BIG DATA AS A SOLUTION TO EXPLOIT THE ENORMOUS EXISTING VOLUME OF HEALTH DATA

28.4.2020

Author's note: This article is part of a series published in SeAMK eJournal about cutting-edge technologies (Artificial Intelligence, Big Data, Robotics, and Internet of Things) that will revolutionize the healthcare in the near future.

Currently an immense volume of medical data is being generated through heterogeneous, and sometimes unstructured, data sources as: electronic health records; medical imaging; genomic sequencing; clinical records; medical devices and Internet of Things, clinical and pharmaceutical research, smartphones and wearables, web and social media, healthcare administrative services, health-related publications and clinical reference data. It is estimated that the sheer volume of health data reaches approximately 25 000 Petabytes (1Petabyte equals to 10^6 Gigabyte) during 2020 [1]. In addition, the emergence of powerful software has created conditions for large and high-quality healthcare datasets to be efficiently collected and analyzed, even in a real-time manner, and so to allow a transition from experience-based clinical decision making to an evidence and personalized driven with the aim at reducing cost and improving patient satisfaction and quality of care.

The concept of Big Data is defined as datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze [2]. Furthermore, Big Data is characterized by the 6 V's: *Volume* (Big Data produces a humongous amount of data); *Variety* (data coming from miscellaneous sources); *Velocity* (data comes and is processed at different rates); *Veracity* (determining the usefulness of data by the quality factor); *Valence* (the connectedness of data); *Value* (Data processed should give valuable insights) [3]. From a healthcare perspective, the Big Healthcare Data Analytics (BHDA) can be defined as the use of statistical, cognitive, predictive, contextual and quantitative models for efficient and fast decision making that allow to plan, forecast, manage resources, discover business values and insights in a timely fashion [2]. The different operations over raw health data provided by that BHDA can compose a work pipeline with the following stages: capture, cleaning, curation, integration and interoperability, storage, pre-processing, indexing, search, sharing, transfer, mining, analysis, and visualization.

One of the main values created by BHDA in healthcare is the development of analytical techniques (Data Mining, Machine Learning, Neural Networks, Natural Language Processing) which provides personalized health services to users and, by using automated algorithms, supports: human decision-making, identification of patient care risk, clinical recommendation for patient empowerment, disease prediction models, medical errors avoidance, ER workflow enhancement and epidemics tracking . Moreover, it allows to achieve cost-effectiveness and efficiency in the healthcare process delivery, coordination and administration. The automated algorithms are developed by using many frameworks and tools as Hadoop, HDFS, MapReduce or Spark which are based on new distributed architectures along with high memory capacity and processing

power [2]. Examples of some of BHDA solutions are “MedAware” [4], “GEMINI”[5], “Google Flu Trends”[6], “COHESY”[7], “CARE” [8], “AEGLE” [9]; and they have a miscellaneous field of application as genomics, elderly care, chronic diseases (cardiovascular, cancer, respiratory, diabetes, Parkinson), gynecology, nephrology, oncology, ophthalmology or urology.

BHDA must face multiple challenges to become reality in clinical routine. BHDA must prevent to provide healthcare clinicians with “black box” answers that will lead to a misuse and lack of interest of such solutions, so the interpretability of results should be a priority. Moreover, data trustworthiness and privacy are of utmost importance in BHDA along with interoperability protocols and standards to ensure a correct confidentiality and integration among existing healthcare IT systems that sometimes lack of an appropriate IT infrastructure. From a technical point of view, data quality and consistency, semantic interoperability or real-time process capabilities are yet other challenges for BHDA. To cope these issues several strategies can be deployed, namely: implementing data governance, developing an information sharing culture, employing security measures, training key personnel to use big data analytics or incorporating cloud computing into the organization’s Big Data analytics.

Pedro A. Moreno Sánchez
RDI Expert/ Researcher
PhD, PMP®
Seinäjoki University of Applied Sciences
School of Health Care and Social Work

References

[1] G. Harerimana, B. Jang, J. W. Kim, and H. K. Park, “Health Big Data Analytics: A Technology Survey,” *IEEE Access*, vol. 6, pp. 65661–65678, 2018, doi: 10.1109/ACCESS.2018.2878254.

[2] S. Bahri, N. Zoghlami, M. Abed, and J. M. R. S. Tavares, “BIG DATA for Healthcare: A Survey,” *IEEE Access*, vol. 7, pp. 7397–7408, 2019, doi: 10.1109/ACCESS.2018.2889180.

[3] C. Shah, M. Shaikh, D. Shah, and K. Samdani, “A Review on Big Data Practices in Healthcare,” in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, Mar. 2019, pp. 1–6, doi: 10.1109/ICSCAN.2019.8878687.

[4] H. Mack, “MedAware Raises $8 Million for Software to Reduce Prescription Errors,” *Wall Street Journal*, Aug. 18, 2017.

[5] Z. J. Ling *et al.*, “GEMINI: an integrative healthcare analytics system,” *Proc. VLDB Endow.*, vol. 7, no. 13, pp. 1766–1771, Aug. 2014, doi: 10.14778/2733004.2733081.

[6] A. F. Dugas *et al.*, “Influenza Forecasting with Google Flu Trends,” *PLOS ONE*, vol. 8, no. 2, p. e56176, Feb. 2013, doi: 10.1371/journal.pone.0056176.

[7] “Towards Collaborative Health Care System Model – COHESY – IEEE Conference Publication.” https://ieeexplore.ieee.org/abstract/document/5986197 (accessed Apr. 20, 2020).

[8] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási, “Time to CARE: a collaborative engine for practical disease prediction,” *Data Min. Knowl. Discov.*, vol. 20, no. 3, pp. 388–415, May 2010, doi: 10.1007/s10618-009-0156-z.

[9] D. Soudris *et al.*, “AEGLE: A big bio-data analytics framework for integrated health-care services,” in *2015 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS)*, Jul. 2015, pp. 246–253, doi: 10.1109/SAMOS.2015.7363682.

[10] N. Mehta and A. Pandit, “Concurrence of big data analytics and healthcare: A systematic review,” *Int. J. Med. Inf.*, vol. 114, pp. 57–65, Jun. 2018, doi: 10.1016/j.ijmedinf.2018.03.013.