Eur J Breast Health 2018; 14: 61-62
DOI: 10.5152/ejbh.2018.0101

The Big Data Revolution for Breast Cancer Patients

Issam Ibnouhsein[1], Stéphane Jankowski[1], Karl Neuberger[1], Carole Mathelin[2]
[1]Quantmetry, Data Science Consulting 128 rue Faubourg Saint-Honoré, Paris, France
[2]Unité de sénologie, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

Keywords: Breast cancer, big data, machine learning, health data

Cite this article as: Ibnouhsein I, Jankowski S, Neuberger K, Mathelin C. The Big Data Revolution for Breast Cancer Patients. Eur J Breast Health 2018; 14: 61-62.

The origins of Big Data date back to 1941, when the first references were made to the notion of “information explosion” in the Oxford Dictionary of English. James Maar has highlighted in 1996 in a report of the National Academy of Sciences the concept of “massive data set” (1). But it was only in 1997 that the precise term ‘Big Data’ first appeared in an article in the Digital Library of the Association for Computing Machinery (2), referring to the technical challenge of analyzing large sets of data. It has since been used to designate “structured or unstructured data, whose very large volume requires adapted analysis tools”. Web giants (Google, Amazon, Facebook, Apple, Twitter) have developed such tools over the past decade, ensuring a constant marginal cost of data exploitation, regardless of volume.

Today, Big Data is characterized by 5V: Volume, Velocity, Variety, Veracity and Value of the data exploited. The drop in storage prices and the increase in computing capacity are at the origin of the large volumes and the high speed of data processing. The variety of data (images, texts, databases, connected devices, etc.) is mainly due to the increasing digitization of information media. Finally, the truth of the data, from which the value of the work is derived, is a central issue for any project of automated data analysis. Indeed, an algorithm is really powerful if the data are numerous, exact, and well-adapted to the question to be solved. Multiplying sources and crossings without worrying about the quality of the data can only lead to erroneous results, notably in the domain of health. The development of Big Data has been accompanied by the emergence of “Open Data” which correspond to data generated and maintained by various organizations and made available to citizens and businesses.

The 5 V, however, are insufficient to characterize the essence of the innovation brought by Big Data. The mastery of these algorithms is at the heart of the business of data scientists.

The diagnosis and treatment of breast cancer have rapidly evolved during the past three decades. Part of this evolution is due to individual or organized breast screening programs and progress of breast imaging technics. Indeed, a sub-domain of artificial intelligence called “machine learning” makes it possible to build algorithms able to accumulate knowledge and intelligence from experiments, without being human-guided during their learning, nor explicitly programmed to manage a particular task, hence their central role in the data value chain while the rest is due to the evolution of surgical techniques or medical treatments. Recently, the advent of Big Data technologies has generated a lot of interest among the medical community concerned with breast health. Indeed, the available storage capacities increased exponentially during the last three decades, thus leading to bigger volumes and variety of stored medical data (mammography scans, 3D ultrasound, MRI, genomic data, pathological data…). Until now, these data were generally exploited at an individual level during a specific period of time in order to establish a diagnosis, a therapeutic protocol, to follow the disease evolution and to estimate a prognosis for a specific patient. Moreover, only structured data, which represented a small fraction of accessible and interesting information sources, were exploited on a statistical scale. The rest was stored in data graveyards that the medical staff barely sees. The big promise of Big Data is to allow the exploitation of all data sources, including unstructured ones such as textual patients reports or images, thus influencing medical research, and ultimately patient care.

Address for Correspondence :
Carole Mathelin, e-mail: Carole.Mathelin@chru-strasbourg.fr

Received: 11.02.2018
Accepted: 13.03.2018

To understand more precisely how Big Data may revolutionize breast health care, it is necessary to consider two progresses. Firstly, the software landscape that emerged in the past decade allows the implementation of predefined operations on huge volumes and different varieties of data. Secondly, machine learning algorithms and their practical implementations in programming languages, can learn from data in order to extract patterns and correlations, and ultimately produce valuable insights. The so-called data scientists are the experts in juggling these sets of techniques.

Various medical projects based on Big Data techniques were launched in the past few years in the domain of breast health, with many implications on the understanding of prognosis and on decision making (3, 4). One of these ongoing clinical trials (ClinicalTrials.gov Identifier: NCT02810093) consists in analysing textual records of patients suffering from breast cancer. The analysis is performed using machine learning algorithms that extract and structure a wide variety of information, including medical history, risk factors, size of tumours, lymph node involvement, presence of specific biomarkers, use of different treatments or patients' evolution. Once this information is structured, a second iteration of statistical modelling is performed, with many interesting insights on specific subpopulations, the importance of certain biomarkers for prognosis, or the adequacy of the decision criteria used by the medical staff in breast cancer treatment. Their results will probably enhance our understanding of the many intricate mechanisms underlying cancer development, or therapeutic resistance.

These achievements highlight the considerable potential of Big Data techniques in breast cancer care, but also for other pathologies. Consequently, medical time is progressively changing: whilst up to 30 years were necessary to gather data using cohorts to answer a specific question, such as the impact dietary factors, physical activities, alcohol consumption, night work on breast cancer development, Big Data technologies now allow us to analyse all sorts of existing data to isolate the relevant information for answering these numerous medical questions. More generally, the medical research paradigm is slowly shifting from the logic of hypothesis verification on ad hoc constructed populations, to the discovery of interesting correlations after the data collection phase.

Transdisciplinarity is central to the success of these innovative studies. Learning a semantics that is shared between the medical staff and data scientists takes time, and the breast disease units' experience in transversal organizations will be very helpful in defining a frame for these collaborations. In addition to the medical staff and data scientists, Big Data projects should involve patients, and more generally civil society, since only a strict compliance with privacy rules can ensure their success and viability.

Moreover, a truly international vision of the future of breast cancer care is necessary, and more generally of how data exploitation can be at the service of public health policies, while the time frame to develop these projects may be very quick, the huge amount of data being available today. Junior doctors should get involved early in their training in the Big Data research thematic. Indeed, in a very near future, it will be up to them to define the interesting questions that need to be answered, the data sources where to look for answers, the data that need to be collected, and the ethical frames for Big Data projects. More generally, all the medical staff needs to progressively learn how to incorporate the new possibilities offered by the Big Data revolution to the day-to-day practice with patients.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study has received no financial support.

References

1. Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences; Commission on Physical Sciences, Mathematics, and Applications; National Research Council, «Massive Data Sets: Proceedings of a Workshop», National Academy Press, 1996.
2. Cox M, Ellsworth D. Application-controlled demand paging for out-of-corevisualization. Proceedings of the 8th Conference on Visualization'97, 1997.
3. Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E, Tallet A, Chabannon C, Extra JM, Jacquemier J, Viens P, Birnbaum D, Bertucci F. A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Res Treat 2011; 126: 407-420. (PMID: 20490655) [CrossRef]
4. Dheeba J, Singh NA, Selvi ST. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. J Biomed Inform 2014; 49: 45-52. (PMID: 24509074)[CrossRef]