka**](http://whatis.techtarget.com/definition/Apache-Kafka): a distributed publish-subscribe messaging system designed to replace traditional message brokers.
* [**Pig**](http://searchdatamanagement.techtarget.com/definition/Apache-Pig): an open-source technology that offers a high-level mechanism for the parallel programming of MapReduce jobs to be executed on Hadoop clusters.

In some cases, [Hadoop clusters](http://searchbusinessanalytics.techtarget.com/definition/Hadoop-cluster) and No-SQL systems are being used primarily as landing pads and staging areas for data before it gets loaded into a data warehouse or [analytical database](http://searchbusinessanalytics.techtarget.com/definition/analytic-database) for analysis, usually in a summarized form that is more conducive to relational structures. More frequently, however, big data analytics users are adopting the concept of a Hadoop [data lake](http://searchaws.techtarget.com/definition/data-lake) that serves as the primary repository for incoming streams of [raw data](http://searchdatamanagement.techtarget.com/definition/raw-data). In such architectures, data can be analyzed directly in a Hadoop cluster or run through a processing engine like Spark. As in data warehousing, sound data management is a crucial first step in the big data analytics process. Data being stored in the [Hadoop Distributed File System](http://searchbusinessanalytics.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS) must be organized, configured and partitioned properly to get good performance on both [extract, transform and load](http://searchdatamanagement.techtarget.com/definition/extract-transform-load) (ETL) integration jobs and analytical queries.   Once the data is ready, it can be analyzed with the software commonly used in advanced analytics processes. That includes tools for [data mining](http://searchsqlserver.techtarget.com/definition/data-mining), which sift through data sets in search of patterns and relationships; [predictive analytics](http://searchcrm.techtarget.com/definition/predictive-analytics), which build models for forecasting customer behavior and other future developments; [machine learning](http://whatis.techtarget.com/definition/machine-learning), which tap algorithms to analyze large data sets; and [deep learning](http://searchbusinessanalytics.techtarget.com/definition/deep-learning), a more advanced offshoot of machine learning. [Text mining](http://searchbusinessanalytics.techtarget.com/definition/text-mining) and [statistical analysis](http://whatis.techtarget.com/definition/statistical-analysis) software can also play a role in the big data analytics process, as can mainstream BI software and [data visualization](http://searchbusinessanalytics.techtarget.com/definition/data-visualization) tools. For both ETL and analytics applications, queries can be written in batch-mode MapReduce; programming languages, such as [R](http://searchbusinessanalytics.techtarget.com/definition/R-programming-language), [Python](http://searchenterpriselinux.techtarget.com/definition/Python) and [Scala](http://searchbusinessanalytics.techtarget.com/definition/Scala-Scalable-Language); and [SQL](http://searchsqlserver.techtarget.com/definition/SQL), the standard language for relational databases that's supported via [SQL-on-Hadoop](http://searchdatamanagement.techtarget.com/definition/SQL-on-Hadoop) technologies.

**1.2.1.2 Big data analytics uses and challenges**

Big data analytics applications often include data from both internal systems and external sources, such as weather data or demographic data on consumers compiled by third-party information services providers. In addition, streaming analytics applications are becoming common in big data environments, as users look to do [real-time analytics](http://searchcrm.techtarget.com/definition/real-time-analytics) on data fed into Hadoop systems through Spark's Spark Streaming module or other open source stream processing engines, such as [Flink](http://searchdatamanagement.techtarget.com/definition/Apache-Flink) and [Storm](http://whatis.techtarget.com/definition/Apache-Storm).

Early big data systems were mostly deployed on-premises, particularly in large organizations that were collecting, organizing and analyzing massive amounts of data. But cloud platform vendors, such as Amazon Web Services ([AWS](http://whatis.techtarget.com/definition/Amazon-Web-Services-AWS)) and Microsoft, have made it easier to set up and manage Hadoop clusters in the cloud, as have Hadoop suppliers such as Cloudera and Horton works, which support their distributions of the big data framework on the AWS and [Microsoft Azure](http://searchcloudcomputing.techtarget.com/definition/Windows-Azure) clouds. Users can now spin up clusters in the cloud, run them for as long as needed and then take them offline, with usage-based pricing that doesn't require ongoing software licenses. Potential pitfalls that can trip up organizations on big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced data scientists and [data engineers](http://searchdatamanagement.techtarget.com/definition/data-engineer) to fill the gaps. The amount of data that's typically involved, and its variety, can cause data management issues in areas including [data quality](http://searchdatamanagement.techtarget.com/definition/data-quality), consistency and governance; also, data silos can result from the use of different platforms and data stores in a big data architecture. In addition, integrating Hadoop, Spark and other big data tools into a cohesive architecture that meets an organization's big data analytics needs is a challenging proposition for many IT and analytics teams

**1.2.1.3 STAKEHOLDERS**

The various stakeholders in healthcare industry have different expected incentives and hopes from Big Data which can be summarized as follows:

1) Patients want their everyday use of technology to flow seamlessly into their medical care. Some want to comparison shop for medical treatment as they do for consumer products. Everyone wants customer-friendly service, one-stop shopping, and better coordination of care between themselves, caregivers and various providers, with an ultimate goal of error-free, compassionate and effective care.

2) Providers want real-time access to patient, clinical and other relevant data to support improved decision-making and facilitate effective, efficient and error-free care. They want technology to be a transparent tool, not an encumbrance.

3) Researchers want new tools to improve the quality and quantity of workflow – e.g., predictive modeling, statistical tools and algorithms that improve the design and outcome of experiments and provide a better understanding of how to develop treatments that meet unmet needs while successfully navigating the regulatory approval and marketing process.

4) Pharmacy companies want to better understand the causes of diseases, find more targeted drug candidates, and design more successful clinical trials to avoid late failures and market safer and more effective pharmaceuticals. Once in the market, they want accurate formulary and reimbursement information to customize their marketing efforts, as well as less costly post-marketing surveillance.

5) Medical device companies, many of which have been collecting data for some time from hospital and home devices for safety monitoring and adverse event prediction, are beginning to wonder what to do with this data, and how to integrate it with old and new forms of personal data.[11] identify the right mix of technologies and then put the pieces together.

**1.2.1.4 OPPORTUNITIES**

 By digitizing, combining and effectively using big data, healthcare organizations ranging from single-physician offices and multi-provider groups to large hospital networks and accountable care organizations stand to realize significant benefits [12]. Implicit benefits of big data analytics in healthcare include earlier detection of diseases and ailments when they are in early stages and can be controlled and treated more easily and efficiently; individual health management by providing patient centric services; improving the treatment methods and detecting healthcare fraud more quickly and efficiently. McKinsey estimates that big data analytics can enable more than $300 billion in savings per year in U.S. healthcare, two thirds of that through reductions of approximately 8% in national healthcare expenditures. Clinical operations and R & D are two of the largest areas for potential savings with $165 billion and $108 billion in waste respectively [13]. McKinsey believes big data could help reduce waste and inefficiency in the following areas:

**a) Clinical Operations**

1) Comparative effectiveness research to determine more clinically relevant and cost effective ways to diagnose and treat patients.

2) Clinical decision support systems to enhance the efficiency and quality of operations; i.e., providing real-time information to emergency technicians, nurses and doctors to improve triage, diagnosis, treatment choice, prevent iatrogenic infections and readmissions, prescription and other medical errors.

3) Other areas include increasing transparency about medical data, remote patient monitoring, and predictive analytics to identify individuals who would benefit from proactive care. [13]

**b)Research & development**

1) Predictive modeling to lower attrition and produce a leaner, faster, more targeted R & D pipeline in drugs and devices.

2) Statistical tools and algorithms to improve clinical trial design and patient recruitment to better match treatments to individual patients, thus reducing trial failures and speeding new treatments to market.

3) Analyzing clinical trials and patient records to identify follow-on indications and discover adverse effects before products reach the market.

**c) Public health**

1) Analyzing disease patterns and tracking disease outbreaks and transmission to improve public health surveillance and speed response.

2) Faster development of more accurately targeted vaccines, e.g. choosing the annual influenza strains.

3) Turning large amounts of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises, especially for the benefit of populations [14].

**d) Genomic analytics**: Add genomic analysis to the traditional healthcare decision making process by developing efficient and effective gene sequencing technologies. Utilize high throughput genetic sequencers to capture organism DNA sequences and perform genome-wide association studies (GWASs) for human disease and human microbiome investigations.[14]

**e) Fraud detection:** Analyze a large amount of claim requests rapidly by using a distributed processing platform (e.g., MapReduce for Hadoop) to reduce fraud, waste, and abuse, such as a hospital’s overutilization of services, or identical prescriptions for the same patient filled in multiple locations[15]

**f) Device/remote monitoring:** Capture and analyze continuous healthcare data in huge amounts from wearable medical devices both in the hospital and at home, for monitoring of safety and prediction of adverse events.[4]

**1.2.1.4 ISSUES AND CHALLENGES**

Along with the benefits that healthcare has been leveraging from big data, there is certain issues and challenges that act as barriers in successful implementation of big data for healthcare and prove to be a hindrance in proper and maximum extraction in the terms of advantages that big data can actually offer. Big data analytics not only provides charming opportunities but also faces lot of challenges. The challenge starts from choosing the big data analytics platform. While choosing the platform, some criteria like availability, ease of use, scalability, level of security and continuity should be considered [16]. The other challenges of big data analytics are data incompleteness, scalability and security [17,18]. Some of the challenges are:

**1) Privacy and Data Security** : Internet transactions, cloud storage, social media communications and related data exposes the personal and private data to potential and implicit misuse which makes the privacy of the data a grave issue which need to addressed and tackled earnestly. Privacy of data specific to the field of healthcare is in the terms of 1) The legal traditional doctor-patient confidentiality 2) The concern of patients regarding disclosure of their health status to third parties 3) Conflicting desires of third parties (insurers, employers, etc.) to access data [34]. Use of the Internet, cloud computing and pooling of data all raise the data security stakes. Healthcare data contains the intimate details of a person’s life and we must respect and protect it with the highest security possible.

**2) Data Standardization and Data Structure issues**: The data available in the healthcare industry is largely in an unstructured format which is in the format of graphs, prescription notes, images. Apart from this, the nature of structured data is mostly heterogeneous. Leveraging the patient or data correlations in longitudinal record and understanding unstructured clinical notes in the right context is a grave problem .Although the EHRs share data within the same organization, intra-organizational, EHR platforms are fragmented, at best. Data is stored in formats that are not compatible with all applications and technologies [21,24] .This lack of data standardization also causes problems in transfer of that data [25,26].

**3) Data Storage and Transfers**: Data generation is inexpensive as compared to data storage. The real problem is in efficiently storing the data such that different methodologies and technologies can be easily applied to extract the desired information successfully. Once data is generated, the costs associated with securing and storing them remain high [26]. Costs are also incurred with transferring data from one place to another as well as analyzing it [28,24,29].Though many technologies have been devised to store and transfer the structured data, data scientists are still struggling with finding better ways to do the same to unstructured and heterogeneous data. Unstructured data is not as easy as the structured data to be analyzed, processed, stored, or transferred .The need of the hour is finding cheaper and less expensive ways for storage and transmission of secure or insecure data.

**4) Requirement of Appropriate skills**: The expertise in technical skills make the handling, storage, retrieval and implementation of big data cumbersome. The McKinsey Global Institute estimates that there will be a more than 100,000 person shortage through 2020. It means that mean 50–60% of data scientist positions may go vacant. Data scientists need highly technical skill sets. They must possess soft skills such as communication, collaboration, leadership, creativity and more [27].The scarcity of employees and workers in healthcare industry with adequate skill required in handling the big data is a significant barrier.

**1.3 Medical Big Data Analytics**

**1.3.1 Introduction**

“Information is the oil of the 21st century and analytics is the combustion engine “.Peter Sondegaard, Global head of research for Gartner stated that information is the oil of the 21st century and combustion is the engine .So what exactly is analytics and why is it so important to the 21st century health care?

The institute of medicine in their 2012 report titled ,”Best care at lower cost : The Path to continuous learning Health Care in America “ stated that America’s Health Care System has become far too complex and costly to continue business as usual .Pervasive inefficiencies and inability to manage a rapidly deepening clinical knowledge , and a reward system poorly focused on the key patient needs all hinder improvements in the safety and quality of care and threaten the nation’s economic stability and global competitiveness. Achieving higher quality care at lower cost will require fundamental commitments to take incentives, culture and leadership that foster continuous learning as the lessons from research and each care experience are systematically captured assessed and translated into reliable care. They define a learning health care system as a system designed to generate and apply best evidence for the collaborative healthcare choices of each patient and provider. To drive the process of discovery as a natural outgrowth of patient care, and to ensure innovative, quality, safety, and value in healthcare.

A Two-fold objective of Health care is to do the following:

Introduce data mining researchers to the sources available and the possible challenges and techniques associated with using big data in healthcare domain.

 Introduce Healthcare analysts and practitioners to the advancements in computing field to effectively handle and make inferences from voluminous and heterogeneous healthcare data.

**1.3.1.1. The Big Picture of Patient Data**

Big Data has changed the way we manage, analyze and leverage data in any industry. One of the most promising areas where big data can be applied to make a change is healthcare. [Healthcare analytics](https://www.datapine.com/healthcare-analytics) have the potential to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life in general. Healthcare analytics ensure better clinical performance and correspondingly better patient outcomes.

Analytics tools mitigate the effect of errors clerically, financially and above all else on patient health; readmissions and misdiagnoses are avoided. It allows for a reduced operating cost while delivering higher quality services without compromising patient safety or medical staff satisfaction. Modern business analytics encourage a culture of transparency, effectively improving performance, accountability and protecting the medical institution itself. It helps to accurately predict demand for services and workforce supply and planning in case of unforeseen situations that could not have otherwise been expected. Healthcare business intelligence software enables you to develop a thorough patient analysis, giving doctors robust insights and ultimately improving patient care. Moreover, it improves data management and sharing, reducing time spent not treating patients or ensuring their health and satisfaction. All in all healthcare analytics guarantee a better final outcome: safe and healthy patients. Average human lifespan is increasing along world population, which poses new challenges to today’s treatment delivery methods. Healthcare professionals, just like business entrepreneurs, are capable of collecting massive amounts of data and look for best strategies to use these numbers. In this article, we would like to address the need of big data in healthcare: Why and how can it help? What are the obstacles to its adoption? If we consider a hospital system it will likely have an electronic health record system (EHR) as well as specialized departmental systems for laboratory, diagnostic imaging ,pharmacy ,nutrition services, billing , and atomic pathology and so on .Each of these systems is designed and intended  for clinical use .In other words , patient care and so they capture specific data about the patient .However , none of these systems has a complete set of data for any individual patient or for a group of patients such as all patients who were admitted in January with a certain diagnosis that could be used for analysis and reporting thus obtaining deep insight into what is happening with individual patients , as well as groups of patients, requires aggregating data together  from many systems  and performing statistical analysis of this aggregated data .

**1.3.1.2. Why Big Data in health care?**

There’s a huge need for big data in healthcare as well, due to rising costs in nations like the United States. As [a McKinsey report](http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care) states, “After more than 20 years of steady increases, healthcare expenses now represent 17.6 percent of GDP —nearly $600 billion more than the expected benchmark for a nation of the United States’s size and wealth.”

In other words, healthcare costs are much higher than they should be, and they have been rising for the past 20 years. Clearly, we are in need of some smart, data-driven thinking in this area and the current incentives are changing as well. Many insurance companies are switching from fee-for-service plans (which reward using expensive and sometimes unnecessary treatments and treating large amounts of patients quickly) to plans that prioritize patient outcomes. In the previous scheme, healthcare providers had no direct incentive to share patient information with one another, which had made it harder to utilize the power of big data. Now that more of them are getting paid based on patient outcomes as they have a financial incentive to share data that can be used to improve the lives of patients while cutting costs for insurance companies.

The physician decisions are becoming more and more evidence-based, meaning that they rely on large swathes of research and clinical data as opposed to solely their schooling and professional opinion. As in many other industries, data gathering and management is getting bigger, and professionals need help in this matter. This new treatment attitude means there is a greater demand for big data analytics in healthcare facilities than ever before, and the rise of [SaaS (Software – as- a- Service) business intelligence tools](https://www.datapine.com/saas-bi) is also answering that need.

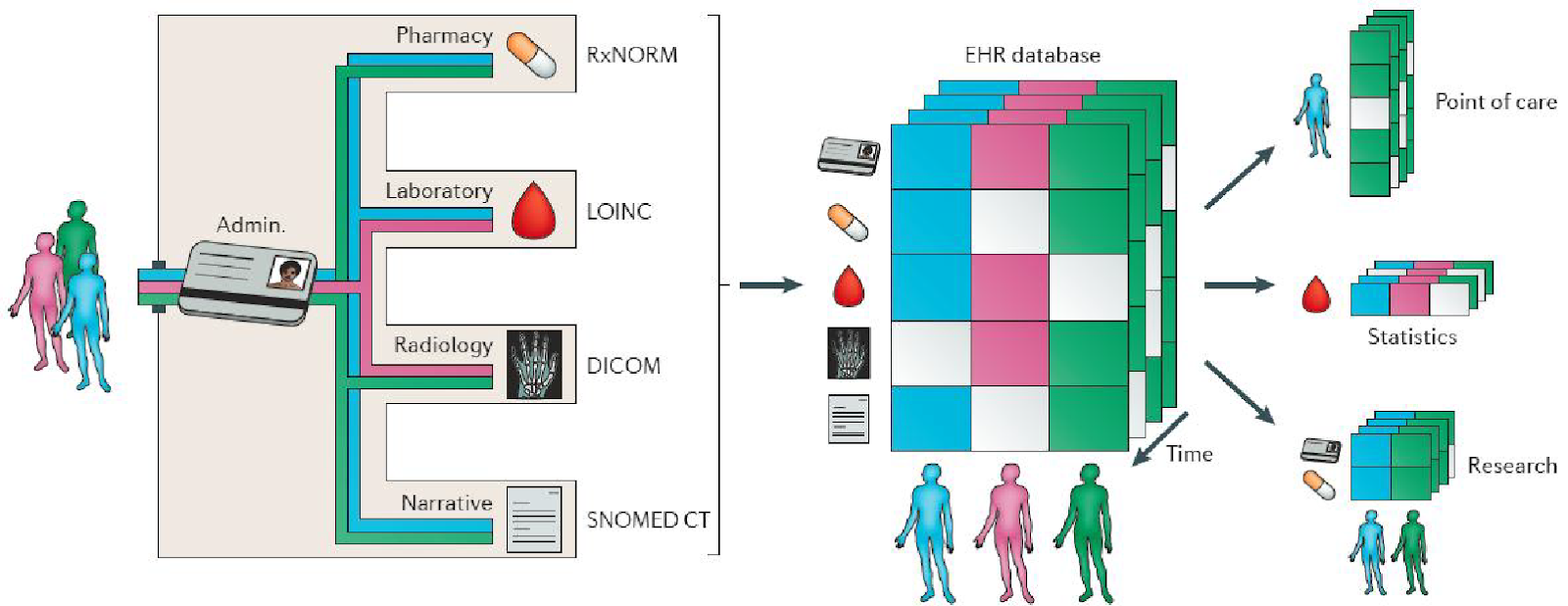
**1.3.1.3. Obstacles to Data-Driven Healthcare**

One of the biggest hurdles standing in the way to use big data in healthcare is how medical data is spread across many sources governed by different states, hospitals, and administrative departments. Integration of these data sources would require developing a new infrastructure where all data providers collaborate with each other. Equally important is implementing new [data analysis tools](https://www.datapine.com/data-analysis-tools) and strategies. Healthcare needs to catch up with other industries that have already moved from standard regression-based methods to more future-oriented like predictive analytics, machine learning, and graph analytics.

However, there are some glorious instances where healthcare doesn’t lag behind, such as EHRs (especially in the US.) So, even if these services are not your cup of tea, you are a potential patient, and so you should care about new healthcare analytics applications. One can observe industries cope with big data and how they can inspire you to adapt and adopt some good ideas.

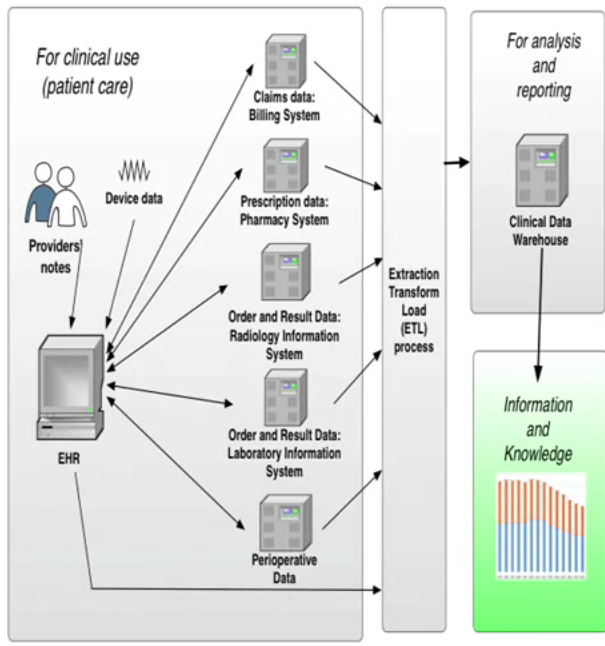
**1.3.1.4. Clinical Data Warehouse**

A clinical data warehouse brings together data for a patient into a single coordinated location and this location is used for analysis and reporting purposes as shown in fig 1.7



**Figure 1.7: Electronic Health record of a patient.**

This is accomplished via a process known as extraction, transform, load or ETL, which retrieves data from various clinical systems, synchronizes formats of data in a process called transformation, and cleans up the data and then imports the data into the database of the clinical data warehouse. The transformation process is especially important, as data can be stored in a variety of forms   across systems as shown in figure 1.8.



**Figure 1.8: clinical data warehouse**

 For eg: a laboratory system might use the letters M, F or U for the patient gender as male, female or unknown .While  the radiology information system might use1, 2 or 9 instead. However, they must match the designations used in the clinical data warehouse. And that process of converting them to match is called transformation .Another important step is ensuring that all of the patients records from various systems are linked together .This typically requires a master patient index, sometimes called a master person index to link a patients various identifiers across systems. Effectively integrating and efficiently analyzing various forms of healthcare data over a period of time can answer many of the impending healthcare problems.

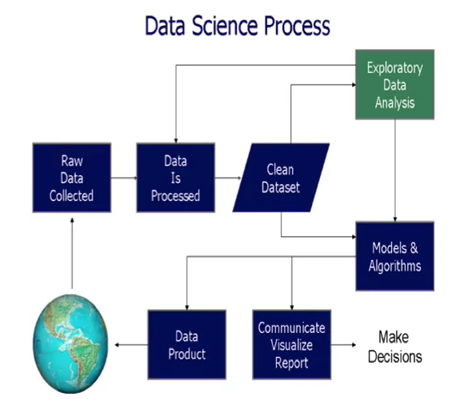
**Analytics**

**1.4.1. Process and Information Extraction**

A centralized coordination is needed for searching the location for patient data that can be used for analysis and reporting .The term Analytics can be defined as “the discovery of meaningful patterns in data, and is one of the steps in the data life cycle of collection of raw data, preparation of information, analysis of patterns to synthesize knowledge, and action to produce value.” (NIST Big Data, 2015).

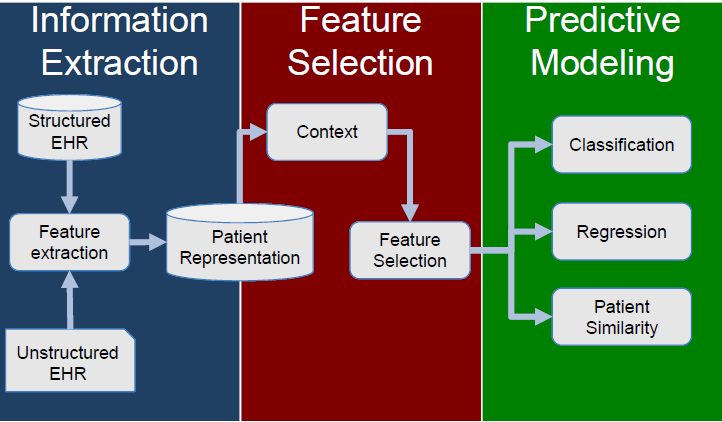
The term analytics has been used in a variety of ways and  with different meaning .In fact , Gartner stated that analytics has emerged as a catchall term for a variety  of different business intelligence ,BI, and application related initiatives .In 2015  the National Institute Of standards issued a formal definition of analytics as follows .The term analytics refers to the discovery of meaningful patterns in data , and is one of the steps in the data life cycle of collection of raw data , preparation of information , analysis of patterns to synthesize knowledge and action to produce value .

 The analytical process is the synthesis of knowledge from information. It includes statistical analysis as one of the steps. as shown in the figure 1.4.1.1 below ,where a raw data is collected from a source and processed by Exploratory Data Analysis and the data is sent for cleaning processed data to select the information required and communicated with the help of visualization tools  such as graphical representations ie. Bar charts, pie charts etc.



**Figure 1.4.1.1: Process of Analytics**

Analytics is the entire process of data collection, extraction, transformation, analysis, interpretation, and reporting .According to the figure 1.4.1.2 , In the information extraction phase , Structured data and unstructured are taken as input and features are extracted and stored as patient representation. The stored patient representation is the input to Feature selection phase  .In the Feature Selection phase data is separated in terms of the context of usage using feature selection process and this output forms the  input to the Predictive modeling. In predictive modeling data is categorized based on three algorithms ie. Classification, regression, and patient similarity.



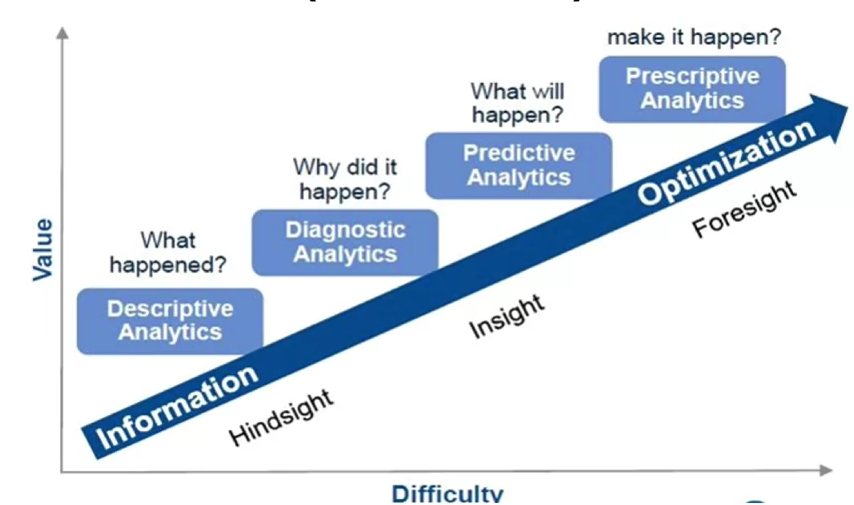
**Figure 1.4.1.2: Steps involved in Information Extraction**

**Types of Analytics**

IBM in 2013 categorized analytics into three types

1. Descriptive: uses business intelligence in data mining to ask what has happened?
2. Predictive:  uses statistical models and forecasts to ask, what could happen?
3. Prescriptive: uses optimization and simulation to ask ”What should we do ?”

Diagnostic: Examines data to answer “Why did it happen?” All the stages of Analytics are as shown in figure 1.4.1.3.



**Figure 1.4.1.3: Types of Analytics**

Gartner added a fourth type of diagnostic analytics which describes a form of advanced analytic which examines data or content to answer the question “why did it happen?” The simplest type of analytics starts in the lower left hand corner with the diagnostic   analytics  are more valuable to the institute, but also more difficult to perform .Even more difficult and also more valuable are the predictive analytics. Finally the most difficult and also the most valuable are the prescriptive analytics.

Descriptive analytics are the simplest type of analytics which simply describe the data .Common statistics are used such as the number of laboratory test. The average age of the patient or the average length of stay in the hospital for the patients with particular diagnosis. Descriptive Analytics are often presented as pie charts, bar or column charts, tables or written narratives. Gartner defines diagnostic Analytics as a form of advanced analytics which examines data or content to answer the question, why did it happen? Tools used for diagnostic techniques include Drill down techniques, data discovery and correlations. Let’s start with an example as explained below.

**Example 1.1:** Kaiser Permanente analyzed data on infants to develop an algorithm for classifying which babies were at risk for developing sepsis  and  conversely which babies did not need to be treated .Sepsis is described by the mayo clinic as a potentially life threatening , complication of an infection .Sepsis occurs when chemicals released into the blood  stream to fight the infection  trigger inflammatory responses throughout the body .This inflammation can trigger a cascade of changes that can damage multiple organ systems , causing them to fail. If Sepsis progresses to septic shock, blood pressure drops dramatically which may lead to death. Kaiser Permanente stated that judicious application of our scheme could result in decreased antibiotic treatment in 80,000 to 240,000 US new borns each year. With that example  in mind , let’s now look at a definition of predictive analytics and how the a Kaiser Permanente case is an example of predictive analytics.

Gartner states that predictive analytics has the  following four attributes .First, an emphasis on prediction rather than description, classifying or clustering .In the Kaiser Permanente example , they were  trying to predict which new borns were at risk of developing a life threatening condition so that they could treat the babies to prevent it .The second attribute defined by Gartner is rapid analysis often in hours or days .Consider again the sepsis example .Sepsis is a rapidly progressing condition that , if it  progress to the most severe stage of septic shock , can have a 50% mortality rate. Therefore analysis of the data to predict which infants are at risk of developing this condition must be done rapidly not over a period of weeks or months.

The third attribute defined by Gartner is an emphasis on the business relevance of the resulting insights .Consider the word relevance and how that would apply to the sample of infants with a life threatening infection. Information that would directly affect the care and prevent infants from dying is relevant. And finally the fourth attribute defined by Gartner is an emphasis on ease of use, thus making the tools accessible to business users. In other words, these tools should be available to the clinical staff to use .However it is important to note that as Michael Woo states, the purpose of predictive analytics is not to tell you what will happen in the future .It cannot do that .In fact, no analytics can do that .Predictive analytics can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature. An example of predictive modeling is explained below. This example shows the prediction of how many days a patient will spend in the hospital. The Heritage Health Prize was an initiative taken by Dr. Merkin to create and develop such a task as explained below.

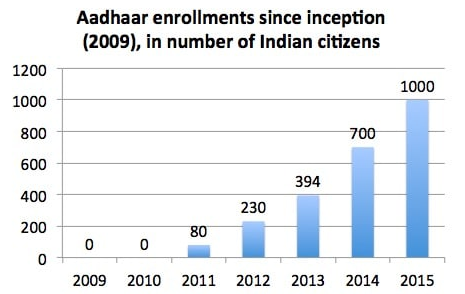
**Example 1.2: Heritage Health Prize**

HPN was recently named one of the global 10 most innovative companies in healthcare for 2012 by “Fast Company Magazine”.  The $3 million Heritage Health Prize is the world's largest predictive modeling contest, challenging entrants to create an algorithm that predicts how many days a patient will spend in the hospital. Created, developed and sponsored by Dr. Merkin, the goal of the prize is to decrease the number of avoidable hospitalizations, saving the country more than $40 billion in avoidable hospitalization costs centric, integrated health care systems that represent the future of health care in the United States. HPN develops programs and services that are responsive to the healthcare needs of today's patients. HPN designs its networks to offer patients a comprehensive range of medical care in convenient locations. It organizes and manages medical groups and independent practice associations, integrating with hospitals and ancillary care providers. This system represents a workable, successful business model that has been duplicated with medical groups, IPAs and medical facilities throughout California. HPN is dedicated to quality, affordable health care and putting patients' wellness first.

Gartner defines prescriptive analytics as a form of advanced analytics which examines data or content to answer the question, what should be done or what can we do to make something happen? As characterised by techniques such as graph analysis, simulation, complex event processing, Neural Networks, Recommendation engines, heuristics, machine learning etc. An application of Machine learning models can be seen in the example below which shows the enrollment of Adhaar cards in India for booking an appointment with the doctor in hospitals through the Adhaar card .The statistics of adhaar card enrollment.

**Example 1.3 Aadhar-based**[**Online Registration System**](http://ors.gov.in/index.html)

In India, the Ministry of Electronics and Information Technology, Government of India, runs Aadhar-based [Online Registration System](http://ors.gov.in/index.html), a platform to help patient’s book appointments in major government hospitals. The portal has the potential to emerge into a source if big data offering insights on diseases, age groups, shortcomings in hospitals and areas to improve. The website claims to have already been used to make 8, 77,054 appointments till date in 118 hospitals. If a researcher has huge sets of data at his disposal, he/she can also find out patterns and simulate it through machine learning tools, which decreases the time required to arrive at a conclusion. Machine learning methods become more robust when they are fed with results analyzed from big data. These data simulation models rely on primary information generated from a study to build predictive models that can help assess how human body would respond to a given perturbation. Due to such models coming into existence, the number of submissions for adhaar card has grown from 0 enrollments in 2009 to 1000 enrollments in 2015 since its inception as shown in the figure 1.3.1 below.



**Figure 1.3.1: Adhaar enrollments**

**(Courtesy by ZDNet, Trak.in,2015)**

Now let us look at the steps in data analysis in more detail. Data Analytics involves a sequence of steps :

1. Identify the problem and the stakeholders

2. Identify what data are needed and where those data are located

3. Develop a plan for analysis and a plan for retrieval

4. Extract / transform or load the data

5. Check, clean, and prepare   the data for analysis

6. Analyze and  interpret the data

7. Visualize the data

8. Disseminate the new knowledge

9. Implement the knowledge into the organization

 In order to identify the problem or question and the stakeholders, the following questions have to be answered:

* Why is this important problem?
* How will the results impact patient care or the institution?
* What is the business case?
* Who are the stakeholders?

In business terms, identify the business case. One must clearly state the problem or the question to guide the rest of the process. We must also find out the impact of the results on for eg : patient care  and who ultimately will be the stake holders

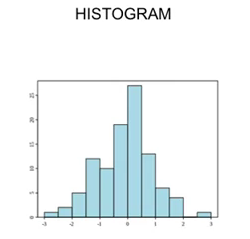
To  identify  what data are  needed  for the analysis  to be identified for eg : What data elements  such as date of birth , gender ,medications , laboratory results and so on are needed Where these elements are located ie. In which systems and database tables? Is there a clinical data warehouse? Who is the contact person for each system and who will be responsible for retrieving the data? All these require multiple extraction steps .A plan for retrieving the data from the various systems along with a plan for checking that all the data required were actually retrieved should be developed .There needs to be some way to determine how many records are expected, and actually retrieved. This may involve cross checking against other systems .This step will require the participation of the individuals who normally perform data retrieval from the systems, involved. An analysis plan needs to be developed and a statistician should be consulted and questions are to be addressed here include, what is the population? What size does the sample need to be? What statistical tests should be performed?

The next step is the actual extraction of the data from the system or the system involved. After the data are retrieved, the data needs to be checked for completeness and then check whether the set of data is complete .Then we analyze whether all the records that should be retrieved are actually retrieved? At a minimum statistics, such as counts, must be performed at this step .At this point, changes to the extraction plan may be needed and another extraction from the source systems may need to take place. Once  a complete set of records is extracted  from the source systems , errors in the  records need to be identified and  corrected .And all data having errors , such as transposed letters and names , and incorrect values .Decisions must be made about how to handle empty fields .Next , the data must also be  synchronized or transformed .One  set of values must be  changed so that all the records are using the same  values .After all the necessary transformation steps have been completed , the data are then imported into the destination system  where the actual data analysis and reporting will take place. This ,may be a system as a complex as a clinical data warehouse , or simply as a desktop computer .The  data now in the system where the  analysis will be run , and it  should be a complete set of data .

One needs to check that everything is ready for analysis .One also needs to check whether we got what we needed? Check and verify this against the analysis plan that was developed in step three and that everything to address the problem that was identified in step one .Now you are ready to do the actual analysis. To execute the analysis plan that was developed earlier, perform the statistical analysis and enlist the assistance of statistician to confirm the interpretations and conclusions of your analysis.

**1.4.1.2 Visualize the data**

Now you need to able to communicate the results of your analysis and how the results address the problem from step one. This communication must be very clear and rapidly understandable the decision makers in the institution .So selecting an appropriate representation for your findings is very essential .For this , choose a visualization that is appropriate  for the type of data for eg: categorical data can be represented with column or bar charts and pivot tables .While quantitative data can be shown  with histograms and a wide variety of other types of graphics such as scatter plots and star plots .Some common tools are Tableau and Microsoft Excel Chart function . Once  the  analysis , interpretation , and any visualizations are complete, a report must be developed .It might be a  formal written document , an  email , or a presentation .Regardless  of then delivery method , the report needs to clearly state the original problem , the process that was used to address the  problem , and then the results of the analysis along with the supporting visualization .this represents  new knowledge and needs to be  distributed to the stakeholders that were  identified in step one .Finally , the new knowledge needs to be implemented to address the original problem. This will require the participation of the stake holders. The different visual representations are as shown in figure 1.13 below.



**Figure 1.13: Data visualization in pie chart and histogram representations.**

**1.4.1.3 Medical data analysis –Technical Issues**

The complexity of healthcare results from the diversity of health-related ailments and their co-morbidities; the heterogeneity of treatments and outcomes; and the subtle intricacies of study designs, analytical methods and approaches for collecting, processing, and interpreting healthcare data [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b6-krcp-36-003)]. There are various sources of medical big data, such as administrative claim record, clinical registries, electronic health records, biometric data, patient-reported data, the internet, medical imaging, biomarker data, prospective cohort studies, and large clinical trials [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b2-krcp-36-003),[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/" \l "b7-krcp-36-003)]. Integration of these data sources causes complementary dimensions of data such as large size (smaller than big data from other disciplines, but larger than data of clinical epidemiology), disparate sources, multiple scales (seconds to years), incongruence’s, incompleteness, and complexity. There is no universal protocol to model, compare, or benchmark the performance of various data analysis strategies [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b6-krcp-36-003)].

Medical big data have several distinctive features that are different from big data from other disciplines. Medical big data are frequently hard to access and most investigators in the medical arena are hesitant to practice open data science for reasons such as the risk of data misuse by other parties and lack of data-sharing incentives [[4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b4-krcp-36-003)]. Medical big data are often collected based on protocols (i.e., fixed forms) and are relatively structured, partially due to the extraction process that simplify raw data [[9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b9-krcp-36-003)]. Another important feature is that medicine is practiced in a safety critical context in which decision-making activities should be supported by explanations. It can be costly due to involvement of the personnel, use of expensive instrumentation, and the potential discomfort of the patients involved. Medical big data are relatively small compared to data from other disciplines, and may be collected from a non-reproducible situation. They can be further affected by several sources of uncertainty, such as measurement errors, missing data, or errors in coding the information buried in textual reports. Therefore, the role of the domain knowledge may be dominant in both analyzing the data and interpreting the results [[10](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b10-krcp-36-003)]. Other distinctive features of medical big data in analytic aspects includes the different types of patient characteristics, which sometimes may require weighting; the time structure, which may be an additional dimension; and treatment information, time point of treatment decision and change (i.e., time-dependent confounding) [[11](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b11-krcp-36-003)].

A big data project involves making sense out of all accumulated data on as many variables as possible due to increasing availability and decreasing expense of computing technology [[12](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b12-krcp-36-003)]. Iwashyna and Liu [[13](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b13-krcp-36-003)] pointed that there were four ways in which a project might be “big data”: material, question, analytic method, and aspiration. Medical big data may include data from new sources as materials for analysis, such as the internet, social media, and so on. It  can give answers to questions focusing on the usefulness of locally stable associations and correlations even in the absence of causal evidence, with analytic methods such as new, often nonlinear, tools for pattern recognition from computer science and other fields, in addition to the conventional statistical tools. Finally, big data technology has been increasingly viewed as the catalyst for a continuously learning health system allowing bidirectional flow between research and operations [[13](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b13-krcp-36-003)]. As previously discussed big data can be the fuel flowing in a continuously learning healthcare system [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b2-krcp-36-003)] as stated in section 1.3.1.

Medical big data can be broadly classified into three common forms, such as large *n* and small *p* (*n* = sample numbers, *p* = parameter numbers); small *n* and large *p*; and large *n* and large *p* [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b5-krcp-36-003)]. Data with large *n* and small *p* can be dealt with classical statistical methods. One example of this kind of data is administrative claim data. Because this kind of data tend to be incomplete, noisy, and inconsistent, data cleaning such as defining cases to be analyzed is not trivial and understanding the context of data collection is essential. Spurious association can be another problem. Discrimination between statistical and scientific significance of domain expertise may be crucial. Second form is data with small *n* and large *p*. Microarray analysis datasets are typical examples and classical statistical tests may not be able to deal with this type of data efficiently. Curse of dimensionality and multiple testing issues are the main problems with this type of data. Last, some data has large *n* and large *p*, where the issues of the first and second types may be raised according to circumstances. Although medical big data analysis and clinical epidemiology share many features, there are several differences between these two, some of which are summarized in [Table 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/table/t1-krcp-36-003/).1

**Table 1.1 Medical big data analysis vs. classical statistical analysis**

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Medical big data analysis** | **Classical statistical analysis** |
| Application | Hypothesis-generating | Hypothesis-testing |
| Questions of interest | Overcoming the limitation of locally or temporally stable association with continually updating the data and algorithm | Trying to prove causal relationships |
| Domain knowledge | More important in interpretation of the results | Important both in collection of data and interpretation of the results |
| Sources of data | Any kind of sources; frequently multiple sources | Carefully specified collection of data; usually single source |
| Data collection | Recording without the direct supervision of a human | Human-based measurement recording |
| Coverage of data to be analyzed | Substantial fraction of entire population | Small data samples from a specific population with some assumptions of their distribution |
| Data size | Frequently huge | Relatively small |
|  | Medical big data analysis | Classical statistical analysis |
| Nature of data | Unstructured and structured | Mainly structured |
| Data quality | Rarely clean | Quality controlled |
| Research questions of data analysis | May be different from those of data collection | Same as those of data collection |
| Underlying assumption of the model | Frequently absent | Based on various underlying probability distribution function |
| Analytic tools | Frequently automated with data mining algorithm | Manually by expert with classical statistics |
| Main outputs of analysis | Prediction, models, patterns identified | Statistical score contrasted against random chance |
| Privacy & ethics | Concerns about privacy and ethical issues | Data collection according to the pre-approved protocol; informed consent from the participants |

**1.4.1.4 How can medical big data be analyzed?**

Big data analysis exploits various algorithms of data mining, which can be defined as the automatic extraction of useful, often previously unknown information from large databases or datasets using advanced search techniques and algorithms to discover patterns and correlations in large pre-existing databases [[18](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b18-krcp-36-003)]. The tasks of data mining can be summarized as description, finding human-interpretable patterns and associations, and prediction, foretelling some response of interest [[10](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b10-krcp-36-003)]. Clinical data mining can be defined as the application of data mining to a clinical problem [[19](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b19-krcp-36-003)].

The algorithms of data mining are categorized as supervised, unsupervised, and semi-supervised learning. Supervised learning means to predict a known output of target, using a training set that includes already classified data to draw inference or classify prospective, testing data. In unsupervised learning, there is no output to predict, so analyzers try to find naturally occurring patterns or grouping within unlabeled data. Semi-supervised learning means to balance performance and precision using small sets of labeled or annotated data and a much larger unlabeled data collection [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b6-krcp-36-003),[20](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/" \l "b20-krcp-36-003)].

Analytic goals of medical big data are prediction, modeling, and inference; classification, clustering, and regression are common methods exploited in these contexts [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b5-krcp-36-003)]. Classification is a kind of supervised learning and can be thought as predictive modeling in which the output vector or predicting variable is categorical. Classification means to construct a rule to assign objects to one of a pre-specified set of classes (predicting variable) based solely on a vector of measurements taken on these objects. Classification techniques include logistic regression, naïve Bayesian methods, decision trees, neural networks, Bayesian networks, and support vector machine. The classification performance can be evaluated by various performance metrics tested in a test set or an independent validation set. These techniques can be used to develop a decision support system assigning a diagnosis among several possible diagnoses or to build models to predict a prognosis based on data from analysis of many biomarkers. Clustering is unsupervised learning used to find groupings in the data through the use of distance metrics. Clustering techniques include k-means clustering, principle components-based clustering, and self-organizing maps. Clustering performance can be evaluated by its performance in a subsequent supervised learning task. Clustering is frequently used in microarray data analysis or phylogenetic analysis, and also can be used in redefining of disease according to pathophysiologic mechanisms providing more specific therapeutic options. Regression is supervised learning where output variable is continuous and is a statistical analysis tool that quantifies the relationship between a dependent variable and one or more independent variables to depict trends in the data. Linear regression is the most commonly used technique in this category. Examples of its applications include a longitudinal analysis of patients’ data or decision support system [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b5-krcp-36-003),[20](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/" \l "b20-krcp-36-003)].

Iavindrasana et al [[19](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b19-krcp-36-003)] summarized nine steps in the data mining process: 1) learning of the application domain, such as determining the relevant prior knowledge of the domain and the goal of the data mining application; 2) dataset selection; 3) data cleaning and preprocessing; 4) data reduction and projection; 5) matching of the objective defined in step 1 to a data mining method; 6) choice of the algorithm and search for data patterns; 7) pattern extraction; 8) evaluation and interpretation; and 9) use of the discovered knowledge. The issues in data cleaning and pre-processing step includes data type issues such as binary, nominal, ordinal or numerical; variable domination issues in case of numerical data; redundancies among several variables; temporality issues; missing value issues; and outlier issues. Data reduction and projection step include reducing the number of variables for computation efficiency and overcoming the curse of dimensionality. During pattern extraction, the dataset can be divided into training and testing sets and the model developed in the training set is then tested in the testing set. There are many methods to split the dataset, such as cross validation, stratified cross validation, leave-one-out, and bootstrapping. The most commonly used performance metrics for evaluation are accuracy, sensitivity, specificity, receiver operating characteristic curve, precision, recall, f-measure, number of positive predictions, and number of false positives [[19](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b19-krcp-36-003)]. During the pattern extraction step, various algorithms can be tried and the algorithm showing the best performance can be chosen (so called “bake-off”); but in medical domain, transparency (or understandability) is another critical issue other than performance to be considered, as well as performance.

Medical big data have several issues related to the data themselves which although not specific to big data, needed to be considered during analyses. The issue of multiple comparisons will not be discussed in this review.

a)Missing value

As data is collected from different sources, the simplest and most overused way to handle missing values is to remove the cases with missing values, or complete-case analysis, it is valid only when missing values are assumed to be independent of both observed and unobserved data .This assumption is not realistic in most situations. Therefore, complete-case analysis in these cases may bias the conclusion. Another major drawback of the complete-case analysis is that reducing the number of data points available for analysis generally is very inefficient [[21](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b21-krcp-36-003)]. Missingness may exhibit various relationships with data already observed or unobserved data. Missing data are classified into three types: 1) missing completely at random (MCAR), 2) missing at random (MAR), and 3) not missing at random (NMAR). MCAR is missingness of which probability does not depend on either observed or unobserved data. If data are MCAR, the probability of a missing observation is the same for all entities. In these situations, complete-case analysis does not bias the scientific inference. This is rarely met in practice. MAR is missing ness of which probability does not depend on unobserved data but depend on observed data. In these cases, the process of missingness should be adjusted for all the variables that affect the probability of missingness. NMAR is missingness of which probability depends on unobserved data. There are many tool kits to handle these types of missingness including NMAR, such as in SAS, R, Stata, and WinBUGS [[21](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b21-krcp-36-003)]. There is no unique way to analyze NMAR data, nor will there ever be a program that will work well for all NMAR datasets [[21](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b21-krcp-36-003),[22](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/" \l "b22-krcp-36-003)]. It has been reported that if fewer than 10% of values were missing, many of the commonly used methods would result in similar conclusions. If between 10% and 60% of values were missing, multiple imputation was recommended. If missingness for more than 60% of the values, no method was found to give satisfactory results [[21](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b21-krcp-36-003)]. More details on incomplete data are reviewed in Wong et al [[21](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b21-krcp-36-003)].

b) Curse of dimensionality

High dimensional data are data with too many attributes compared to the number of observational units. Microarray data or next generation sequencing data are typically high dimensional datasets. In high-dimension datasets, many numerical analyses, data sampling protocols, combinatorial inference, machine learning methods, and data managing processes are susceptible to the “curse of dimensionality” [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b6-krcp-36-003)] which describes the difficulty of optimization in high dimensional datasets [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b5-krcp-36-003)].

Sparsity, multicollinearity, model complexity, computational cost to fit model, and model overfitting are the issues accompanied by high dimensional datasets [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b5-krcp-36-003)]. The space volume increases rapidly as data dimension increases; thus, the distance between data points increases accordingly. The stability of distance metrics is critical in statistical inference; therefore, this sparsity between data points affects most quantitative analyses, even for big data [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b6-krcp-36-003)]. Multi-collinearity is a phenomenon in which two or more predictor variables in a model, such as the multivariate regression model, are not independent. It violates the common regression technique assumption that requires the predictor variables to be independent of the error term (model residuals). Multi-collinearity makes a model unreliable or underpowered. Although in traditional statistical analyses with standard datasets, multicollinearity is exceptional, it may be ubiquitous in big data analyses [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b6-krcp-36-003)]. Model overfitting may cause the problem of generalizability. High dimensional data can be handled with dimension reduction [33] or feature selection [[24](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b24-krcp-36-003)]. It is important to recognize that reducing dimensionality or feature selection may cause loss of key mechanistic information. There is an overall tradeoff between a false positive rate and the benefit of identifying novel insights [[25](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b25-krcp-36-003)].

c) Bias control

Randomized controlled trials minimize bias and control confounding and are therefore considered the gold standard of design validity [[26](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b26-krcp-36-003)]. Every dataset, however, has limitations. Randomized controlled trials are frequently showing the lack of generalizability because randomized controlled trials generally are conducted under ideal conditions, among highly selected patients followed by highly qualified physicians. It is practically impossible to perform a randomized intervention for a novel biomarker without specific measures to control its multi-levels in human. It also is frequently lengthy and costly to obtain an answer for its question. Such  trials often produce heterogeneous results and a single randomized trial cannot be expected to provide a gold-standard result that applies to all clinical studies [[27](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b27-krcp-36-003)]. Well-designed observational studies may be less prone to heterogeneous results than randomized controlled trials, possibly due to a broad representation of the population at risk and less opportunity for differences in the management of subjects among observational studies, which already are diverse with respect to disease severity, treatment protocols, and coexisting illnesses. In contrast, each randomized controlled trial may have a distinct group of patients according to its specific inclusion and exclusion criteria, and the experimental protocol for therapy may not be representative of clinical practice [[27](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b27-krcp-36-003)]. Clinical studies, both observational and interventional, frequently lack the ability to provide reliable answers to their research questions because of inadequate sample sizes. Underpowered studies are subject to multiple sources of bias, may not represent the larger population, and are regularly unable to detect differences between treatment groups. Most importantly, underpowered studies can, moreover, lead to incorrect conclusions [[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b7-krcp-36-003)].

Big data analyses on various data from administrative claim database or national registries can be used to overcome these limitations. Administrative claim data have broad generalizability, large numbers of patient records, and less attrition than clinical trials; they are faster and less costly than primary data collection, and can often be linked with other datasets [[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b7-krcp-36-003)]. Big data analyses are basically observational studies, and thus share the limitations of observational studies in addition to the limitations inherent to the big data and should be considered as hypothesis-generating. It is recommended that the results of observational studies should not influence clinical practice until these hypotheses are tested in adequately powered randomized controlled trials [[28](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b28-krcp-36-003)].

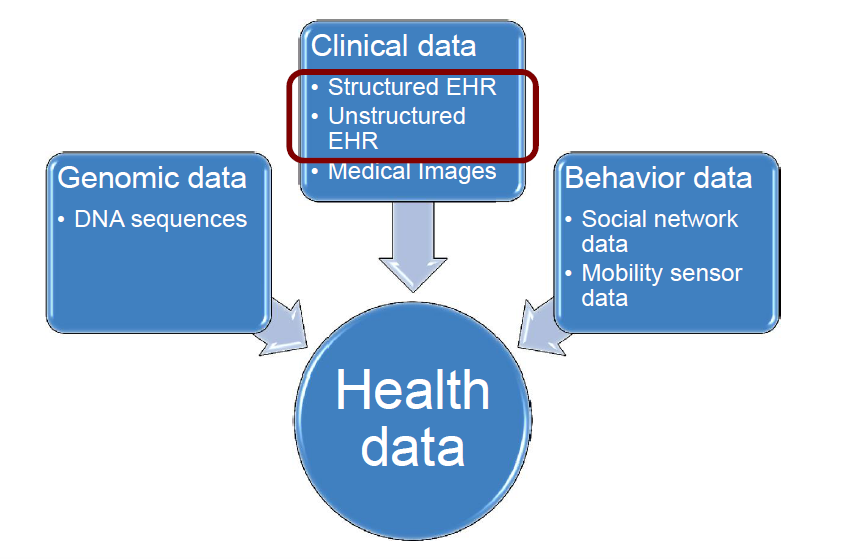
In addition to large observational studies, it can produce implausibly precise estimates of effect size that are highly statistically significant but clinically unimportant [[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b7-krcp-36-003),[28](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/" \l "b28-krcp-36-003)]. To minimize the impediments to drawing valid inferences, specific scientific best practices should be adopted, such as generation of a priori hypotheses in a written protocol, detailed analytical plans noting specific methods and safeguards against bias, and transparent reporting with justification of any changes in plans. Potential clinically important effects should be defined a priori and the results discussed accordingly [[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b7-krcp-36-003)]. There are two analytic techniques to address the problem of confounding in observational studies; propensity score analysis and instrumental variable analysis [[26](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b26-krcp-36-003)]. Propensity score is the likelihood of a patient being assigned to an intervention on the basis of his or her pre-intervention characteristics, and propensity score analysis is performed by creating pseudo-randomization of all possible measured confounders using the propensity score. Instrumental variable analysis is comparing patient groups according to an instrumental variable which is randomly distributed, rather than comparing patients with respect to the actual intervention received. A critical step in instrumental variable analysis is to find an appropriate instrument. An instrumental variable should meet three requirements: 1) to be associated with the intervention or exposure (relevancy assumption); 2) not to directly affect the outcome of interest, but to only indirectly affect the outcome through the intervention assignment (exclusion restriction); and 3) to be independent of confounders [[26](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b26-krcp-36-003)]. Theoretically, this technique aims to control for unmeasured or unknown confounders [[26](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b26-krcp-36-003)]. Recently genetic variants are used as instrumental variables to circumvent the issues of both unmeasured confounding and reverse causation in observational studies [[30](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b30-krcp-36-003)].

**1.4.1.4 What are the challenges for medical big data?**

Although the potential of big data analytics is promising, assessing the “state of science” and recognizing that, at present, the application of big data analytics is largely promissory is important [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b2-krcp-36-003)]. Therefore, it is critical to delineate some of challenges for big data applications in healthcare. First, the evidence of practical benefits of big data analytics is scarce. Second, there are many methodological issues, such as data quality, data inconsistency and instability, limitations of observational studies, validation, analytical issues, and legal issues. An effort to improve the data quality of electronic health records is necessary. In the nephrology area, although chronic kidney disease is one of the hottest area of research, its codes are not assigned in many of administration claim databases; most cases of acute kidney injury not requiring dialysis therapy are not coded in claim databases. Therefore, these practices need to be corrected. Last, clinical integration and utility is a big issue. Big data analytics need to be integrated into clinical practice to reap the substantial benefits, and clinical integration requires the validation of clinical utility of big data analytics which have been largely overlooked [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/#b2-krcp-36-003)]. It is critical to solve these challenges to fasten the application of big data technology in medical sector and thus to improve patient outcome and to reduce waste of resources in healthcare, which should be the real value of big data studies.

**1.5 Sources and Techniques for Big Data in Healthcare**

**1.5.1 Clinical data**: Data in hospitals can be stored in various ways .Interpretation of type of data depends on the nature of the data required by the doctors or the patients .Data in Health care can be divided into the following types as shown in the figure1.14 below



**Figure 1.5.1 Types of data in Medical Big data**

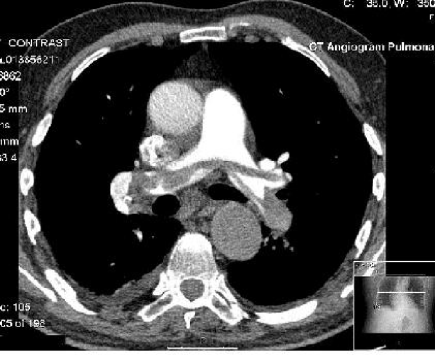
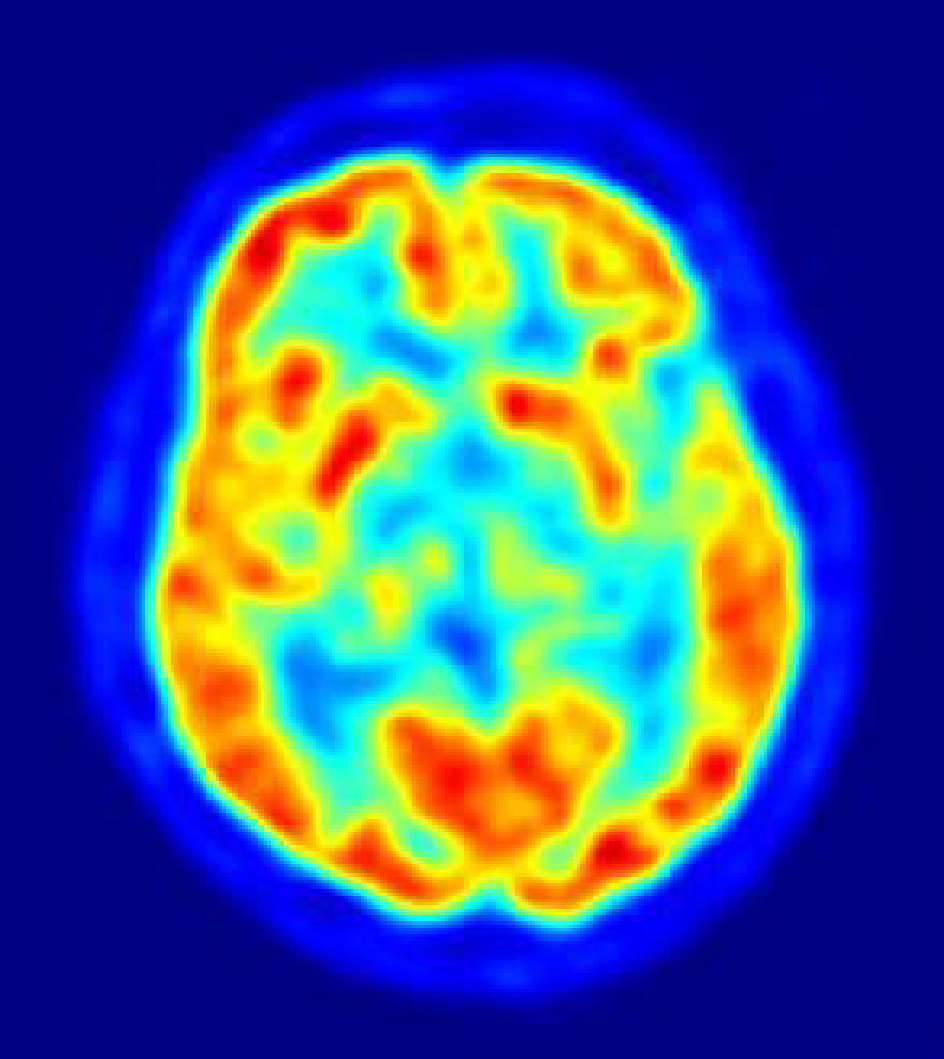
**1.5.1.1 Medical Imaging data**

Medical data is increasing by the day. By 2015, the average hospital will have two-thirds of a *petabyte* (665 terabytes) of patient data, 80% of which will be unstructured image data like CT scans and X-rays. Medical Imaging archives are increasing by 20%-40%. While much of this data is coming from files held in electronic health records, even more of it appears to be coming from [image archiving](https://medcitynews.com/2013/01/diagnostic-image-sharing-record-retrieval-business-raises-6m-in-fresh-capital/) and [communications applications](https://medcitynews.com/2013/03/startup-enabling-hipaa-compliant-mobile-image-sharing-for-on-call-docs-looking-for-1m/). 60 - 80 % of data is estimated to be unstructured and exists in the form of images, video and email. Data helps doctors make the most informed treatment decisions. The main challenge with the image data is that it is not only huge, but is also high-dimensional and complex.  Disk remains the most popular destination for healthcare data, according to survey results. 67.5% of respondents who archive data do so via disk, 50.0% use tape, 27.5% use optical media and 7.5% use cloud storage . Disk storage is expensive and to address growing data volumes, a strategy cannot be sustained over the long-term. The survey showed that most hospitals do not archive data holistically based on its age and value.

a) Clinical text analysis and mining: Automatic speech recognition (ASR) is used to create clinical notes to reduce costs associated with note creation for EMR. One needs to ensure note quality and reduce the time required to edit ASR transcripts detection models using logistic regression and conditional random field models, exploring a variety of text-based features that consider the structure of clinical notes and exploit the medical context. Different medical text resources are used to improve feature extraction.

b) Patient record analysis (x-rays)

Radiology has been a distinct medical speciality with unique technical challenges from its inception. Technical nature of X-ray image capture and perhaps more significantly the difficulty of exposing, transporting and developing images on fragile glass plates for subsequent interpretation is tedious task .Therefore, radiologists have been clinical specialists, who have been obliged to also become experts in image capture technology, Radiographic image interpretation and reporting, broad-based advances in engineering and, more recently, applications of information technology for healthcare. Improved image clarity and tissue differentiation in a number of situations has dramatically increased the range of diagnostic information and in many cases the demonstration of pathology without the requirement of invasive tissue sampling is achieved. The use of imaging for functional evaluation and cellular activity has created a new challenge for radiologists whose training has predominantly been based on the anatomical and pathological model with limited experience in physiology and cell function. The different types of image representations in radiology are as  shown in figure 1.15.

Computed Positron       Magnetic Tomography(CT) Emission Imaging (MRI)

Tomography (PET)

**Figure 1.15 types of image representations in radiology**

**1.5.1.1.1 Medical Image Retrieval System**

Current medical image techniques can be classified according to the type and nature of the features used for indexing and nature of the features used for indexing. They are text-based, content-based and semantic-based approaches. The PACS system is widely used in medical institutions where it is dedicated to the storage, retrieval, distribution and presentation of medical images. Images in PACS are annotated manually by medical officers and indexed by text keywords, hence limiting its features to textual descriptions only, besides unable to sufficiently describe the visual features of the images.

 CBIR techniques could be valuable to radiologists in accessing medical images by identifying similar images in large archives that could assist with decision support.

Two components are:

1. Image features/descriptors - bridging the gap between the visual content and its numerical representation. These representations are designed to encode color and texture properties of the image, the spatial layout of objects, and various geometric shape characteristics of perceptually coherent structures.

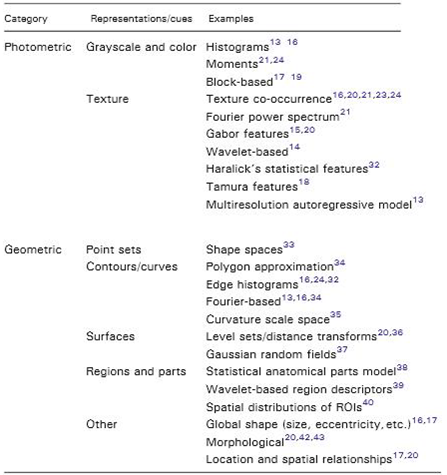
2. Assessment of similarities between image features based on mathematical analyses, which compare descriptors across different images. Vector affinity measures such as Euclidean distance, Mahalanobis distance, KL divergence, Earth Mover’s distance are amongst the widely used ones. Texture – based descriptors have become very important because they may reveal potentially fine details contained within an image structure.

**1.5.1.1.2 Geometric image features**

1. Photo-metric features:  exploit color and texture cues and they are derived directly from raw pixel intensities.

2. Geometric features: cues such as edges, contours, joints, poly lines, and polygonal regions.  A suitable shape representation should be extracted from the pixel intensity information by region-of- interest detection, segmentation, and grouping. Due to these difficulties, geometric features are not widely used.

A summary of image features and descriptors used in medical images is as shown in the table 1.2 below.



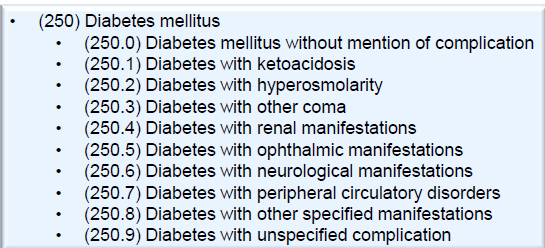
**Table 1.2 : Image Features /Descriptors used in Medical domain [44]**

**1.5.1.2 EHR(Electronic Health Records )**

It is a collection of patient and population electronically-stored health information in a digital format. These records can be shared across different [health care](https://en.wikipedia.org/wiki/Health_care) settings. EHRs may include a range of data, including [demographics](https://en.wikipedia.org/wiki/Demographics), medical history, medication and allergies, [immunization](https://en.wikipedia.org/wiki/Immunization) status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information. EHR systems are designed to store data accurately and to capture the state of a patient across time. It eliminates the need to track down a patient's previous paper medical records and assists in ensuring data is accurate and legible. It can reduce risk of data replication as there is only one modifiable file, which means the file is more likely up to date, and decreases risk of lost paperwork. Due to the digital information being searchable and in a single file, EMRs are more effective when extracting medical data for the examination of possible trends and long term changes in a patient. Population-based studies of medical records may also be facilitated by the widespread adoption of EHRs and EMRs.

1. **Billing Data –ICD codes**

ICD stands for International Classification of Diseases. ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO).  In US, most people use ICD-9, and the rest of world use ICD-10 .the advantage of this is that it is universally available its disadvantage is that it has medium recall and medium precision for characterizing patients .The figure below shows the clinical data of patients suffering Diabetes mellitus with their respective ICD codes which are universally accepted  as shown in figure 1.16.



**Figure 1.16 ICD codes of the patient suffering with Diabetes mellitus**

b)CPT codes

CPT stands for Current Procedural Terminology created by the American Medical Association. CPT is used for billing purposes for clinical services. The basic advantage is it shows high precision and the disadvantage is it has a low recall. The CPT code is as shown in the figure 1.17 below.

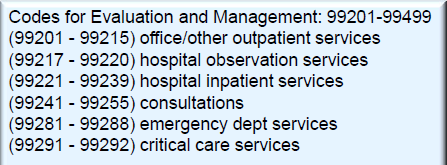


Figure 1.17 CPT codes

c) Quality standards for medicines approved by FDA

Standard code is National Drug Code (NDC) by Food and Drug Administration (FDA), which gives a unique identifier for each drug not used universally by EHR systems and too specific, drugs with the same ingredients but different brands that have different NDC. RxNorm is a normalized naming system for generic and branded drugs by National Library of Medicine. Medication data can vary in EHR systems .Either it can be in both structured or in an unstructured form. The availability and completeness of medication data vary depending upon the inpatient medication data which are complete, but the outpatient medication data are not complete. Medications usually only store prescriptions but we are not sure whether patients actually filled those prescriptions.

Clinical notes contain rich and diverse source of information. Challenges for handling clinical notes are:

1. Ungrammatical, short phrases
2. Abbreviations
3. Misspellings
4. Semi-structured information : Lab results, vital signs

**1.5.1.2.1 Structured EHR**

Structured data is entered and coded by a registered health data professional that has the ability to input the data in a way that makes it easy for you to view, share and access. Enable health information exchanges, made more popular by the Affordable Care Act correlate population health trends standardize how providers and patients share health information across different    doctors, care giving groups and health systems

**1.5.1.2.2 Unstructured EHR**

With unstructured health data, one has to open every single doctor’s note for the past few visits and take notes of your blood sugar level, for example, and then look back at the trends of all these notes to understand if you were getting healthier or not visit over visit.

Advantages and Disadvantages of Structured and Unstructured Data

• Despite evidence of positive effects such as document completenessand ease of billing,documentation approaches based solely on coded data entry also have significant limitations, including clinician time for note completion,difficulty ascertaining medical relevance and loss of information.

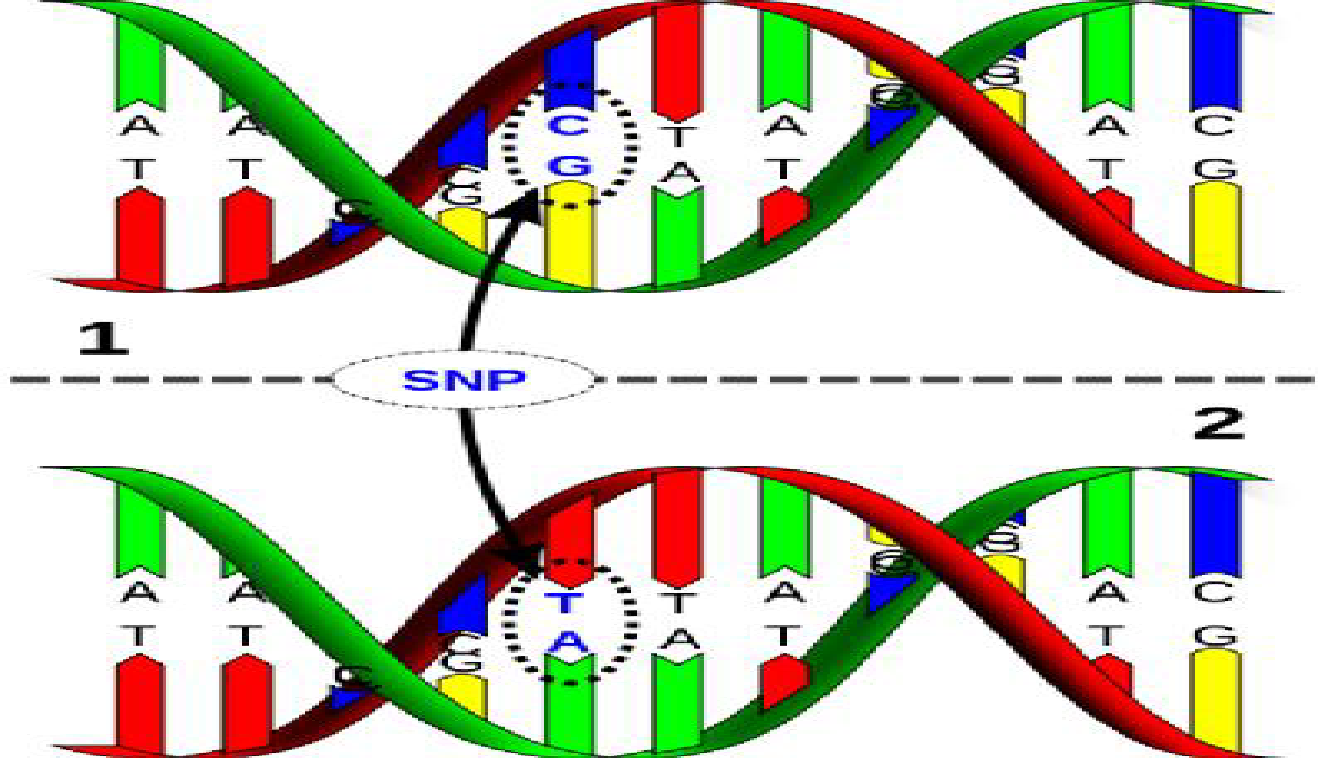
Narrative is a critical factor in evaluating medical evidence, making management decisions, and communicating medical knowledge must be more accurate and comprehensive. This should reduce length of stay of a patient in a hospital and diminish unnecessary tests. A Well-written narrative can be easier to comprehend, more edifying, and even more convincing than structured data. Data entry gives errors and due to the freedom of data it can lead to inconsistency of EHR.

**1.5.1.3 Genetic Data**

**1.5.1.3.1 Genetic data retrieval**

The human genome is made up of DNA which consists of four different chemical building blocks (called bases and abbreviated A, T, C, and G). It contains 3 billion pairs of bases and the particular order of As, Ts, Cs, and Gs is extremely important.  Size of a single human genome is about 3GB. Thanks to the Human Genome Project (1990-2003). The goal was to determine the complete sequence of the 3 billion DNA subunits (bases). The total cost was around $3 billion. The whole genome sequencing data is currently being annotated and not many analytics have been applied so far since the data is relatively new.. It costs around $5000 to get a complete genome. It is still in the research phase and heavily used in the cancer biology.

Genome-wide association studies (GWAS) :It is more relevant to healthcare practice. Some clinical trials have already started using GWAS.. Genome-wide association studies (GWAS) are used to identify common genetic factors that influence health and disease. These studies normally compare the DNA of two groups of participants: people with the disease (cases) and similar people without (controls)  (One million Loci) as shown in the figure 1.18 below . Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence differs between individuals.  SNPs occur every 100 to 300 bases along the 3-billion-base human genome.



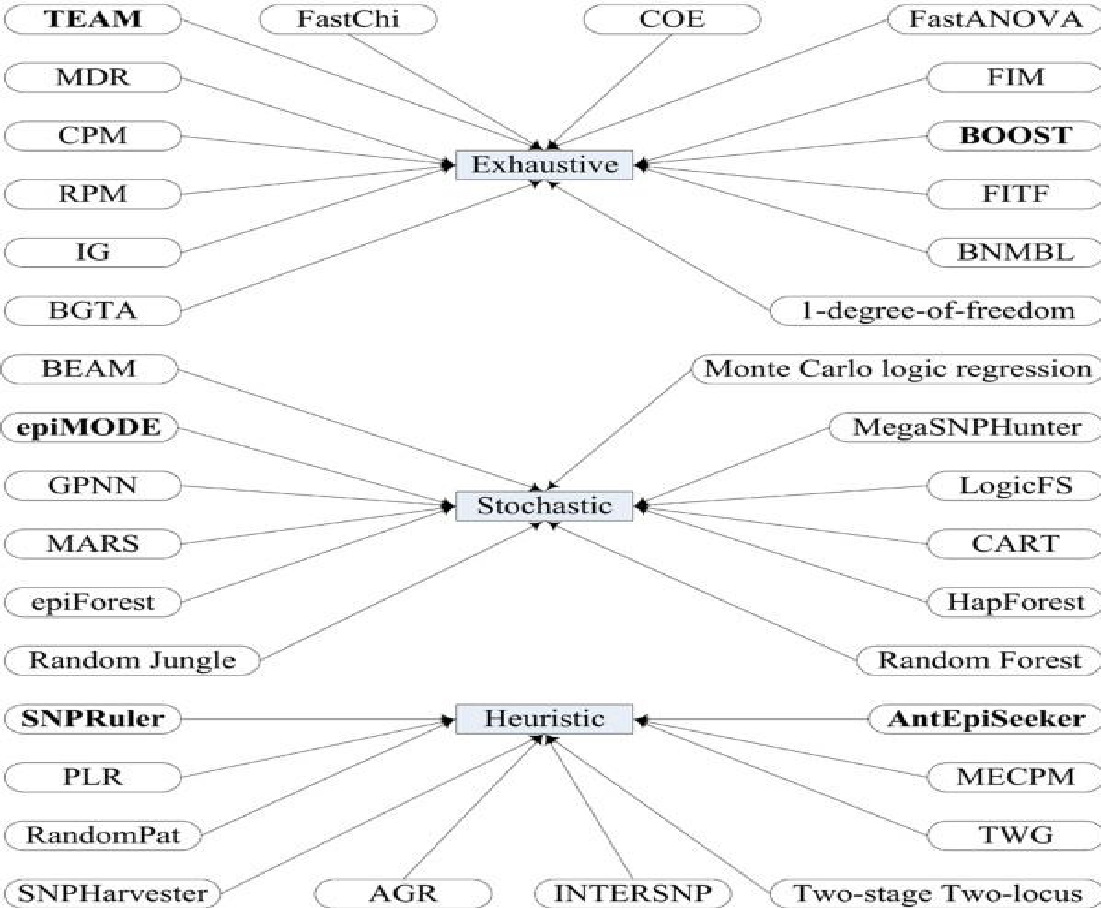
**Figure :1.18 DNA of two groups of participants**

**a) Epistasis Detection**

For simple Mendelian diseases, single SNPs can explain phenotype very well. The complex relationship between genotype and phenotype is inadequately described by marginal effects of individual SNPs. Increasing empirical evidence suggests that interactions among loci contribute broadly to complex traits.  The difficulty in the problem of detecting SNP pair interactions is the heavy computational burden. To detect pair wise interactions from 500,000 SNPs genotyped in thousands of samples, a total of 1.25 X 10 statistical tests are needed.

**b) Epistasis detection methods**

1. Exhaustive:  Enumerates all *K*-locus interactions among SNPs. Efficient implementations mostly aiming at reducing computations by eliminating unnecessary calculations [45].
2. Non-Exhaustive :
3. Stochastic: randomized search. Performance lowers when the # SNPs increase.
4. Heuristic: greedy methods that do not guarantee optimal solution.
5. The various methods are shown in the figure 1.19 below.

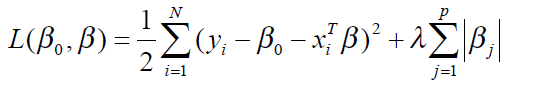


**Figure 1.19: Epistasis detection methods**

**(Courtesy Shang,Junliang et al.)**

**1.5.1.3.1.2 Sparse methods for SNP data analysis**

Successful identification of SNPs strongly predictive of disease promises a better understanding of the biological mechanisms underlying the disease. Sparse linear methods have been used to fit the genotype data and obtain a selected set of SNPs. Minimizing the squared loss function (*L*) of *N* individuals and *p* variables (SNPs) is used for linear regression and is defined  in the equation below.



where *xi* ∈ ℝ *p* are inputs for the *i*th sample, *y* ∈ ℝ*N* is the *N* vector of outputs, β*0* ∈ ℝ is the intercept, β ∈ ℝ*p* is a *p*-vector of model weights, and λ is user penalty. Efficient implementations that scale to genome-wide data are available. Various data sets for this research are available in sparSNP[42].

A recent advancement in Genomics has shown that big data can be used in scientific research .Human genome Project has taken research to the next level in studying genetics and personalized medicine .This project is explained in example 1.3.

**Example 1.3: Genomics (Big data in scientific research)**

The developments  grown around big-data in healthcare has broken the silos in scientific research. For example, the field of genomics has taken a giant stride in evolving personalized and genetic medicine with the help of big data. A good example of how big data analytics can help modern medicine is the [Human Genome Project](https://www.genome.gov/12011238/an-overview-of-the-human-genome-project) and the innumerous researches on genetics, which paved way for personalized medicine, would have been difficult without the democratization of data, which is another boon of big data analytics. The study shows that in the year 2008 there were only 5 personalized medicines available and it has increased to 132 in the year 2016.

Collecting data on patient demographics, tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status of the patient is done . It is widely used for understanding disparities related to race, age, and gender. It can be used to overlay information with other sources of data (such as water/air pollution, climate, socio-economic) to identify any correlations. It cannot be used for predictive analysis, but mostly used for studying trends.

Collecting data based on the behavioral patterns of a person in terms of emotions, actions of a person and how the thoughts of a person reflect the nature of his diseases can be captured and diagnosed .US president Obama started an initiative to understand the human brain and emotions and based on that diagnose the disease of a person called as the BRAIN initiative as explained in example 1.4.

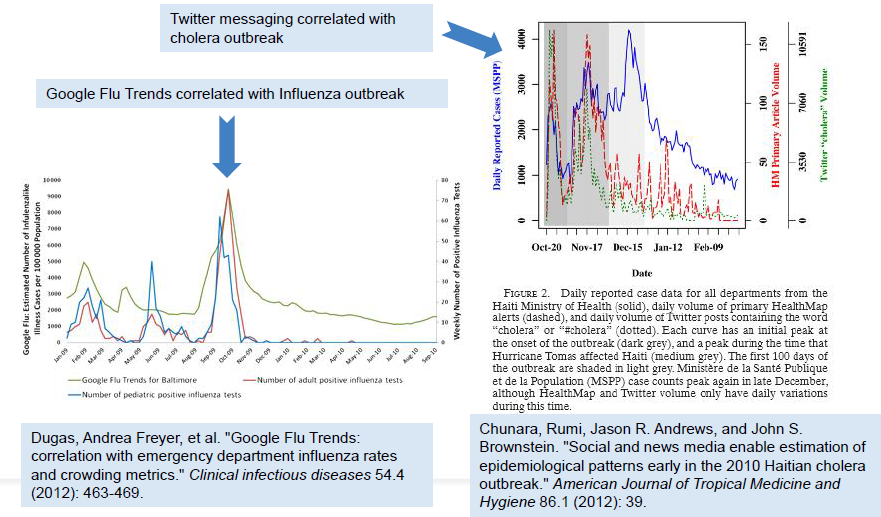
**Example 1.4: White house unveiling BRAIN initiative**

The US President unveiled a new bold $100 million research initiative designed to revolutionize our understanding of the human brain. BRAIN (Brain Research through Advancing Innovative  Neuro-technologies) Initiative. Find new ways to treat, cure, and even prevent brain disorders, such as Alzheimer’s disease, epilepsy, and traumatic brain injury. *“Every dollar we invested to map the human genome returned $140 to our economy... Today, our scientists are mapping the human brain to unlock the* *answers to Alzheimer’s.”--* President Barack Obama, 2013 State of the Union. “Advances in "Big Data" that are necessary to analyze the huge amounts of information that will be generated and increased understanding of how thoughts, emotions, actions and memories are represented in the brain.” (NSF  Joint effort by NSF, NIH, DARPA, and other private partners)

**a) Social media can sense public data !!**

During infectious disease outbreaks, data collected through health institutions  and official reporting structures may not be available for weeks, hindering early epidemiologic assessment. Social media can get it in near real-time. For  example Social networks for patients are initiated . PatientsLikeMe1 is a patient network online data sharing platform started in 2006; now has more than  200,000 patients and is tracking 1,500 diseases. Its main objective was : “Given my status, what is the best outcome I can hope to achieve, and how do I get there?”  People connect with others who have the same disease or condition, track and share their own experiences, see what treatments have helped other patients like them, gain insights and identify any patterns. Patient provides the data on their conditions, treatment history, side effects, hospitalizations, symptoms, disease-specific functional scores, weight, mood, quality of life and more on an ongoing basis.

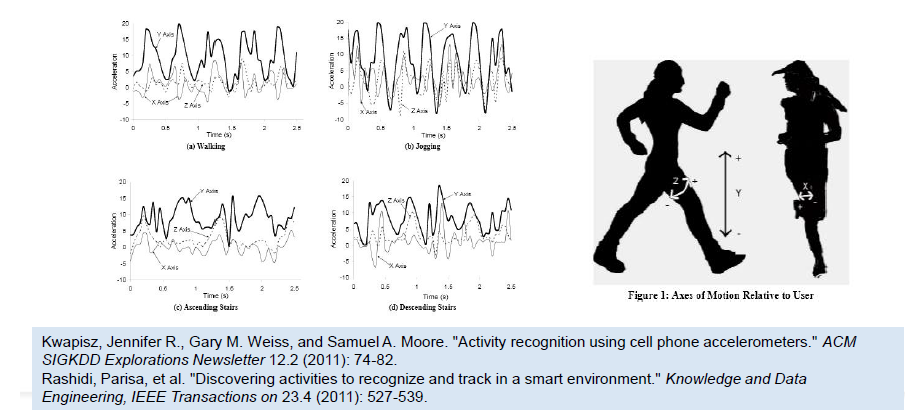
Case study 1: Patients with influenza are tracked by Google Flu to check the rate of outbreak of the disease and the crowding metrics is shown in Figure 1.20 (a) and the impact of social and news media to detect epidemiological patterns for patients suffering with cholera and its outbreak is shown in figure 1.20 (b) below.



(a)           (b)

**Figure 1.20 : a)Checking influenza rates and crowding metrics b)The effect of using Social and mews media to estimate epidemiological patterns in a cholera outbreak**

**Mobility Sensor Data** : Advancements in sensing technology are critical for developing effective and efficient home-monitoring systems Sensing devices can provide several types of data in real-time. Activity Recognition using Cell phone Accelerometersis as shown in Figure 1.21 below.



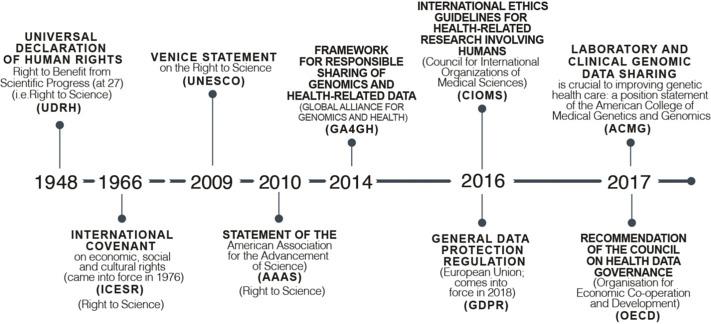
**Figure 1.21: Activity recognition using cell phones accelerometers**

**1.6 Legal and ethical issues**

Legal and ethical issues are of the foremost importance in the field of Medical big data .The recent scale of data collection and use in biomedical research and clinical care is seemingly limitless in the quest for precision medicine via the use of next generation [sequencing](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sequencing) technologies. The question one needs to ask is “will health data flow to and from medical record to the research context and back in a learning healthcare system?” [41] .With the entry onto force in 2018 of the European Union's 2018 General Data Protection Regulation (GDPR), the stage is set for an international debate on the use of Big Data in both research and health care [[41]](https://www.sciencedirect.com/science/article/pii/S2452310017300264#bib6). Perhaps it is time to move to a new paradigm where we catalyze and activate the right of all citizens to benefit from advances in science via data sharing, its benefits, and its applications as probabilistic at-risk health information becomes the “treatment”. Governance policies and security mechanisms for Big Data will emerge as equally important.  The Policy recommends transparent policies and efficient processes to ensure the timely availability and quality of data. The community needs to trust in the high quality of data if they are to inform future research, the approval of tests and drugs, and finally, the diagnosis and treatment of patients.

The twelve principles of the 2017 Organization for Economic Co-operation and Development (OECD)’s *Recommendations on Health Data Governance* on the governance of health data foresee measures to ensure more (not less) data sharing so as to maximize potential health benefits and “manage the risks while maintaining the utility of personal health data for the public interest in effective and sustainable health care systems”.

 For example from a national professional body, the American College of Medical Genetics and Genomics (ACMG) considers data sharing “crucial to improving genetic healthcare” and that research and clinical lab data should be contributed to public databases” [[38]](https://www.sciencedirect.com/science/article/pii/S2452310017300264#bib29). It maintains that the “responsible sharing of genomic variant and phenotype data will provide the necessary information to improve patient care and to empower those who are developing tests and treatments for patients to continue to improve genetic testing”. A summary of the various policies are given in the figure 1.22 below.



**Figure 1.22: A summary of policies**

**1.6.1 Clinical integration and utility issues**

Many hospitals and physicians are turning to clinical integration as a viable option to (1) increase quality, (2) reduce cost and waste in the current system to maintain margins, (3) sustain independence for physicians not ready for hospital employment and (4) position providers to take on higher levels of accountability to effectively manage utilization and the health of populations in the future. CI is commonly defined as a health network working together, using proven protocols and measures, to improve patient care, decrease cost and demonstrate value to the market[41]. The 7 key aspects are :

1. Legal options.  Defining  performance improvement initiatives to provide demonstrated value to the market.

2. Physician leadership a robust communication strategy must be there across the network and its partners. Clear goals and objectives by both employed and independent physicians will encourage dialogue and partnership formation as the strategy is implemented.   
3. Participation criteria. Member physicians or groups in the CI network must sign a participation agreement. This agreement outlines the expectations and requirements for participation in the CI program.

4. Performance improvement. Clinical quality and operational improvement projects are necessary components of a CI program CI also allows physicians to take an active role in care redesign and protocol development to increase quality, more effectively manage costs, reduce variation and eliminate unnecessary waste within the delivery system.

5. Information technology An EHR is a medical record for a patient in a physician office, hospital, ancillary care facility or ambulatory care facility. The EHR is intended to replace paper-based patient records for recording encounter-based information on each patient who receives care from the provider entity and includes electronic: data entry, order entry, prescribing and transcription.

6. Contracting options. The purpose of CI is to provide higher quality care. Creating a CI network for the sole purpose of negotiating better rates is not the purpose of CI.

7. Flow of funds. Funds are distributed based on meeting performance objectives and performance can be defined in a variety of ways.

**1.7 Big Data Challenges in Medical big data**

Inspite of the methods available in Analytics the main question still remains .The various challenges in health care to be addressed are:

1. Inferring knowledge from complex heterogeneous patient sources.
2. Leveraging the patient/data correlations in longitudinal records.
3. Understanding unstructured clinical notes in the right context.
4. Efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers.
5. Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity.
6. Capturing the patient’s behavioral data through several sensors; their various social interactions and communications.

**1.8 Enhanced Treatment Methods with Big Data**

 Focusing and joining patient clinical records to claim datasets for synthesis, by furnishing the data and supplying to third parties. Digitizing is the present trend where evolvement of mobile applications helps to function and address healthcare of patient. Using mobile apps, anyone from anywhere can synchronize the patient's data and use for the latter. Big Data in healthcare has major part for forecast catching, antidote disease, by improving quality of life and reducing obviate deaths. With the rapid growth of world's population, treatment methods are quickly changing. All the changes behind these decisions are driven out by data. In very near future, consulting a doctor even requires sharing of data as a part of his or her diagnostic toolbox.  Even to access the immense-ever growing databases, the allowable problems about the public healthcare to be spotted and rectified before they start arising. Then, studying of those diseases and finding appropriate remedy is to be prepared well in advance detecting disseminate diseases beforehand.

The groundbreaking work, often by linkup between medical and data professionals, with the potential to peer into the future and predict problems before they happen [17].

When it comes to diagnose the spread of infectious diseases, there is a problematic curve because the information from public health sources is reappraisal. In real time, present challenges like data sources on disease imitable tools from web and social networking may be helpful. Developing applications for effective forecasting of information on widespread diseases and studying how communication of social media is being used operatively.

**1.9 New Generation of Digital Health Advisors**

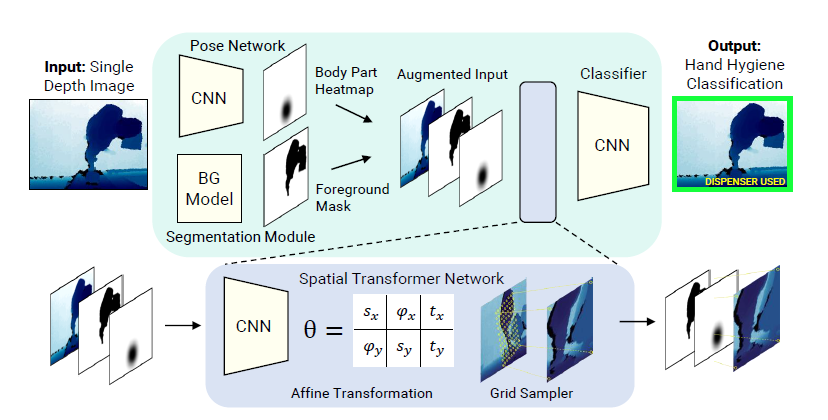
1.Artificial Intelligence: Once a data store has been built from many different sources—EHR data, payer data, device , patient survey responses, consumer health data—and has been integrated into a unified data structure, then AI can yield meaningful insights. AI, after all, is about pattern recognition, comparing a particular pattern of data around a given individual with similar (not necessarily identical) patterns found elsewhere, and making predictive recommendations based on what happened in those other situations. This is very much what clinicians do when exercising "clinical judgment"—identifying a pattern, taking into account medical problems, medications, labs values, personal and family history, and comparing it to similar patterns from the clinician's experience.

A new generation of "Health Coaches", Tele-Carers or Digital Health Advisors can be trained to make these AI-derived recommendations useful [[26](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981575/#B26)]. They need to be easy-to-use, consumer-orientated persons who can connect to the aggregated data store and the AI analytics engines that sit on top of that. They can empower consumers/patients, and reduce the demand burden on clinicians. Will they replace clinicians? No, of course not. But they will help filter the demand to those who truly need to be seen, while empowering patients with real-time, believable and personalized guidance for the more common things in day-to-day life.

So what stands in the way of Digital Health Advisors? Policy (how we pay for healthcare) needs to encourage self-care and facilitate healthy behaviors, rather than encourage in office doctor visits. And, simultaneously, health data needs to become reorganized in order to empower AI and drive the emergence of new apps and related technologies. It will be a while before we get there, but we can see the path to that new generation of healthcare technology.

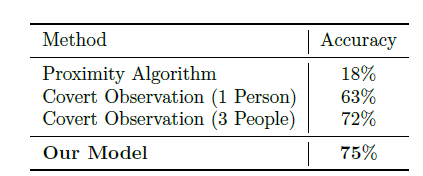
2.Computer Vision : There has been recent interest in creating smart hospitals with the aim of increasing operational efficiency and improving patient care . One use case of smart hospitals is in the prevention of HAIs, or more specifically, for monitoring and tracking hand hygiene of hospital staff. Current technologies that track hand hygiene include RFID-based systems . However such systems are limited in resolution and precision . Computer vision-based tracking systems have shown promising results in non-clinical applications such as self-driving cars and sports analytics .This example demonstrates the effectiveness of a computer vision-based method for tracking and monitoring hand hygiene compliance and identifies promising directions for future smart hospital research.

An experiment conducted by Stanford research lab [46] uses  a non-intrusive vision-based system for tracking people's activity in hospitals. It measures hand hygiene compliance and hand hygiene activity classification results. Intuitive, qualitative results that analyze human movement patterns and conducting spatial analytics conveys the interpretability of the system designed . Figure 1.23 shows the hand hygiene activity and using classifiers and transformation parameters .

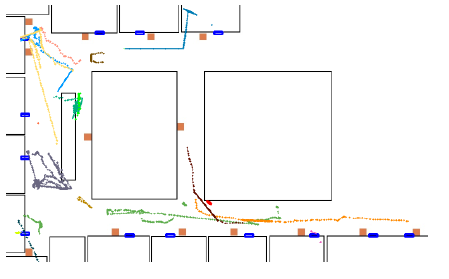


**Figure 1.23: Hand Hygiene activity classifier using various transformation parameters**

Test Cases: Ground truth hand hygiene compliance data was collected on a friday from 12 pm to 1 pm. (peak lunch time), there were a large number of visitors in the hospital. A total of 351 ground truth tracks were collected and annotated, of which 170 involved a person entering a patient room, of which 30 were compliant (i.e., followed correct hand hygiene protocol). Of the 181 tracks exiting a room, 34 were compliant. The data used for the classifier consists of 150,400 images, of 4 which 12,292 images contained pedestrians using the dispenser. The training set contained 80% of the images with the remaining 20% allocated to the test set as shown in figure :1.24 and  figure :1.25



**Figure 1.24: Comparison of hand hygiene**



**Figure 1.25: Top-down view of tracks**

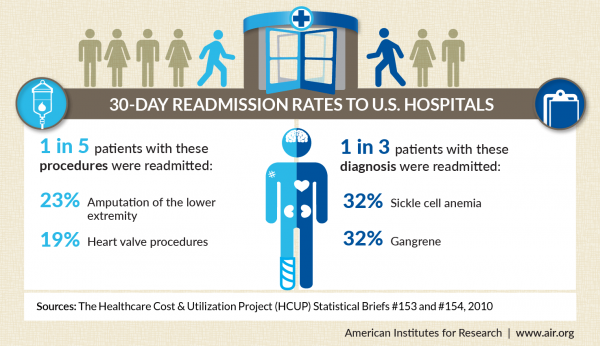
Top-down view of tracks. Blue rectangles are  doors, orange squares are dispensers, and black

Lines are walls. Different track colors denote different people. The system presented in this work is a first step towards vision-based smart hospitals and demonstrates promising results for reducing HAIs and ultimately improve the quality of patient care.

**1.10 Other Examples**

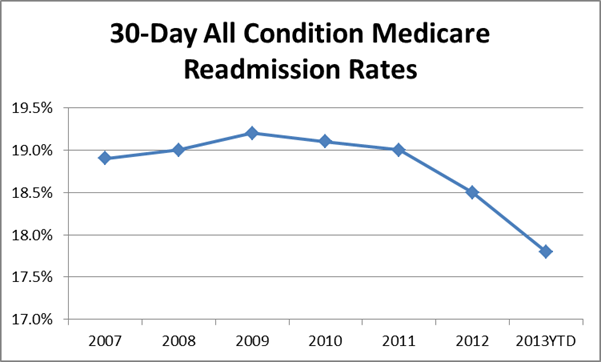
**Example 1:  Penalties for poor care 30-day readmissions**

Hospitalizations account for more than 30% of the 2 trillion annual cost of healthcare in the United States. Around 20% of all hospital admissions occur within 30 days of a previous discharge as shown in figure 1.26. This is not only expensive but are also potentially harmful, and most importantly, they are often preventable.  Medicare penalizes hospitals that have high rates of readmissions among patients with heart failure, heart attack, and pneumonia. Identifying patients at risk of readmission can guide efficient resource utilization and can potentially save millions of healthcare dollars each year. Effectively making predictions from such complex hospitalization data will require the development of novel advanced analytical models.



**Figure  1.26 :Readmissions in U.S hospitals within 30 days .**

According to the statistics nearly 19% of readmissions occurred in 2007 which later on came down to 17.5% in the year 2013 .The graph reveals these findings below in figure 1.27



**Figure 1.27:   30 day readmission rates from 2007-2013**

**Example 2: Frequency of usage of Health care apps**

Study shows that the download of health apps have increased worldwide in 2016 to nearly 1,200 million from nearly 1,150 million in the last year and 36 percent of these apps belong to the fitness and 24 percent to the diseases and treatment ones. A, 7 minute workout, a health app with three million users helps one get that flat tummy, lose weight and strengthen the core with 12 different exercises. Fooducate, another app, helps keep track of what one eats. This app not only counts the calories one is consuming, but also shows the user a detailed breakdown of the nutrition present in a packaged food.

**Example 3: Health Apps in pharma Industry**

This is an age where customers determine what they want.”*myTomorrows”* is one example of the changing look of business models, in this case, directly connecting customers and pharma [36]. Interestingly, there are signs that pharma is reaching out from its traditional medicine-centric approach. Glaxo recently announced that it is investing in electroceuticals, bioelectrical drugs that work by micro-stimulation of nerves. J&J has teamed up with Google to develop robotic surgery. In addition, they are collaborating with Philips on wearable devices such as blood pressure monitors. Novartis is working with Google (again) on sensor technologies, such as the smart lens, and a wearable device to measure blood glucose levels [30]. Sensors can provide a lot of information to support pharma development, but it is particularly important to recruit the right patients for the right clinical trials. Body sensors, once gadgets that were mainly used by athletes and runners, are now rapidly entering the general market, and consumers and pharma will soon have access to a wealth of information including not only pulse, blood pressure, ECG and respiratory rate, but also more advanced data, such as inflammation, sleep patterns, etc. A number of mobile apps which support device handling have emerged, including myDario and among others [31,32]. The Hacking Medicine Institute recently announced RANKED Health, a program to critically evaluate and rank health-focused applications and connected devices [[18](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981575/#B18)]. It has been predicted that in the near future we will look at our phone or smart watch to check health outcomes more often than we do now to check our mail or WhatsApp. A typical situation might involve an elderly person[35], recovering from a medical condition at home, linked to a combination of several connected services streaming data towards different parties, such as family members, tele-carer and physicians as shown in fig: 1.28.

An external file that holds a picture, illustration, etc.
Object name is hir-22-156-g002.jpg

**Figure 1.28: A typical situation involved an elderly person, recovering from a medical condition at home, linked to a combination of several connected services streaming data towards different parties, such as family members, tele-carer and physicians.**

**Example 4: Unnecessary ER visits**

Saving time, money and energy using big data analytics in healthcare is necessary. What if we told you that over the course of 3 years, one woman visited the ER more than 900 times? A woman who suffers from mental illness and substance abuse went to a variety of local hospitals on an almost daily basis. This woman’s issues were exacerbated by the lack of shared medical records between local emergency rooms, increasing the cost to taxpayers and hospitals, and making it harder for this woman to get good care .In order to prevent future situations like this from happening,  PreManage ED is introduced which shares patient records between emergency departments. This system lets ER staff know things like:

If the patient they are treating has already had certain tests done at other hospitals, and what the results of those tests are. If the patient in question already has a case manager at another hospital, preventing unnecessary assignments What advice has already been given to the patient, so that a coherent message to the patient can be maintained by healthcare providers.

**Example 5: CMS services**

The Centers for Medicare & Medicaid Services (CMS) have vast stores of billing data that can be mined to promote *high value care*; the same is true of private health insurers. And hospitals have attempted to reduce re-admission rates by targeting patients where predictive artificial intelligence (AI) algorithms indicate people who may be at highest risk based on an analysis of available data collected from existing patient records as shown in figure 1.29.

An external file that holds a picture, illustration, etc.
Object name is hir-22-156-g003.jpg

**Figure 1.29: The Centers for Medicare & Medicaid Services (CMS) data system.**

Underlying these and many other potential uses, however, are a series of technology, legal and ethical challenges relating to, among other things, privacy, discrimination, intellectual property, tort, and informed consent, as well as research and clinical ethics [[24](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981575/#B24)].

**References**

[1]  I.A.T. Hashem, et al, “The rise of “big data” on cloud computing: Review and open research issues,” Information Systems, 2014

[2] Jared Crapo. 2014. Big data in healthcare: Separating the hype from the reality. https://www.

[3] healthcatalyst.com/healthcare-big-data-realities. (2014).

[4] Bonnie  Feldman, Ellen M. Martin, and Tobi Skotnes. 2012. Big data in healthcare hype and hope. Technical

[5]<https://www.researchgate.net/profile/Tauseef_Naqishbandi/publication/271515514_Big_Data_CEP_and_IoT_Redefining_Holistic_Healthcare_Information_Systems_and_Analytics/links/54ca5ea70cf2c70ce521d024.pdf>

[6]  Report, Dr. Bonnie 360 (2012).

[7] 4 Kevin Hung, Yuan-Ting Zhang, and B. Tai. 2004. Wearable medical devices for tele-home healthcare

[8] Lake, P. and Drake, R. (2014). Information Systems Management in the Big Data Era. Cham: Springer International Publishing.

 [9] Pääkkönen, P. and Pakkala, D. (2015). ‘Reference Architecture and Classification of Technologies,  Products and Services for Big Data Systems‘. Big Data Research, 2 (4), pp. 166-186.

[10] Demchenko, Y., De Laat, C. and Membrey, P. (2014). ‘Defining architecture components of the Big Data Ecosystem‘. Proceedings of 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, USA, pp. 104-112.

 [11] Vol. 2. IEEE, 5384–5387.

[12] Bonnie Feldman, Ellen M. Martin, and Tobi Skotnes. 2012. Big data in healthcare hype and hope. Technical

[12] In 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEMBS’04).

 [13] Report, Dr. Bonnie 360 (2012).

[14]Techcrunch. 2014. Healthcare’s Big Data Opportunity. http://techcrunch.com/2014/11/20/healthcaresbigdata-opportunity/. (2014)

[15] Bonnie Feldman, Ellen M. Martin, Tobi Skotnes (2012). Big Data in Healthcare Hype and Hope, online <https://www.ghdonline.org/uploads/big-data-inhealthcare_B_Kaplan_2012.pdf>

[16] Burghard C: Big Data and Analytics Key to Accountable Care Success. IDC Health Insights; 2012

[17] Manyika J, Chui M, Brown B, Buhin J, Dobbs R, Roxburgh C, Byers AH: Big Data: The Next Frontier for Innovation, Competition, and Productivity. USA: McKinsey Global Institute; 2011.

[18] Manyika J, Chui M, Brown B, Buhin J, Dobbs R, Roxburgh C, Byers AH: Big Data: The Next Frontier for Innovation, Competition, and Productivity. 2011, USA: McKinsey Global InstituteGoogle Scholar

 [19] Data in Healthcare Hype and Hope, Bonnie Feldman Ellen M. Martin Tobi Skotnes,October 2012

[20] Daniel A. Reed and Jack Dongarra 2015. Exascale computing and big data. Communications of the ACM 58, 7 (2015), 56–68.

 [21]  Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: Promise and potential. Health Information Science and Systems 2, 1 (2014), 3.

 [22] Alexandros Labrinidis and H. V. Jagadish, “Challenges and opportunities with big data,” Proc. VLDB Endow. 5, pp. 2032-2033, August 2012

[24] Hsieh JC, Li AH, Yang CC. Mobile, cloud, and big data computing: contributions, challenges, and new directions in telecardiology. Int J Environ Res Public Health. 2013 Nov 13;10(11):6131–53.

[25] Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med. 2013 Sep;28(Suppl3):S660–5. doi: 10.1007/s11606-013-2455- 8.http://europepmc.org/abstract/MED/23797912.

[26] 25. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry.

[27]Overcoming Healthcare Big Data Challenges, retrieved from <http://www.mckesson.com/healthcareanalytics/healthcare-big-data-challenges/#footNote>

 [28] Song TM, Song J, An JY, Hayman LL, Woo JM. Psychological and social factors affecting Internet searches on suicide in Korea: a big data analysis of Google search trends. Yonsei Med J. 2014

[29]Naqishbandi T, Imthyaz Sheriff C, Qazi S. Big data, CEP and IoT: redefining holistic healthcare information systems and analytics. Int J Eng Res and Technol. 2015;4(1):1 Jensen, Peter B., Lars J. Jensen, and Soren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* (2012)

[30] Senior M. Novartis signs up for Google smart lens. Nat Biotechnol. 2014;32(9):856. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/25203024)]

[31] SleepBot [Internet] New York (NY): SleepBot; c2013. [cited at 2016 Jul 1] [32] RANKED Health [Internet] place unknown: publisher unknown; [cited at 2016 Jul 1].

[33] Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making sense of big data in health research: towards an EU action plan. Genome Med. 2016;8(1):71.

 [34] Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res. 2016;8(3):1560–1580.

[35] Giraldo-Rodriguez L, Torres-Castro S, Martinez-Ramirez D, Gutierrez-Robledo LM, Perez-Cuevas R. Tele-care and tele-alarms for the elderly: preliminary experiences in Mexico. Rev Saude Publica. 2013;47(4):711–717.

 [36] MyTomorrows [Internet] Amsterdam: MyTomorrows; c2016. [cited at 2016 Jul 1]. Available from:[https://mytomorrows.com](https://mytomorrows.com/).

 [37] J. Tasioulas Big data, human rights and the ethics of scientific research ABC Religion and Ethics (November 30, 2017)

[38]C. Auffray, R. Balling, I. Barroso, L. Bencze, M. Benson, J. Bergeron, E. Bernal-Delgado, N. Blomberg, C. Bock, A. Conesa, *et al.*Making sense of big data in health research: towards an EU action plan Genome Med, 8 (2016), p. 71

[39]P.N. Schofield, J. Eppig, E. Huala, M. HrabedeAngelis, M. Harvey, D. Davidson, T.Weaver, S. Brown, D. Smedley, N. Rosenthal, *et al.*Sustaining the data and bioresource commons Science, 330 (2010), pp. 592-593

[40] J.L. Contreras Bermuda's legacy: policy, patents, and the design of the genome commons Minn J L Sci Technol, 12 (2011), pp. 61-125

[41]R.R. Faden, N.E. Kass, S.N. Goodman, P. Pronovost, S. Tunis, T.L. Beauchamp ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics European Commission General data protection regulation. Regulation (EU) (2016), p. 679

[42] SparSNP package http://bioinformatics.research.nicta.com.au/software/sparsnp/

[43] Wu, Tong Tong, et al. "Genome-wide association analysis by lasso penalized logistic regression."  *Bioinformatics* 25.6 (2009): 714-721.

[44] Akgul, Ceyhun Burak, et al. "Content-based image retrieval in radiology: current status and future directions." Journal of Digital Imaging 24.2 (2011): 208-222.

[45] Shang, Junliang, et al. "Performance analysis of novel methods for detecting epistasis." *BMC.*

[46] Terry Platchek, Arnold Milstein, Li Fei-Fei, Towards Vision-Based Smart Hospitals: A    System for Tracking and Monitoring Hand Hygiene Compliance, 1Proceedings of Machine Learning for Healthcare 2017 JMLR W&C Track Volume 68.