# Big data in Clinical Trials

Mohan Mahendrakar
Indiana University
P.O. Box 1212
Bloomington, Indiana 43017-6221
mmahendr@iu.edu

## ABSTRACT

This paper will help us to understand about Clinical Trials and how Big data is impacting Clinical Trials. Oncology (Trials) is undergoing a data-driven metamorphosis. Armed with new and ever more efficient molecular and information technologies, we have entered an era where big data is helping us spearhead the fight against various deceases. This technology driven data explosion, often referred to as "big data". [3]

## KEYWORDS

I523, HID 326, Big data, Clinical, Trials, Health care, Data integration, Analytics

## 1 INTRODUCTION

A primary objective of clinical trials is gaining knowledge from studying a subset of patients which can then be applied to a much wider group of patients to improve care. In routine practice, patient care is delivered within a rich background of intrinsic and endemic confounding factors and biases associated with practices and patients. [2]

The data collected around the world from various patients, deceases form big data (collection of large data sets). Big data is currently being used on a limited basis in the clinical trials arena, but experts believe its widespread use is coming in the near future. Some hail the great promise it holds in furthering drug discovery. Others are skeptical that it will bring much value and say that enthusiasm should be tempered. [5]

According to IBM, 2.3 trillion gigabytes of data are created every dayfi?!so much that 90% of the data in the world today has been created in the last two years alone. Digital Universe estimates that by 2020, there will be 5,200 gigabytes of data for every man, woman and child on Earth.

It is predicted that the market for Big Data technology and services will reach $16.9 billion in 2015, up from $3.2 billion in 2010. This is an annual growth rate of 40 percent, which is about seven times the rate of the overall information and communications technology market. According to CB insights, health care investments in Big Data totaled $274.5 million in 2012, and it went to $371.5 million in 2013. [6]

## 2 BIG DATA & CLINICAL RESEARCH

Discovering clinical trials hidden patterns and associations within the heterogeneous data, uncovering new bio markers and drug targets. Allowing the development of predictive disease progression models. Analyzing Real World Data (RWD) as a complementary instrument to clinical trials, for the rapid development of new personalized medicines. The development of advanced statistical methods for learning causal relations from large scale observational data is a crucial element for this analysis. [4]

### 2.1 Data Integration

Having access to consistent, reliable, and well linked is one of the biggest challenges facing pharmaceutical clinical trials. The ability to manage and integrate data generated at all phases of the value chain, from discovery to real-world use after regulatory approval, is a fundamental requirement to allow companies to derive maximum benefit from the technology trends. Data are the foundation upon which the value-adding analytics are built. Effective end-to-end data integration establishes an authoritative source for all pieces of information and accurately links disparate data regardless of the source be it internal or external, proprietary or publicly available. Data integration also enables comprehensive searches for subsets of data based on the linkages established rather than on the information itself. "Smart" algorithms linking laboratory and clinical data, for example, could create automatic reports that identify related applications or compounds and raise red flags concerning safety or efficacy. [2]

Implementing end-to-end data integration requires a number of capabilities, including trusted sources of data and documents, the ability to establish cross-linkages between elements, robust quality assurance, workflow management, and role-based access to ensure that specific data elements are visible only to those who are authorized to see it. Pharmaceutical companies generally avoid overhauling their entire data-integration system at once because of the logistical challenges and costs involved, although at least one global pharmaceutical enterprise has employed a "big bang" approach to remaking its clinical IT systems. [2]

Data is being generated by different sources and comes in a variety of formats including unstructured data. All of this data needs to be integrated or ingested into Big Data Repositories or Data Warehouses. This involves at least three steps, namely, Extract, Transform and Load (ETL). With the ETL processes that have to be tailored for medical data have to identify and overcome structural, syntactic, and semantic heterogeneity across the different data sources. The syntactic heterogeneity appears in forms of different data access interfaces, which were mentioned above, and need to be wrapped and mediated. Structural heterogeneity refers to different data models and different data schema models that require integration on schema level. Finally, the process of integration can result in duplication of data that requires consolidation.

The process of data integration can be further enhanced with information extraction, machine learning, and semantic web technologies that enable context based information interpretation. Information extraction will be a mean to obtain data from additional

sources for enrichment, which improves the accuracy of data integration routines, such as duplication and data alignment. Applying an active learning approach ensures that the deployment of automatic data integration routines will meet a required level of data quality. Finally, the semantic web technology can be used to generate graph based knowledge bases and oncologies to represent important concepts and mappings in the data. The use of standardized oncologies will facilitate collaboration, sharing, modelling, and reuse across applications. [4]

### 2.2 Exascale computing

After data integration is completed, the big question is how to process such huge volume of the data? There will be use cases, e.g. precision medicine, where the promises brought by Big Data will only be fulfilled through dramatic improvements in computational performance and capacity, along with advances in software, tools, and algorithms. Exascale computers-machines that perform one billion calculations per second and are over 100 times more powerful than today's fastest systems will be needed to analyses vast stores of clinical and genomic data and develop predictive treatments based on advanced 3D multi-scale simulations with uncertainty quantification. Precision medicine will also require scaling these systems down, so clinicians can incorporate research breakthroughs into everyday practice. [4]

### 2.3 Data-driven metamorphosis

Data collected in clinical trials undergoing a data-driven metamorphosis. Armed with new and ever more efficient molecular and information technologies, we have entered an era where data is helping us spearhead the fight against cancer. This technology driven data explosion, often referred to as "big data", is not only expediting biomedical discovery, but it is also rapidly transforming the practice of oncology into an information science. This evolution is critical, as results to-date have revealed the immense complexity and genetic heterogeneity of patients and their tumors, a sobering reminder of the challenge facing every patient and their oncologist . This can only be addressed through development of clinicomolecular data analytics that provide a deeper understanding of the mechanisms controlling the biological and clinical response to available therapeutic options. Beyond the exciting implications for improved patient care, such advancements in predictive and evidence-based analytics stand to profoundly affect the processes of cancer drug discovery and associated clinical trials. [3]

### 2.4 Big data analytics

Medical research has always been a data-driven science, with randomized clinical trials being a gold standard in many cases. However, due to recent advances in omics-technologies, medical imaging, comprehensive electronic health records, and smart devices, medical research as well as clinical practice are quickly changing into Big Data-driven fields. As such, the healthcare domain as a whole - doctors, patients, management, insurance, and politics - can significantly profit from current advances in Big Data technologies, and from analytics. [4]

### 2.5 Machine Learning

Many healthcare applications would significantly benefit from the processing and analysis of multimodal data - such as images, signals, video, 3D models, genomic sequences, reports, etc. Advanced machine learning systems can be used to learn and relate information from multiple sources and identify hidden correlations not visible when considering only one source of data. For instance, combining features from images (e.g. CT scans, radiographs) and text (e.g. clinical reports) can significantly improve the performance of solutions. [4]

## 3 CHALLENGES

Big pharma companies typically keep their cards close to the vest because it costs so much to develop a drug throughout its lifetime. From discovery to prescription pad, a typical medication can take twelve years and $4 billion to shepherd through its lifecycle, a significant investment that would be hard to recoup if everyone had the secret to the newest blockbuster pill, especially since only ten percent of drugs ever make it to market. [1]

Although there is already a huge amount of healthcare data around the world and while it is growing at an exponential rate, nearly all the data is stored in individually. Data collected by a clinic or by a hospital is mostly kept within the boundaries of the healthcare provider. Moreover, data stored within a hospital is hardly ever integrated across multiple IT systems. For example, if we consider all the available data at a hospital from a single patientfis perspective, information about the patient will exist in the EMR system, laboratory, imaging system and prescription databases. Information describing which doctors and nurses attended to the specific patient will also exist. However, in most of cases, every data source mentioned here is stored in separate silos. Thus, deriving insights and therefore value from the aggregation of these data sets is not possible at this stage. It is also important to realize that in today's world a patient's medical data does not only reside within the boundaries of a healthcare provider. The medical insurance and pharmaceuticals industries also hold information about specific claims and the characteristics of prescribed drugs respectively. Increasingly, patient-generated data from IoT devices such as fitness trackers, blood pressure monitors and weighing scales are also providing critical information about the day-to-day lifestyle characteristics of an individual. Insights derived from such data generated by the linking among EMR data, vital data, laboratory data, medication information, symptoms (to mention some of these) and their aggregation, even more with doctor notes, patient discharge letters, patient diaries, medical publications, namely linking structured with unstructured data, can be crucial to design coaching programs that would help improve people's lifestyles and eventually reduce incidences of chronic disease, medication and hospitalization. [4]

## 4 CONCLUSION

The recent surge in big data initiatives in health care is expected to have a positive impact on clinical trials. Increased standardization of common data elements and nomenclature should assist in streamlined trial design and exchange of data. Standardize between trials and will allow easier multi-study analysis. Standardization

and quality improvement efforts go hand in hand with a maturing big data infrastructure providing collateral benefits to data curation for trials. [3]

## ACKNOWLEDGMENTS

The authors would like to thank to Professor and TAs for guiding in making the better paper.

## REFERENCES

[1] Jennifer Bresnick. 2014. Big pharma opens up big data for clinical trials, analytics. (July 2014). https://healthitanalytics.com/news/big-pharma-opens-up-big-data-for-clinical-trials-analytics

[2] Jamie Cattell, Sastry Chilikuri, and Michael Levy. 2013. *How big data can revolutionize pharmaceutical R&D.* White Paper. McKinsey Center for Government. https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/public%20sector/regulatory%20excellence/how_big_data_can_revolutionize_pharmaceutical_research.ashx

[3] Taglang G and Jackson DB. 2016. *Use of "big data" in drug discovery and clinical trials.* Article. Molecular Health GmbH, 69115 Heidelberg, Germany. https://doi.org/10.1016/j.ygyno.2016.02.022

[4] Dr. Adrienne Heinrich, Aizea Lojo, Dr. Alejandro Rodrguez Gonzlez, Dr. Andrejs Vasiljevs, Chiara Garattini, Cristobal Costa-Soria, Dirk Hamelinck, Elvira Narro Artigot, Prof. Ernestina Menasalvas, PD Dr. habil. Feiyu Xu, Dr. Felix Sasaki, Prof. Frank Mller Aarestrup, Gisele Roesems fi?? Kerremans, Jack Thoms, Marga Martin Sanchez, Marija Despenic, Mario Romao, Matteo Melideo, Prof. Dr. Miguel A. Mayer, Prof. Dr. Milan Petkovic, Dr. Nenad Stojanovic, Nozha Boujemaa, Patricia Casla Mag, Paul Czech, Prof. Roel Wuyts, Sergio Consoli, Dr. rer. Nat. Stefan Rping, Stuart Campbell, Dr. Supriyo Chatterjea, Prof. Dr. Ir. Wessel Kraaij, Wilfried Verachtert, Dr. Wouter Spek, and Ziawasch Abedjan. 2016. *Big Data Technologies in Healthcare.* techreport. Big data value association. http://www.bdva.eu/sites/default/files/Big%20Data%20Technologies%20in%20Healthcare.pdf

[5] F. Hoffmann-La Roche Ltd. 2013. *Understanding Clinical Trials.* techreport. GPS Public Affairs, 4070, Basel, Switzerland. https://www.roche.com/dam/jcr:1d4d1b52-7e01-43ac-862f-17bb59912485/en/understanding_clinical_trials.pdf

[6] Dr. Sarika Vanarse. 2014. *BIG DATA BREATHES LIFE INTO NEXT-GEN PHARMA R&D.* techreport. Wipro, DODDAKANNELLI, SARJAPUR ROAD, BANGALORE - 560 035, INDIA. http://www.wipro.com/documents/big-data-breathes-life-into-next-gen-pharma-RD.pdf