JHI JOURNAL OF HEALTH INFORMATICS

EDITORIAL

Big Data and Health: Some challenges and opportunities for Brazil

Umberto Tachinardi
Associate-Dean for Biomedical Research – School of Medicine and Public Health - UW-Madison – U.S.A.

Introduced in the early days of the 21st century, "Big Data" (BD) is a common term used after META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being "three-dimensional"[1]. The "3Vs" defined by Laney, were: Volume, Variety and Velocity. There is also reasonable consensus that *Complexity* is probably the single most important feature of anything worthy of the BD label[2]. The field of medicine is clearly a producer and a consumer of BD, particularly since the monumental efforts to decode the Human Genome that coincidentally were also concluded in the early 2000s[3]. The growing popularity of Electronic Health Record (EHR) systems, high-resolution imaging, real-time recording of physiological signals, complex laboratory tests, and wearable and home care sensors all contribute to volume and complexity of data collected and used in Medicine[4].

Human biology is very complex, the result of an extraordinary number of factors and intricate interactions among them[5]. Integrating data from multiple "*omics"* sources (e.g. genomics and proteomics) and clinical systems, advances understanding why, how, where and when diseases develop[6]. The ability to factor a larger number of contributing parameters can lead to production of personalized diagnosis, treatments and prevention schemes[7]. Targeted cancer therapies are a good example of advancements generated by sophisticated those analytics[8]. BD analytics can play a key role in helping us meet the new challenges (e.g. aging population, raising costs of healthcare, new environmental threats). The translation of discoveries into effective solutions entails sophisticated new tools and methods. Health professionals, from all backgrounds, need to be trained to understand the new paradigms (e.g. precision/personalized medicine, pharmacogenomics, evidence based-medicine, and clinical-decision systems)[9] and adapt to them.

The proliferation of information requires re-engineering of most of the infrastructural elements of data storage, transmission and computing. This means new hardware (e.g. clusters, grid) and networking solutions, but also new software solutions (e.g. unconventional database architectures like Hadoop). This is a challenge to virtually every country in the world and Brazil is no exception.

BD is only a "Big Deal" if it can generate *value*. The gigantic datasets (e.g. genomes, sensors, and images) are only as valuable as our capacity to extract new knowledge or actionable information from them. But most of the familiar tools and methods currently available are inadequate in the BD era. For most of the information buried in unstructured data (e.g. texts, images and genomes), new approaches must be used (e.g. natural language processing and machine learning) to convert these large stores of data to computable forms. This new business requires highly skilled developers (data scientists) and operators (data miners), as well as data-savvy users (e.g. clinicians and epidemiologists) that are able to understand, produce and consume BD analytics. This is an international problem, but each nation must develop and implement appropriate strategies.

National public health databases, like the ones hosted by DataSUS in Brazil[10], are good examples of costly resources that can be leveraged by using modern analytic methods. DataSUS's impressive collection of health data (including epidemiology systems, socio-economic registries and vital records), is sourced from dozens of health systems used across the whole country. In terms of complexity, variety and volume, the datasets are big enough to reach BD status. While some basic visualization and cohort building tools are available, the heavy data crunching necessary for much more complex tasks like predictive analytics and risk analysis requires computing power and methods that are not currently available and must be built.

BD in health raises important privacy and security issues[11]. This is not an easy task: even de-identified datasets sanitized to obscure Protected Health Information (PHI) can be vulnerable to unauthorized malicious re-identification. Analysts, health care providers, and researchers must seriously protect against those threats, while ensuring that those very protections still enable the promising future of health analytics. We must engage all stakeholders, including citizens, scientists, and politicians, in a discussion that will develop

the appropriate safeguards, taking into account socio-cultural considerations. In the United States the Health Insurance Portability and Accountability Act (HIPAA)[12], although far from perfect and not particularly attuned to BD challenges, nevertheless provides a framework for such a discussion. There is currently no equivalent of HIPAA in Brazil, but one is needed to provide legal guidelines to the appropriate use of patients' data.

BD is the journey not the destination and there is a wonderful world of new discoveries ahead of us. The idea of a *Learning Healthcare System*[13] will certainly be closer to reality if we succeed in this journey. Trough JHI, the Brazilian Society for Health Informatics (SBIS) is helping this future happen. This road is paved with creativity, ingenuity and hard work. A new science is developing, and the Health Informatics community is equipped to help improve health with data, Big Data.

REFERÊNCIAS

1. https://en.wikipedia.org/wiki/Big_data
2. Kaisler S, Armour F, Espinosa JA, Money W. Big Data: Issues and Challenges Moving Forward. System Sciences (HICSS). 2013: 995-1004.
3. https://www.genome.gov/11006929
4. N. Versel. How Hospitals are Dealing With Big Data.http://health.usnews.com/health-news/hospital-of-tomorrow/articles/2013/10/15/how-hospitals-are-dealing-with-big-data
5. Wolkenhauer O, Fell D, De Meyts P, Blüthgen N, Herzel H, Le Novère N, Höfer T, Schürrle K, van Leeuwen I. SysBioMed report: advancing systems biology for medical applications. IET Syst Biol. 2009 May;3(3):131-6.
6. Special Issue on Big Data - JAMIA, Volume 19, Issue e1, 1 June 2012
7. Ashley EA. The Precision Medicine Initiative: A New National Effort. JAMA. 2015;313(21):2119-2120.
8. Li L, Wang H. Heterogeneity of liver cancer and personalized therapy. Cancer Lett. 2015 Jul 23. pii: S0304-3835(15)00473-5.
9. Lehmann CU, Longhurst CA, Hersh W, Mohan V, Levy BP, Embi PJ, Finnell JT, Turner AM, Martin R, Williamson J, Munger B. Clinical Informatics Fellowship Programs: In Search of a Viable Financial Model: An open letter to the Centers for Medicare and Medicaid Services. Appl Clin Inform. 2015 Apr 15;6(2):267-70.
10. http://datasus.saude.gov.br/
11. Shoenbill K, Fost N, Tachinardi U, Mendonca EA. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. J Am Med Inform Assoc. 2014 Jan-Feb;21(1):171-80.
12. http://www.hhs.gov/ocr/privacy/index.html
13. Institute of Medicine. The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine). Washington, DC: The National Academies Press, 2007.