Cancer Letters 382 (2016) 110–117

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Mini-review

Big Data and machine learning in radiation oncology: State of the art and future prospects

Jean-Emmanuel Bibault [a,b,*], Philippe Giraud [a], Anita Burgun [b,c]

[a] Radiation Oncology Department, Georges Pompidou European Hospital, Assistance Publique – Hôpitaux de Paris, Paris Descartes University, Paris Sorbonne Cité, Paris, France
[b] INSERM UMR 1138 Team 22: Information Sciences to support Personalized Medicine, Paris Descartes University, Sorbonne Paris Cité, Paris, France
[c] Biomedical Informatics and Public Health Department, Georges Pompidou European Hospital, Assistance Publique – Hôpitaux de Paris, Paris Descartes University, Paris Sorbonne Cité, Paris, France

ARTICLE INFO

Keywords:
Radiation oncology
Big Data
Predictive model
Machine learning

ABSTRACT

Precision medicine relies on an increasing amount of heterogeneous data. Advances in radiation oncology, through the use of CT Scan, dosimetry and imaging performed before each fraction, have generated a considerable flow of data that needs to be integrated. In the same time, Electronic Health Records now provide phenotypic profiles of large cohorts of patients that could be correlated to this information. In this review, we describe methods that could be used to create integrative predictive models in radiation oncology. Potential uses of machine learning methods such as support vector machine, artificial neural networks, and deep learning are also discussed.

Introduction

Level I evidence-based medicine relies on randomized controlled trials designed for large population of patients. But the increasing number of clinical and biological parameters that need to be explored to achieve precision medicine makes it almost impossible to design dedicated trials [1]. New approaches are needed for all subpopulations of patients. Clinicians need to use all the diagnostic tools (medical imaging, blood tests and genomics) in order to decide the appropriate combination of treatments (radiotherapy, chemotherapy, targeted therapy and immunotherapy). Each patient has an individual set of molecular abnormalities responsible for their disease or correlated with treatment response and clinical outcome. The concept of tailored treatments relies on identifying and leveraging these aberrations for each patient. This shift to molecular oncology has driven cancer research in the last 25 years and has allowed significant progress in poor-prognosis diseases such as non-small cell lung cancer (through the use of EGFR inhibitors [2]) or melanoma (through the use of immunotherapy [3]). But the burden of variant mutations can involve up to several hundred genes in a single tumor. Next-Generation Sequencing can be focused on specific regions, on whole-exome (all coding genes are sequenced) or whole-genome (all DNAs are sequenced). The same approach can be used to study the transcriptome. In any case, exploring as many genes as possible will be mandatory as we unravel the complexity of the molecular circuits involved in primary or secondary treatment resistance or radiation response [4]. The intricacy involved makes it almost certainly impossible to create specific trials for each and every case. It is traditionally considered that our cognitive capacity can integrate up to five factors in order to take a decision. By 2020, a decision will rely on up to 10,000 parameters for a single patient [5].

As sequencing costs have significantly decreased [6–8] and computing power has steadily increased (Fig. 1), the only factor preventing us from discovering factors influencing the disease's outcome is the lack of large phenotyped cohorts. The generalization of Electronic Health Records (EHR) gives us a unique opportunity to create adequate phenotypes. Data science has an obvious role in the generation of models that could be created from large databases to predict outcome and guide treatments.

Moreover, the similarity between clinical research patients and routine care patients regarding comorbidities, severity, time before initiation of treatment and tumor characteristics has been questioned [9]. A new paradigm of data driven methodologies reusing routine healthcare data to provide decision support is emerging. To quote I.S. Kohane, "Clinical decision support algorithms will be derived entirely from data … The huge amount of data available will make it possible to draw inferences from observations that will not be encumbered by unknown confounding" [10].

Integrating such a large and heterogeneous amount of data is in itself a challenge that must be overcome before we can actually create accurate models. The objective of this review is to explain the main informatics challenges in the implementation of a precision medicine program in radiation oncology and describe

* Corresponding author. Tel.: +33 156093403; fax: +33 156093506.
E-mail address: jean-emmanuel.bibault@aphp.fr (J.-E. Bibault).

http://dx.doi.org/10.1016/j.canlet.2016.05.033
0304-3835/

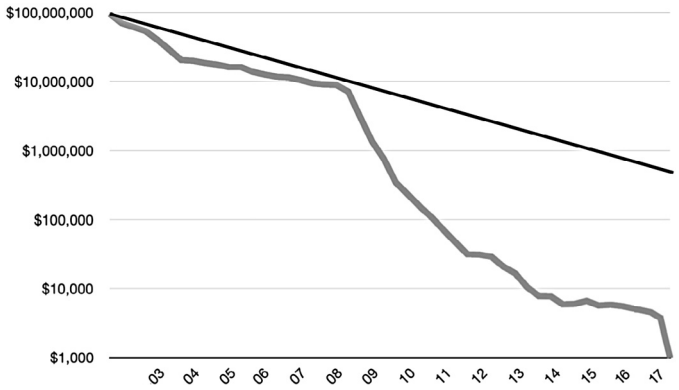

Fig. 1. Whole Genome Sequencing (actual cost, gray line) and computer power (Moore law, black line) costs.

approaches to address these challenges. We will discuss the methods available to create models predicting the outcome after radiotherapy or chemoradiation.

Which data should be considered and how should they be managed?

Lambin et al. have described in details the features that should be considered and integrated into a predicting model [11]. They include:

- Clinical features (patient performance status, grade and stage of the tumor, blood test results, patient questionnaires).
- Treatment features: planned spatial and temporal dose distribution, associated chemotherapy. For this, data could be extracted directly from the record-and-verify software to be analyzed.
- Imaging features: tumor size and volume, metabolic uptake (more globally included into the study field of "radiomics").
- Molecular features: intrinsic radiosensitivity [12], hypoxia [13], proliferation and normal tissue reaction [14]. In that part, genomic studies play a key role to determine these characteristics.

Data collection and management

State of the art radiation oncology provides a clear digital representation of the treatment performed. For each patient, we record the radiation regimen that has actually been performed. For each patient and treatment session, we know very well where photons go in the body and, by definition, we already have it in a digital format for every patient. Daily variability is also taken into account by onboard imaging, so we know where the dose is actually delivered. These systems can give the temporal and spatial distribution of the treatments performed. Data are prospectively collected for every patient in the record-and-verify software in each department. This highly digital nature lends itself to quantifying and analyzing the care delivery process. The quality of data gathered is far better than in most other fields of medicine. Extracting these data to integrate it in clinical data warehouse (CDW) in hospitals can be performed at different levels. Raw data provide detailed information on dose volume histograms, treatment volumes, time between each fraction, overall treatment time, dose rate, and images produced by onboard systems. Another approach that would consist of extracting only the data that are considered relevant before integrating it into the CDW would greatly decrease the richness of information and should be avoided [15].

Beyond the data described earlier, follow-up is very important in radiation oncology and medicine in general in order to detect toxicity. In that regard, online and mobile, but also wearable device inputs should be encouraged. Patients would then be able to provide detailed, real-time information on adverse events during and after the treatment without having to wait for their next appointment with the radiation oncologists. Several studies in that field have already shown the interest of patient reported outcomes to improve follow-up [16,17].

Table 1
Data types and approximate sizes for a single patient.

Data type	Format	Approx. size
Clinical features	Text	10 MB
Blood tests	Numbers	1 MB
Administrative	ICD-10 codes	1 MB
Imaging data	DICOM	450 MB
Radiation oncology data (planning and on-board imaging)	DICOM, RT-DICOM	500 MB
Raw genomic data	BAM: Position, base, quality	6 GB
Total		**7.9 GB**

The volume of data that need to be collected and managed is rapidly growing. Today, we can estimate that data for a single patient would amount to 7 GB, including the raw genomic data that would account for roughly 70% of it (Table 1). Health data security and accessibility is a major challenge for any institution. They should be accessible with ease and velocity from anywhere, without compromising their safety. Remote access to the data requires that the architecture takes into account high security constraints, including a strong user authentication and methods that guarantee traceability of all data processing steps. Relevant healthcare professionals' login procedures require scalable process with a significant cost, but they should certainly not be overlooked [18]. Medical record linkage and data anonymization are very often necessary steps to provide data for research. They often require a trustworthy third party that takes care of these procedures. In general, to provide healthcare data for research, the data must be moved from the care zone, where data are under the control of the trusted relationship between physician and patient, to the none-care zone, where data are under the control of special data governance bodies, to be anonymized and made available for analysis.

Existing solutions to support the storing and access of care include translational research platforms. These platforms are able to integrate large data sets of clinical information with omics data [19]. Despite technological advances, some authors believe the increases in data volume could be outstripping the hospitals' ability to cope with the demand for data storage [20]. One solution would consist of managing these data as most hospitals manage old medical files, i.e. moving the oldest and biggest files to external storage. For digital data, in order to maintain fast and easy access, we would need to move the most voluminous data to a secondary storage-optimized platform, separate from the query platform. Fig. 2 shows a proposal for a system integrating data from the hospital and data directly provided by the patients.

Use of ontologies for quality data extraction

Standardization in the fields and terms used in the EHR, treatment procedures, and genomic annotations increase the quality and comparability of the data used to create models. Diversity in these features results in an almost impossible challenge to extract and aggregate quality data. An ontology, i.e. a set of common concepts, is a key component of any data collecting system and predictive models. There are currently around 440 biomedical ontologies. The most commonly used include SNOMED [21], the NCI Thesaurus [22], CTC AE [23] and the UMLS meta-thesaurus [24].