Big data for better cancer care

Cancer is defined by the abnormal and uncontrollable growth of cells and is the second leading cause of death worldwide (You and Henneberg, 2018). The heterogeneity in disease manifestations, various treatment options, patients' preferences and deciding which treatment is optimal for an individual patient are some of the most significant challenges in health care.

To make reliable decisions, we need large amounts of data. According to IBM, approximately 2.5 quintillion bytes of digital data are generated from humans' daily activities (e.g. from groceries, demographic and administrative medical records) (Arora, 2019). The acquisition or extraction of these extensive and voluminous data is colloquially known as big data. There is no standard definition for big data, but in simplistic terms, it is a significantly large and complex dataset, which is impossible to adequately manage and process with traditional software (Deng, 2014). However, Gartner defines big data as:

> **'Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.' (https://www.gartner.com/it-glossary/big-data/)**

Mr Akuli B Osong, PhD Candidate, GROW-School for Oncology and Developmental Biology, Department of Radiation Oncology (MAASTRO), Maastricht University Medical Center, Maastricht, The Netherlands

Professor Andre Dekker, Professor of Clinical Data Science, GROW-School for Oncology and Developmental Biology, Department of Radiation Oncology (MAASTRO), Maastricht University Medical Center, Maastricht, The Netherlands

Dr Johan van Soest, Postdoctoral Researcher, GROW-School for Oncology and Developmental Biology, Department of Radiation Oncology (MAASTRO), Maastricht University Medical Center, Maastricht, The Netherlands

Correspondence to: Mr AB Osong (biche.osong@maastro.nl)

Although big data is perceived by many as just an extensive collection of datasets, there is more to the definition of 'big' than just the volume aspect.

Big data anatomy in health care

Like the human body, big data has several distinct features. To have a better grasp of big data in health care, it helps to look at four aspects which are regularly used to describe big data.

Volume

With the advances in medical technology, a considerable amount of patients' data can be generated from clinical practice (e.g. diagnostic and therapeutic procedure information, omics) and public health (e.g. wearables and mobile devices).

Velocity

The switch from recording patients' medical information on paper to an electronic health records system has increased the amount of data which can be accessed and shared via a secure information system at any time. This reduces the time between data generation and processing for real-time (or near real-time) decision making.

Variety

Health-care providers use different media formats (e.g. numbers, images, videos, text or audio) which can be classified into three main groups (e.g. structured, semi-structured or unstructured).

Veracity

The quality and source of data are pivotal to being able to make the right inference. The mere fact that data comes from different sources opens up the possibilities of bias, uncertainty, inconsistency, incompleteness or imprecision. All these affect the accuracy of the results and conclusions drawn from these heterogeneous sources.

Big data in health care

In 2006, Clive Humby, a UK mathematician, stated:

> **'Data is the new oil. It is valuable, but if unrefined it cannot be used.'**

This analogy is not without merit as data fuels most if not all industries (e.g. sports, agriculture, education). This quote and the accompanying hype is also influencing the health-care sector as it continuously strives to add value to patient care. These efforts can be split into two aims: to improve operational efficiency and to increase clinical insights and excellence (in terms of personalized or precision medicine).

Improve operational efficiency

Owing to the increased incidence of cancer worldwide, combined with rising health-care costs, operational efficiency in clinical practice becomes important. Optimizing or automating clinical workflows could reduce human workload, increasing the number of patients which can be treated and reducing the time spent receiving medical care.

Radiotherapy, one of the most widely used and effective cancer treatment options, is already using big data to automate image contouring and treatment planning tasks (Lustberg et al, 2018). This automation results in less variability in treatments, because it limits unnecessary human variation. In the end, big data and proper application of methods can result in a shift of repetitive and time-intensive tasks usually performed by humans to computer supervision.

Staff allocation per shift has been one of the challenges in clinical management. Finding the perfect balance between over-staffing departments (while providing satisfactory patient care) and under-staffing (with unsatisfactory care) is a challenging task. Big data can help to solve this dilemma by estimating the expected number of patients visiting an emergency department on a daily and hourly base (Morzuch and Allen, 2006).

> “...big data can be used to predict treatment outcomes, allowing humans to make decisions between various treatment options (and their outcomes).”

Clinical excellence

As our understanding of the human body increases, selecting the best treatment option for a particular patient becomes more complicated. The information at hand is much larger than the human brain can comprehend (Abernethy et al, 2010), resulting in treatment outcomes which are unpredictable by humans (Oberije et al, 2014). Fortunately, big data can be used to predict treatment outcomes, allowing humans to make decisions between various treatment options (and their outcomes). These predictions can be used to target patients who need rapid interventions or those will benefit from (neo)adjuvant treatments.

Eventually, applying such big data strategies could have benefits in terms of costs, quality and time. For example, knowing which patients do not need (neo) adjuvant treatment strategies would save costs for this treatment, improve quality of life (in terms of treatment burden or adverse treatment effects), and reduce treatment time.

This also results in the concept of value-based care (Gentry and Badrinath, 2017), where the interventions are weighed against their actual benefit to the patient. As an example, in personalized treatment where the patient's specific tumour information is used to ascertain treatment response, big data can identify tumour-specific patterns information contained within the pixels of medical images which could assist the physician and radiologist in their diagnostics and decision-making process (Gillies et al, 2016).

Impediments of big data in health care

William Cowper said in the late 18th century, 'Variety is the spice of life'. However, too much variety without a proper definition of terms can be detrimental for big data in health care. Big data within this domain are coming from a variety of sources which when combined can give a richer insight or help provide better care. But the major challenge in data sharing among different sources within health care (both within and without health-care organizations) is how to preserve participants' privacy while still benefiting future patients. Data are currently stored in systems which are optimized for a specific (clinical) purpose, which generally does not target re-use by users who are not the data holder or vendor of the system (Lustberg et al, 2017).

To address these problems of data sharing in health care will require proper data descriptions with formal and well-defined ontologies (terminologies, properties and their respective relationships) within the domain. These ontologies are standardized in a machine-readable format (with human-readable representations) and shared among the different participants (Lustberg et al, 2017). The emphasis on having a common controlled vocabulary or standardizing an ontology in health care is essential to avoid participants being unable to use each other's data.

Benefits of FAIR big data in health care

To fully benefit from big data in health care, data and databases should firmly adhere to the FAIR principles (Wilkinson et al, 2016), i.e. they should be findable, accessible, interoperable, reusable (FAIR) but still respecting patients' privacy and medical confidentiality. These principles do not state that all data should be publicly available but urge that at least their descriptions (e.g. what kind of data are available) are published, including the means to contact the data (owners or maintainers), and which (semantic) representations are used. Using rich metadata descriptions of the actual data should improve data reuse and make data FAIR.

Conclusions

Big data is an emerging field which has come to revolutionize the way we think and act in the health-care sector. Without a standard definition of the term, many have taken this to mean collecting a vast amount of data without looking at the true meanings of big data. The complexity of big data is an essential factor in health care, as data are coming from different sources, with accompanying different data privacy regulations, ownership and security implications. **BJHM**

KEY POINTS

- Health care always had big data, before the adoption of the term itself.
- Big data still needs semantics.
- The use of big data is needed to advance personalized medicine.
- Big data in health care is scattered over multiple institutes, with the complexity of managing data privacy, ownership and security.
- Making data findable, accessible, interoperable, reusable (FAIR) is an essential catalyst for the use of big data (in health care).

Abernethy AP, Etheredge LM, Ganz PA et al. Rapid-learning system for cancer care. J Clin Oncol. 2010 Sep 20;28(27):4268–4274. https://doi.org/10.1200/JCO.2010.28.5478

Arora S. 2019. Data Science vs. Big Data vs. Data Analytics. (accessed 13 May 2019) https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article

Deng J. Big data in radiation oncology: challenges and opportunities. Cancer Sci Res Open Access 1(2): 1–2. https://doi.org/10.15226/csroa.2014.00111

Gentry S, Badrinath P. Defining health in the era of value-based care: lessons from England of relevance to other health systems. Cureus. 2017 Mar 6;9(3):e1079. https://doi.org/10.7759/cureus.1079

Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016 Feb;278(2):563–577. https://doi.org/10.1148/radiol.2015151169

Lustberg T, van Soest J, Jochems A et al. Big Data in radiation therapy: challenges and opportunities. Br J Radiol. 2017 Jan;90(1069):20160689. https://doi.org/10.1259/bjr.20160689

Lustberg T, van Soest J, Gooding M et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol. 2018 Feb;126(2):312–317. https://doi.org/10.1016/j.radonc.2017.11.012

Morzuch BJ, Allen PG. 2006. Forecasting hospital emergency department arrivals. Presented at 26th Annual Symposium on Forecasting, Santander, Spain: 11–14 June

Oberije C, Nalbantov G, Dekker A et al. A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: A step toward individualized care and shared decision making. Radiother Oncol. 2014 Jul;112(1):37–43. https://doi.org/10.1016/j.radonc.2014.04.012

Wilkinson MD, Dumontier M, Aalbersberg IJJ et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. https://doi.org/10.1038/sdata.2016.18

You W, Henneberg M. Cancer incidence increasing globally: the role of relaxed natural selection. Evol Appl. 2018 Feb;11(2):140–152. https://doi.org/10.1111/eva.12523