ORIGINAL RESEARCH PAPER
Volume-7 | Issue-7 | July-2018 | PRINT ISSN No 2277 - 8179
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

BIG DATA - BIG CHALLENGE, BIGGER OPPORTUNITY

Clinical Research

Dr Omar Salim Akhtar* MBBS, MS (General Surgery), DNB (General Surgery), MRCS (Edin, UK), DNB (Urology)Lecturer in Urology,Super Specialty Hospital,Government Medical College, Srinagar, J&K – 190014 *Corresponding Author

Dr Sabahat Rasool MBBS, MS (OBGYN), DNB (OBGYN), MRCOG (UK) Infertility Specialist and Reproductive Medicine Consultant, Srinagar, J&K – 190014

ABSTRACT

Introduction. Big Data is defined as data that is so large that normal data analytics cannot process it adequately. The spectre of the 'Big Data' Revolution hangs above us. Recently, the ability of gadgets to store immense amounts of data electronically has increased and so has the capability to analyse them in ways that were hitherto impossible.

Methods. Recent scientific articles, webpages and news items were studied to assess the role of Big Data in Medical Science in general and in Urology and Infertility in particular.

Results. This article introduces the concept of Big Data, highlights examples of its use in medical science, explains its limitations, and the various ways to use it to further clinical research. Big Data will find use in diverse fields, but nowhere will its effects be felt more than in the field of Medical Science. It will possibly open up new fields of observation, correlation, and analysis, that were impossible to understand earlier. Urology, and Infertility, as fields that have embraced new technology in the past, will not be spared from the over-arching effect of Big Data.

Conclusion. The ability of Urologists and Infertility Professionals to use the explosion of available data to their use will determine the future direction of this specialty.

KEYWORDS

The Digital Age and Big Data

The world passed into the 'Digital Age' in 2002. In that year, the information contained within Digital Information storage systems (e.g., Hard Drives, Compact Discs) surpassed the information contained within Analog Information storage systems (e.g., Books, Videotapes). Since then, there has been an exponential rise in the amount of digital information that is generated and stored[1]. By the end of 2016, the data spreading through Internet Protocol (IP) traffic crossed the zettabyte (ZB) threshold (1 zettabyte = 1 trillion gigabytes), generating 1.1 ZB in 2016, which will double to 2.0 ZB in 2019[2]. In the period of 2014-2019, the mobile data traffic is expected to grow 10-fold, to reach 300 exabytes (EB) annually in 2019, or about 24.3 EB per month[3] (1 exabyte = 1 billion gigabytes). The Digital Age also saw major changes in the field of healthcare, with the development of Electronic Health Records (EHR) and Picture Archiving and Communication Systems (PACS) which have become ubiquitous in the modern hospital setting.

The Digital Age also brought with it the ability to generate, store and analyse this enormous amount of data generated. It is these large sets of data which are generated every day, that have been given the name, 'Big Data.' According to the McKinsey Global Institute, 'Big Data' refers to, 'datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.'[4] Big data has also been defined as, 'large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information.'[5] It is generated across every industry, and as a service industry, Healthcare is no different. Healthcare in the United States is believed to have generated over 150 exabytes of data in 2011[5]. On further analysis, the potential of this data to modify the daily practise of medicine is immense - from preventive measures after picking up early signs of an epidemic through social media, to proactive intervention in intensive-care settings by picking up early-warning signs of impending infection through active monitoring of patient-care data, to establishing norms and standards by studying outcomes that are deemed to be the best.

Table 1. Definition of Big Data

Definition of Big Data
'The totality of data related to patient care and well-being make up the 'big data' in healthcare6'
'Ultralarge bodies of data that have not been prospectively limited in size or scope by the intent to address specific research questions or disease conditions, and that grow continuously and rapidly7'

What is Big Data in Healthcare?

According to Raghupathi W, 'The totality of data related to patient care and well-being make up the 'big data' in healthcare.'[6] In an Editorial[7], the term 'Big Data' has been described as referring to, 'ultralarge bodies of data that have not been prospectively limited in size or scope by the intent to address specific research questions or disease conditions, and that grow continuously and rapidly.' Such data can be generated from EHR, PACS, medical devices, patient-reported outcomes, social media use, and online patient forums[7]. According to the McKinsey Report[4], the healthcare industry generates data in all four major areas of Big Data - Video, Image, Audio, and Text/Numbers, with large amounts generated in Image and Text/Numbers. In an average Urology Department, the data generated can be from all four areas, with video generated from Endourology and Laparoscopic Urologic procedures, images from X-rays, Ultrasound scans (USG), Computed Tomography (CT) scans, etc, text/numbers from patient files, doctor's' notes, prescriptions, and audio from teaching and discussion sessions and patient feedback.

Table 2. Difference between Normal Data and Big Data

Quality	Normal Data	Big Data
Data time	Static	Continuum
Data type	Structured	Structured, Unstructured and Semi-structured
Data size	Smaller (Gigabytes)	Larger (Exabytes)
Analysis	Possible using desktop-based software, biostatistics	Needs multiple nodes or super-computers, complex computational programming
Scale in healthcare	Studying sets of individuals	Studying entire populations

Characteristics - The 4V's

Big data, besides being obviously large in terms of size, has certain characteristics. Data is collected in various forms - patient records, imaging, laboratory tests, doctor notes, patient-entered information - over the course of time, this represents the '*Volume*' of data. This data is being collected in real time - across medical devices, prescription orders, admission and discharge recording, for example, representing the '*Velocity*,' of data. Previously static medical data, like X-rays and lab reports, are now becoming dynamic - with lab report graphs comparing day-to-day and even hour-to-hour parametres of patients, and 3D imaging showing real-time motion of body organs. In intensive-care settings, the monitoring of multiple parametres of sick patients generates real-time data of multiple types across different units of measurements - this heterogeneity of data is called, '*Variety*' of data. In health care, another quality of Big Data has been added - that of '*Veracity*.'[6] Veracity is data assurance - that the data is correctly entered and correctly evaluated - and is of paramount importance as life-saving decisions may have to be made based on such data. Incorrect data, followed by incorrect data analysis, poses a grave danger to patient outcomes.

In another paper[8], the fourth 'V' has been defined as '*Value*.' Value has been defined as the ability to use Big Data to improve operational efficiency and business processes, thereby making patient care effective and affordable, which is the ultimate aim of Big Data analysis.

The data that is collected can be structured, unstructured or semi-structured. *Structured data* includes input record fields such as patient name, date of birth, address, physician's name, hospital name and address, and other information easily coded into and handled by automated databases. It is easy to store, access, recall, analyse, and manipulated by machine[6]. *Unstructured data* is that which is not compartmentalised - file notes, CT scans, paper prescriptions are examples of data that is not structured in a manner to make it easily usable[6]. *Semi-structured data* is data that is generated across devices - such as fitness devices, cell-phone based monitors, which is generated continuously, but not integrated with older EHR systems. There is immense potential in collating this data with structured forms of data.

Architecture and Framework of Big Data

Theoretically, Big Data cannot be broken down and analysed on a stand-alone machine, like a laptop or desktop. It requires to be broken down into parts and accessed and analysed on separate systems, called '*Nodes*.' The availability of online, free-to-use, platforms for assessing Big Data has made access to analytics easier, but the limitations remain in terms of trained manpower and data entry. Big Data analytics are complex, require programming skills, and are intensive[6].

The steps to using Big Data[6]:

1. Concept formulation - Is there a need for Big Data Analytics in the case?
2. Proposal - What is the problem being addressed? Background material? Big Data approach?
3. Methodology - Proposition, Data collection methods, Data analysis methods.
4. Deployment - Evaluation, validation and testing.

An example can be the study of patient parametres in intensive-care settings - to determine the at-risk category of patients who are prone to develop hospital-acquired infections. Already this has been put to use at the Hospital for Sick Children in Toronto (SickKids) - by using Big Data analytics from continuous body sensors which generate thousands of signals every minute, the Hospital was able to identify infants who developed nosocomial infections at least 24 hours earlier[9]. In Big Data analytics, there is a continuum of data analysis - of which the critical are two types of analytics. *Clinical analytics* bring together clinical, financial and operational data to answer questions and perform *retrospective* analysis about how healthcare organizations are running, the state of their patient populations and the effectiveness of programs. *Advanced analytics*, which are more *predictive* and forward–looking in nature, often focus on making predictions regarding at-risk patients[9].

Impact and Use of Big Data

According to an article published online, *'Big Data will transform approaches in healthcare that have long defined the industry. No longer will we pool data from individuals to predict what happens at the population level. Instead, population data will be so comprehensive that it will accurately predict what happens to an individual patient.*[10]*'* There will be six major areas where Big Data will change healthcare - supporting research, transforming data into information, supporting self-care, supporting providers, increasing awareness, and pooling of data[10].

Supporting Research

There are many examples of Big Data being used to expand the frontiers of knowledge. An illustrative example is that of the Virtual Physiological Human (VPH)[11]. Once established it will enable collaborative investigation of the human body as a single unit - at present, anatomy, physiology, pathology, and as an extension, medicine, surgery and other clinical sciences, are organ-system or organ-based. The understanding of the complex interactions between systems that actually defines normal human physiology is yet to be completely elucidated. The VPH aims to be descriptive, integrative and predictive.

Transforming Data into Information

Colossal amounts of data are generated each day - Big Data can potentially convert these immense quantities of data into useful bits of information. This is illustrated by the Google Flu study, which used Google Search entries to track the flu epidemic in 2009 and showed that by the study of search engine entries, it was possible to predict flu incidence with the accuracy of a Governmental organisation with its own population-based screening methods[12].

Supporting Self-care

This is a major aim of Big Data - by the analysis of 'wearable' monitors - it may be possible to predict important health care events before they become clinically obvious. An example of this is wearable technologies to monitor heart-rate, blood pressure, and respiratory rate - which can continuously transmit data to a processing centre, and alert care providers to a sudden rise in blood pressure or a rise in heart rate, signalling the start of a cardiac event. Another example is various Android Apps which have been developed to locate and alert nearby certified emergency response providers after an event has taken place.

Supporting Providers

An example of this was illustrated in the article by Dr Gawande in 2012[13]. In this article he talks about his visit to a 'control room' of multiple intensive-care units in the United States - where a doctor and two nurses monitor care in intensive-care units of different hospitals in real-time. The 'control room' would communicate with the teams in the hospitals and inform them of any 'warning signs' that might turn up on the monitoring equipment. In the long-term, the purpose of such 'control rooms' is to monitor intensive-care and make sure standards of care are up to a standard across hospitals, and also pooling resources to bring better results in care. Big Data helps develop benchmarks for the 'warning signs' and develops the standards of care expected, by continuous monitoring of patient-related outcomes.

Increasing Awareness

Kaiser Permanente, a healthcare provider company based in California, USA, used analysis of clinical data with cost data to generate a data set - further analysis of which revealed an important adverse drug effect and the subsequent withdrawal of a drug from the market[5,6]. Analysis of Twitter feeds have been used to track epidemics[5].

Pooling of Data

The pooling of data collected across multiple sources will generate important information - nowhere is this more evident than in the field of genomics. In the coming decades, almost every patient will have their genome analysed - already infants are having portions of their genomes analysed in some cases[10]. The standard 20-variable blood chemistry panel could realistically be replaced with one that tests for tens of thousands of known markers. This data will be able to tell physicians if a person with genetic risk factors for disease development is starting to manifest illness[10].

Table 3. Analytics continuum - From Reporting to Analytics[9]

Steps	1 ➔	2 ➔	3 ➔	4
Type of data analysis	Transaction Reporting	Data Integration Data Warehouse	Clinical Analytics	Advanced Analytics
Sources (examples)	- Spreadsheets - Data sources - Manual collection	- Unified data sources - Clinical data sources - Departmental data	- Enterprise-wide data - Enterprise analytics - Clinical outcomes	- Evidenced-based medicine - Personalised healthcare - Fraud detection - Behaviour
Typical questions asked	What happened? How many? Where?	What is the problem? What action is needed?	What is happening? What is the productivity?	What could happen next? How can we achieve the best outcomes?

Impact of Big Data on Urology

Big Data has already been put to use in some studies related to Urology. A recent study by Dr Caroli and his group used a patient-specific computational model to determine vascular flow rates post arterio-venous reconstruction. They were able to show that the computational tool, 'reliably predicted patient-specific blood flow increase resulting from vascular access surgery and subsequent vascular adaptation.' The innovative approach may help plan the 'most appropriate fistula' configuration to optimize access blood flow for hemodialysis, potentially reducing the incidence of vascular access dysfunctions and the need of patient hospitalisation[14].

In the study of male infertility, the recent explosion of 'omics,' ('omics' - the study of the collective and characterisation of biomolecules that translate into structure, function, and dynamics of the organism) integrated with Big Data analytics has the potential to uncover the exact role of the >2300 genes involved in regulating spermatogenesis. According to the Carell et al, improvements in technology and the progressive lowering of costs to perform large-scale 'omic' studies has led to important advancements in biological knowledge, including an improved understanding of spermatogenesis, sperm function, and post-fertilization events. It is likely that future studies will uncover low frequency polymorphisms and epigenetic variations that result in abnormal protein expression resulting in infertility or risk to the offspring[15].

The American Urological Association (AUA) also started a new initiative called the AUA Quality Registry (AQUA). The AQUA Registry is a US-based national urologic disease registry designed to measure and report healthcare quality and patient outcomes. Through the aggregation and organization of both clinician- and patient-reported data on diagnostic and therapeutic interventions, outcomes and resource utilization, AQUA is expected to provide urologists with a definitive resource for informing and advancing urology within the United States[7,16].

With the cost of sequencing the entire genome of a human being coming down by manifold in the last decade, the ability to identify genetic associations with disease processes has been multiplied. In Urology, which is plagued by multifactorial and heterogenous disease with a common clinical presentation - such as prostate cancer or calculous disease - this 'omics' revolution pitted with Big Data will lead to greater understanding of disease processes. With the completion of the VPH, it is expected that chronic kidney diseases will be identified earlier through better understanding and monitoring of 'high-risk' subsets.

Recently, Google and Ethicon, a subsidiary of Johnson and Johnson, formed a new company called, Verb Surgical[17,18]. This new company hopes to lead to, 'A new future... which will involve machine learning, robotic surgery, instrumentation, advanced visualization, and *data analytics*.' It is believed that the latest technology and data will be put together to develop 'New Robotic Surgery products.'[17] The use of this new 'Robot' is expected to change the technology currently in use for Robotic surgery, and the possibility of the development of an autonomously functioning robot sometime in the future cannot be ruled out[19].

Challenges of Big Data

In the IHTT report[5] on Big Data in Healthcare it was mentioned that there are challenges on many fronts - industry readiness (Is the healthcare industry willing to collect and share data?), data usability and trustworthiness (Is the data authentic?), data fragmentation (Each Department - IT, Labs, Outpatients, building its own data), architectural and infrastructural issues (Old computers, old software compatibility), and data ownership, security and privacy concerns while using Big Data. The plethora of data collection centres begs the question - *'Whose data is it anyway?'*

Many Big Data scientists have clearly spelled out other formidable challenges that need to be overcome before Big Data can reach its potential. Analytic software should provide the support that is necessary and needed to solve health care related problems, the software should be easily available, there should be continuity, easy to use, privacy should be respected, and security maintained. Big Data will realise its full force, if 'real-time' analytics are deployed. Another major challenge is the standardisation of data[6].

In its editorial, European Urology has highlighted that Big Data is not a panacea for all medical problems[7]. An increase in data size does not reduce bias in all cases - and gives rise to *unmeasured confounding* - which arises when a causal factor is incorrectly assigned to a disease because of its statistical association alone - the true variable may actually be related to this variable, but buried in the data. Big Data brings it with challenges of reducing data 'noise' by determining essential elements and interactions.

Piai and Claps have highlighted in their paper the difficulty of using unstructured data and the need to combine unrelated sets of data and information to develop a complete picture[8]. The challenges they have mentioned are those of governance that favours collaboration, as opposed to segregation, maintaining the veracity of data and contextual meaningfulness, the protection of data, and attracting the right talent to Big Data analysis.

In their article on Big Data analytics on in healthcare, Belle et al have highlighted the problems of proper data acquisition, data storage and retrieval, and data aggregation, which will be essential to use of Big Data[20].

In a recent online article, the worry of healthcare managers regarding Big Data preparedness was considered the 'top challenge' facing managers in 2016. Only 10% of managers in the US felt that their systems were equipped well enough to take advantage of Big Data[21].

Table 4. Challenges of Big Data

Challenges of Big Data
Industry readiness
Data usability and trustworthiness
Data fragmentation
Architectural and infrastructural issues
Manpower - talent and skill shortage
Data ownership
Security
Privacy

Steps to developing Big Data Enabled Systems in Healthcare

According to the white paper by IBM[9], there are five major steps towards becoming a data-driven healthcare organisation.

1. Define data sources

Each healthcare organisation should determine which sets of data - both internal and external - are of importance to its functional improvement.

2. Set data quality metrics and assess and improve quality

Data entry should be standardised. Once data quality has been benchmarked, it must be normalised to formats, structures and quality to enable comparison.

3. Integrate data sources

All data can then be transmitted to a storage facility. Outsourcing data storage will also enable the organisation to claim high-end 'real-time' data analytics.

4. Identify analytic needs

This is very important - without an aim for data collection - all the organisation collects is data 'noise.' The healthcare organisation needs to determine exactly what it is looking for - is it predictive models for future trends? Is it early detection of intensive-care infections? - and accordingly the data analytics will need to be applied. Executive reports and data scientist reports require two vastly different sets of data.

5. Secure and manage the data lifecycle

Data security must be confirmed, and decisions made on any future use of the data for audit, retrospective study, or quality improvement, and further storage of data once used. There are expenses involved in storage of such large amounts of data - the organisation needs to be aware of the cost of historical data storage as well.

Conclusions

Big Data is transformative technology at its best. It has the ability to give insights into common problems that previously baffled for want of conclusive data. The internet altered the way most healthcare professionals searched for information on medical science. In Urology, the lithotriptor revolutionised calculous disease care, the laparoscope challenged open surgery, and the robot is a marriage between the laparoscope and modern technology. In the same vein, Big Data, although nascent, has great potential, but at the same time, this potential will be realised when major challenges, such as those outlined above, are overcome[6]. Big Data is not a panacea for all the ills of healthcare, and it is important to ask the right questions when dealing with Big Data and its analytics[7].

In Urology, Big Data has the potential to uncover the underlying genetic interactions which result in male infertility, chronic kidney disease progressing to renal failure, and the heterogeneity of prostate cancer. Better understanding of diseases at the 'omic' level will ultimately lead to better, personalised, treatment options. The ability of Big Data to be shared widely and simultaneously across borders, is an opportunity for doctors to overcome cultural and linguistic barriers which hitherto prevented collaborative research.

The authors believe that Big Data, by virtue of its size and scale, is eventually going to replace the way medical scientific knowledge is stored, referenced, and recalled. Just as Index Medicus is now a relic of the past, Big Data promises to make printed textbooks, printed journals, and online libraries a thing of the past as well. Large-scale computation will present the most meaningful, relevant and up-to-date information to answer any query, as opposed to the 'key-word' search algorithms in place today. Just like in sport, where the players' and teams' records are constantly updated in real time, patient-related data will be updated in real time - enabling quick access. The author calls this, *'Research at the Speed of Thought.'*

References

1. Hilbert, M & López P. 'The Worlds Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025), 60-65.
2. Cisco Visual Networking Index: The Zettabyte Era - Trends and Analyses. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html
3. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019 White Paper. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html
4. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, USA. Executive Summary, p1.
5. IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-report-download-today/
6. Raghupathi and Raghupathi. Big data analytics in healthcare: promise and potential. Health Information Science and Systems. 2014, 2:3.
7. Ghani KR, Zheng K, Wei JT, Friedman CP. Harnessing Big Data for Health Care and Research: Are Urologists Ready? European Urology. 2014; 66: 975-977.
8. Piai S, Claps M. Bigger Data for Better Healthcare. IDC Health Insights, September 2013
9. IBM: Data Driven Healthcare Organizations Use Big Data Analytics for Big Gains; 2013.http://www03.ibm.com/industries/ca/en/healthcare/documents/Data_driven_healthcare_organizations_use_big_data_analytics_for_big_gains.pdf
10. http://www.healthcare.siemens.com/magazine/mso-big-data-and-healthcare-1.html?stc=wwhc101014
11. Virtual Physiological Human. Wikipedia. (Online) https://en.wikipedia.org/wiki/Virtual_Physiological_Human
12. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," Nature, vol. 457, no. 7232, pp. 1012–1014, 2009
13. Gawande A. Big Med. Annals of Health Care. The New Yorker. August 13, 2012
14. Caroli A, Manini S, Antiga L, Passera K, Ene-Iordache B, Rota S et al. Validation of a patient-specific hemodynamic computational model for surgical planning of vascular access in hemodialysis patients. Kidney Int. 2013 Dec;84(6):1237-45
15. Carrell DT, Aston KI, Oliva R, Emery BR, De Jonge CJ. The "omics" of human male infertility: integrating big data in a systems biology approach. Cell Tissue Res. 2016 Jan;363(1):295-312.
16. AQUA. (Online) https://www.auanet.org/resources/aqua.cfm
17. New York Times, Online. http://bits.blogs.nytimes.com/2015/03/27/google-and-johnson-johnson-team-for-robotic-surgery-projects/?_r=3
18. Verb Surgical, Online. http://www.verbsurgical.com/about-us/
19. http://spectrum.ieee.org/automaton/robotics/medical-robots/google-verily-johnson-johnson-verb-surgical-medical-robots
20. Belle A, Thiagarajan R, Soroushmehr R, Navidi F, Beard DA, Najarian K. Big Data Analytics in Healthcare. BioMed Research International. 2015 Article ID 370194
21. http://managedhealthcareexecutive.modernmedicine.com/managed-healthcare-executive/news/survey-big-data-presents-top-challenge-managed-care-executives