BJR https://doi.org/10.1259/bjr.20160689

Received: 17 August 2016 | Revised: 12 October 2016 | Accepted: 20 October 2016

Cite this article as:
Lustberg T, Van Soest J, Jochems A, Deist T, Van Wijk Y, Walsh S, et al. Big Data in radiation therapy: challenges and opportunities. *Br J Radiol* 2017; **90**: 20160689.

COMMENTARY

Big Data in radiation therapy: challenges and opportunities

TIM LUSTBERG, BSc, JOHAN VAN SOEST, MSc, ARTHUR JOCHEMS, PhD, TIMO DEIST, MSc, YVONKA VAN WIJK, MSc, SEAN WALSH, PhD, PHILIPPE LAMBIN, MD, PhD and ANDRE DEKKER, PhD

Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, Netherlands

Address correspondence to: Dr. Prof. Andre Dekker
E-mail: *andre.dekker@maastro.nl*

ABSTRACT

Data collected and generated by radiation oncology can be classified by the Volume, Variety, Velocity and Veracity (4Vs) of Big Data because they are spread across different care providers and not easily shared owing to patient privacy protection. The magnitude of the 4Vs is substantial in oncology, especially owing to imaging modalities and unclear data definitions. To create useful models ideally all data of all care providers are understood and learned from; however, this presents challenges in the guise of poor data quality, patient privacy concerns, geographical spread, interoperability and large volume. In radiation oncology, there are many efforts to collect data for research and innovation purposes. Clinical trials are the gold standard when proving any hypothesis that directly affects the patient. Collecting data in registries with strict predefined rules is also a common approach to find answers. A third approach is to develop data stores that can be used by modern machine learning techniques to provide new insights or answer hypotheses. We believe all three approaches have their strengths and weaknesses, but they should all strive to create Findable, Accessible, Interoperable, Reusable (FAIR) data. To learn from these data, we need distributed learning techniques, sending machine learning algorithms to FAIR data stores around the world, learning from trial data, registries and routine clinical data rather than trying to centralize all data. To improve and personalize medicine, rapid learning platforms must be able to process FAIR "Big Data" to evaluate current clinical practice and to guide further innovation.

Primarily because of the ubiquity of imaging in oncology, as well as many other diagnostic and therapeutic procedures, cancer data are firmly in the realm of "Big Data".[1] To make an estimate, in the past 10 years, approximately 140 million patients were diagnosed with cancer in about 100,000 hospitals globally. If one assumes a data volume (depending on the hospital) of 0.1–10 Gb of data per patient, the total volume of cancer patient data in the world is estimated to be 14–1400 petabyte of data. Specifically, data in radiation oncology can be classified as "Big Data" because: the use of data-intensive imaging modalities (Volume), the imaging archives are growing rapidly (Velocity), there is an increasing amount of imaging and diagnostic modalities available (Variety), and interpretation and quality differs between care providers (Veracity). With this deluge in data, it becomes increasingly hard to translate all these data into knowledge and subsequently leverage that knowledge to guide clinical decisions.[2] The radiation oncologist is overwhelmed with scientific literature, swiftly evolving treatment techniques and the exponentially increasing amount of clinical data.[2] To provide high-quality individualized treatments, radiation oncologists need help translating all these data into knowledge that supports decision-making in routine clinical practice.[3] Collecting these data provides its own set of challenges. The data are spread over care providers around the world, difficult to share while protecting patient privacy, non-interoperable and varying in quality.

The gold standard to assess the utility of innovations that directly affect patients is clinical trials. However, clinical trials only provide information about a select patient population, which often represents only a small percentage of the actual population. Also, clinical trials provide the radiation oncologist with little information when making clinical decisions for someone who does not (exactly) fit the trial population owing to age, comorbidities etc. On the other hand, clinical trials do provide high-quality reusable data owing to the clear definitions that are provided by trial protocols. Initiatives such as IBM Watson attempt to simplify accessing knowledge garnered from scientific literature for physicians. Patient characteristics provided by the physician are used to find and retrieve relevant publications (and possible other sources), which can aid the

physician in making precision decisions for that particular patient.[4]

To fill the gap of evidence between clinical trials and the common patient (*i.e.* one who does not fit trial inclusion criteria), data registries are being created around the world.[5] In general, the goal is to register a select set of parameters for all patients treated for a certain cancer. This results in a large patient population with high-quality data.[5] However, this requires great effort from care providers to collect these data, which limits the number of elements recorded, as someone has to fill in a form, digital or paper, to provide the registries with the data specified in the registry protocol. There are some early initiatives[6] to automatically provide the registries with the data they require by data mining the Oncology Information Systems. In theory, this should work well for all structured data (*e.g.* the fractionation schema or age), but is challenging for data which are usually recorded in free text (*e.g.* smoking behaviour or comorbidities). Registries in general give insights in practice but are not designed to guide decisions for individual patients. ASCO CancerLinQ is the exception; it aims to create a "super" registry with a learning approach on routine healthcare data in medical oncology.[7] Cancer screening shows promising advancements in identifying patients at high risk using data mining techniques;[8] this particular example shows the power of centralizing data.

A different approach is to use routine clinical data from around the world to transform data into knowledge. As a proof of concept, the euroCAT project created data stores at several cancer centres, which can be accessed using web technologies. The local data are mined, pseudoanonymized, translated, mapped to standard concepts and made available to trusted partners in the network. The trusted partners do not have direct access to the data, but they can send a machine learning algorithm to the different data stores to learn from the data without sharing them (*i.e.* knowledge sharing, not data sharing). To demonstrate the power of this technique, an existing model was improved by combining the data of five centres (www.eurocat.info). However, a lot of time and effort is still required to access and utilize all data generated in clinical routine and translate them into knowledge. The end result of distributed learning is prediction models which can support physicians when making patient-specific choices (www.predictcancer.org). A different successful data mining was started by Public Health England. Data from all linear accelerators in England were collected automatically using the Radiotherapy Dataset tools (http://www.ncin.org.uk/collecting_and_using_data/rtds). This data set was analyzed to examine the variation in given treatments for different regions of the country.

An important topic when discussing collection of healthcare data is patient privacy. All three approaches handle privacy issues differently. Registries are usually hosted in some central location by professional societies or government-related entities, which are often authorized to collect identifiable patient data and securely store them while giving researchers an anonymized view of these data. Clinical trials work with informed consent and with pseudoanonymized data. Distributed systems are privacy-by-design systems, as they simply do not allow data to leave the site where they were collected.

It should be noted that there is a perception among healthcare providers that data must be kept in isolation owing to privacy issues. The fact is that there are existing solutions for these issues. The real barrier to learning from (Big) data in healthcare is that it requires willingness, resources and expertise.[9]

Healthcare data are not yet "Big" enough to apply purely data-driven machine learning approaches and clinical expertise is needed to create useful models that make sense to the clinicians. Clinical trials, clinical registries and routine clinical data all provide unique evidence, which is currently utilized separately. Combining the three evidence sources into interoperable data stores makes them complementary to each other and will enable healthcare to move forward. However, data quality (Volume, Variety, Velocity and Veracity) and sharing issues are hindering progress. To achieve "Big Data" in healthcare, the data have to be Findable, Accessible, Interoperable and Reusable. The Findable, Accessible, Interoperable and Reusable Guiding Principles[10] can be applied to achieve good data management and stewardship, which will enable knowledge discovery and innovation. Eventually, when data-driven machine learning approaches have matured, it will provide a large knowledge base and clinical trials will only be used for a small subset of studies that requires to specific setup or a trial to prove (*i.e.* a new experimental treatment).

IBM have stated that there are numerous ways to improve healthcare using their technology and they provide a conclusion of utmost importance: "Information technology cannot drive change".[4] "Big Data" can be a powerful tool to move healthcare forward, but healthcare providers need to invest resources to make this happen. Consequently, industry leaders in radiation oncology are already exploring this horizon market in anticipation of the opportunities and challenges that "Big Data" in healthcare represents, both in terms of efficiency and efficacy. Our experience is that the technical limitations of sharing data are minimal. Practical reasons are that healthcare providers are not willing, do not have the resources and/or knowledge to share their data. Sharing data can have an effect on the reputation of the healthcare provider because it allows their performance to be compared with others. Furthermore, these data can be used in research that a competing institute is working on as well, possibly creating unwanted competitors. By limiting the access to data to a machine and only sharing the model learned from the data, these issues are eliminated or largely negated. Despite the conflicting interests of healthcare providers, change may be driven by pressure from external institutions (such as government and health insurance companies) to ensure that the highest standard and most cost-effective care is delivered to the patient.

Many world leaders throughout history have referenced the seventeenth century poem by John Donne—"No man is an island", when illustrating the need for collective responsibility and action towards a brighter future for all. This maxim rings as true in healthcare as it does in all other areas of life. We believe that utilizing patient privacy-preserving distributed machine

learning to translate and combine all data sources into knowledge will enable healthcare to move to individualized, high-quality, affordable and safe cancer treatments, ensuring the sustainability of healthcare. This will also allow moving further towards participative medicine with customized patient decision aids.

REFERENCES

1. Zarrouk M. Delivering excellence in patient care with ready access to clinical data. 2012 [Cited 28 July 2016]. Available from: http://www.netapp.com/us/media/wp-7169.pdf
2. Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, et al. A prospective study comparing the predictions of doctors *versus* models for treatment outcome of lung cancer patients: a step towards individualized care and shared decision making. *Radiother Oncol* 2014; **112**: 37–43. doi: https://doi.org/10.1016/j.radonc.2014.04.012
3. Benedict SH, Hoffman K, Martel MK, Abernethy AP, Asher AL, Capala J, et al. Overview of the American Society for Radiation Oncology-National Institutes of Health-American Association of Physicists in Medicine Workshop 2015: exploring opportunities for radiation oncology in the era of big data. *Int J Radiat Oncol Biol Phys* 2016; **95**: 873–9. doi: https://doi.org/10.1016/j.ijrobp.2016.03.006
4. Kohn MS, Sun J, Knoop S, Shabo A, Carmeli B, Sow D, et al. IBM's health analytics and clinical decision support. *Yearb Med Inform* 2014; **9**: 154–62. doi: https://doi.org/10.15265/IY-2014-0002
5. Bilimoria KY, Stewart AK, Winchester DP, Ko CY. The national cancer data base: a powerful initiative to improve cancer care in the United States. *Ann Surg Oncol* 2008; **15**: 683–90. doi: https://doi.org/10.1245/s10434-007-9747-3
6. Efstathiou JA, Nassif DS, McNutt TR, Bogardus CB, Bosch W, Carlin J, et al. Practice-based evidence to evidence-based practice: building the National Radiation Oncology Registry. *J Oncol Pract* 2013; **9**: e90–5. doi: https://doi.org/10.1200/JOP.2013.001003
7. Shah A, Stewart AK, Kolacevski A, Michels D, Miller R. Building a rapid learning health care system for oncology: why CancerLinQ collects identifiable health information to achieve its vision. *J Clin Oncol* 2016; **34**: 756–63. doi: https://doi.org/10.1200/JCO.2015.65.0598
8. Liao LJ, Chou HL, Lo WC, Wang CT, Chou HW, Chen CD, et al. Initial outcomes of an integrated outpatient-based screening program for oral cancers. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2015; **119**: 101–6. doi: https://doi.org/10.1016/j.oooo.2014.09.020
9. Sullivan R, Peppercorn J, Sikora K, Zalcberg J, Meropol NJ, Amir E, et al. Delivering affordable cancer care in high-income countries. *Lancet Oncol* 2011; **12**: 933–80. doi: https://doi.org/10.1016/S1470-2045(11)70141-3
10. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; **3**: 160018. doi: https://doi.org/10.1038/sdata.2016.18