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/

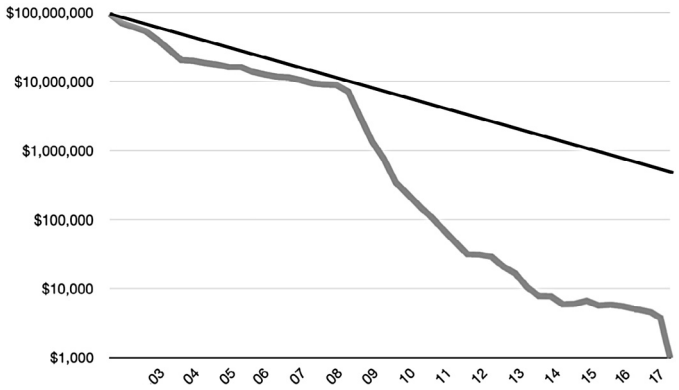

Fig. 1. Whole Genome Sequencing (actual cost, gray line) and computer power (Moore law, black line) costs.

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].

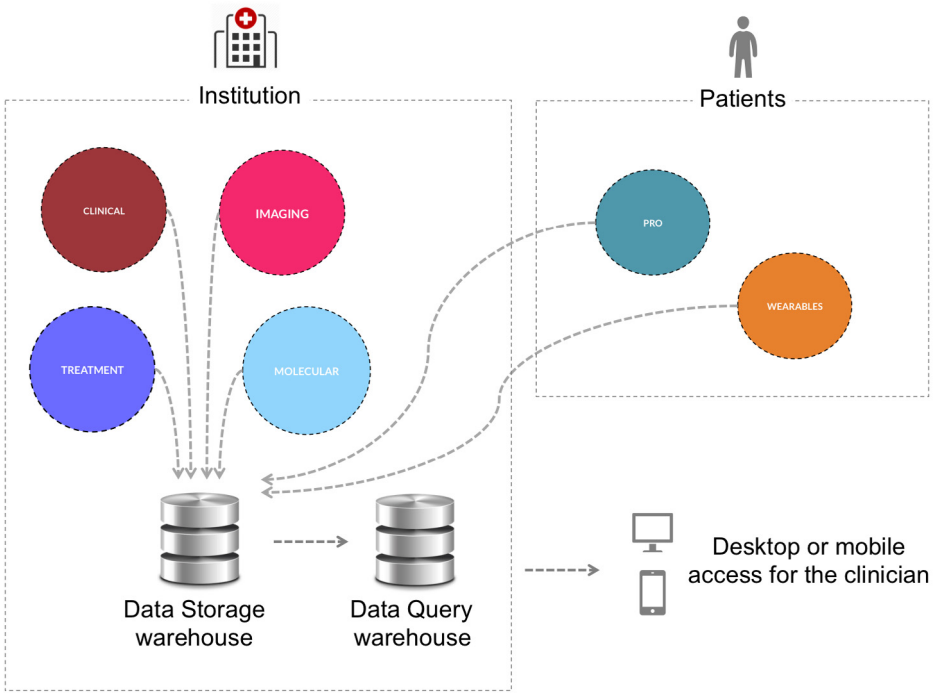

Fig. 2. Data collection and management system (PRO = patient reported outcomes).

These ontologies do not include many radiation oncology terms, which led to the creation of the Radiation Oncology Ontology (ROO) [25], which reused other ontologies but adds RO words such as Region Of Interest (ROI), Target Volumes (GTV, CTV, PTV), and Dose-Volume Histograms (DVH). Universal use of common ontologies will allow automatic multicentric data extraction and integration.

Data set quality and careful feature selection are very important. Independent verification by a second curator or data checker should be used when possible. Further verification by a knowledgeable expert is also very valuable, which means that the collaboration between the physician and the data scientist conducting the experiment is mandatory.

How to create a predictive model

Predictive modeling is a two-step process involving qualification followed by validation. Qualification will consist of demonstrating that the data are indicative of an outcome. Once predictive or prognostic factors have been identified, they should be validated on a different dataset. Once a model has been qualified and validated, further studies must be conducted in order to assess whether treatment decisions relying on the model actually improve the outcome of patients.

Kang et al. have proposed seven principles of modeling [15] in radiation oncology:

1. Consider both dosimetric and nondosimetric predictors
2. Manually curate predictors before automated analysis
3. Select a method for automated predictor selection
4. Consider how predictor multicollinearity is affecting the model
5. Correctly use cross-validation to improve prediction performance and generalization to external data to provide model generalizability with external data sets when possible
6. Assess multiple models and compare results with established models

These principles can be expanded to the whole field of medicine and should be carefully considered before creating a predictive model and validating this model. In order to create a model, we can rely on traditional statistical methods or machine learning methods. We will mostly focus on machine learning methods applied to radiation oncology.

Traditional statistical methods

For models predicting qualitative outcomes (such as toxicities), logistic regression should be considered, while Cox regression is traditionally used for survival-type data.

Logistic regression (LR) maps a combination of predictors to a probability of an outcome on an S-shaped curve (sigmoidal logistic function). LR should be used when exploring few, unrelated predictors (age, sex, tumor size). For example, in lung SBRT, it can be used to determine the optimal radiation dose that would probably achieve local control (one-dimension data, Fig. 3A) or even add-in GTV size (two-dimensions, Fig. 3B) as a predictor. Each feature comes into the model linearly and additively. A decision boundary can be created that is one dimension lower than the number of predictors explored (one dimension line for two predictors, two for three predictors, etc). LR has been used to predict esophagitis and xerostomia after lung or head and neck radiotherapy in several studies [26–28].

Machine learning (ML) methods

Several ML algorithms have been used in oncology:

- Decision Trees (DT) [29] where a simple algorithm creates mutually exclusive classes by answering questions in a predefined order,
- Naïve Bayes (NB) classifiers [30,31], which output probabilistic dependencies among variables,

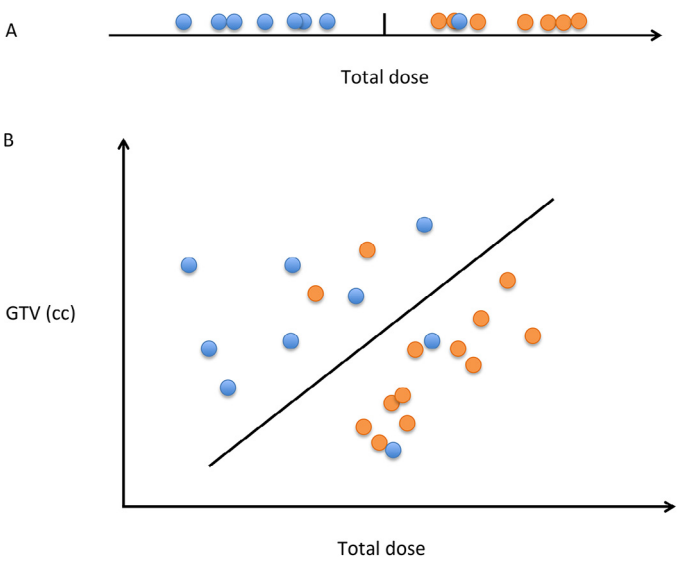

Fig. 3. Logistic Regression can create a linear threshold for one-dimension (A, total dose) or two-dimension (B, total and GTV size) data. Blue dots are local control failure and orange are local control success. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- k-nearest Neighbors (*k*-NN) [32], where a feature is classified according to its closest neighbor in the dataset, are used for classification and regression,
- Support Vector Machine (SVM) [33], where a trained model will classify new data into categories,
- Artificial Neural Network (ANN) [34], where models inspired by biological neural networks are used to approximate functions,
- Deep Learning (DL) [35], a variant of ANNs, where multiple layers of neurons are used.

Each of these methods has advantages and limitations, with different computation power requirements (Table 2). These should be used to choose the relevant method for any data analysis project. We will detail the two methods that were used in radiation oncology studies: SVM, ANN and its variant, DL.

Support vector machine

As mentioned above, LR defines a linear threshold for a limited number of features. If the model needs to integrate a higher number of variables that cannot be separated linearly, SVM can be used to find complex patterns. Similarity functions (or kernels) are chosen to perform a transformation of the data and choose data points or "support vectors". Patients with a combination of vectors are used to compare new patients and predict their outcome (Fig. 4). SVMs have been used in several studies to predict radiation pneumonitis after conformal radiotherapy [36], local control after lung SBRT [37] and chemoradiosensitivity in esophageal cancer [38]. In these studies, the authors classified the input parameters as dose (DHVs, EUD, BED) or non-dose features (clinical or biological features). It should be noted that the exact number and nature of features used is not always provided, which might limit the impact and applicability of the results.

Artificial neural network

In Artificial neural network, several layers of neurons are set up. Each "neuron" has a weight that determines its importance. Each layer receives data from the previous layer, calculates a score and passes the output to the next layer (Fig. 5). Using an ANN requires weighting neurons and connections correctly. A method to achieve this is to assign random weights to neurons and iteratively calculate and adjust these weights to progressively improve the correlation. ANNs have been used to predict survival in advanced head and neck cancers treated with irradiation with or without chemotherapy [39]. A three-layer feed-forward neural network integrating fourteen clinical parameters was trained through a thousand iterations. Bryce et al. showed that ANN was more reliable than LR and used more predictive variables. Six years later, Gulliford et al. used ANN to predict biological outcome and toxicity after radiotherapy for prostate cancer [40]. They used dosimetric parameters (DVH) and three separate ANNs on nocturia, rectal bleeding and PSA measurement. They showed that ANNs were able to predict biochemical control and specific bladder and rectum complications with sensitivity and specificity above 55%. Other studies performed on larger datasets improved sensitivity and specificity [41,42].

In lung radiotherapy, ANNs have also been used to predict pneumonitis [43,44]. In the study by Chen et al., six input features were selected: lung volume receiving >16 Gy (V16), generalized equivalent uniform dose (gEUD) for the exponent a = 1 (mean lung dose), gEUD for the exponent a = 3.5, free expiratory volume in 1 s (FEV1), diffusion capacity of carbon monoxide (DLCO%), and whether or not the patient underwent chemotherapy prior to radiotherapy. All features were then removed from the model to assess their relevance.

Table 2
Benefits and limitations of different machine learning algorithms.

Algorithm	Advantages	Limitations
Decision Tree	• Easy to understand • Fast	• Classes must be mutually exclusive • Results depend on the order of attribute selection • Risk of overly complex decision trees
Naïve Bayesian	• Easy to understand • Fast • No effect of order on training	• Variables must be statistically independent • Numeric attributes must follow a normal distribution • Classes must be mutually exclusive • Less accurate
k-nearest Neighbors	• Fast and simple • Tolerant of noise and missing values in data • Can be used for non-linear classification • Can be used for both regression and classification	• Variables with similar attributes will be sorted in the same class • All attributes are equally relevant • Requires considerable computer power as number of variables increases
Support Vector Machine	• Robust model • Limits the risk of error • Can be used to model non-linear relations	• Slow training • Risk of overfitting • Output model is difficult to understand
Artificial Neural Network and Deep Learning	• Tolerant of noise and missing values in data • Can be used for classification or regression • Can be easily updated with new data	• Output model is difficult to understand (« black-box ») • Risk of overfitting • Requires a lot of computer power • Requires experimentation to find the optimal network structure

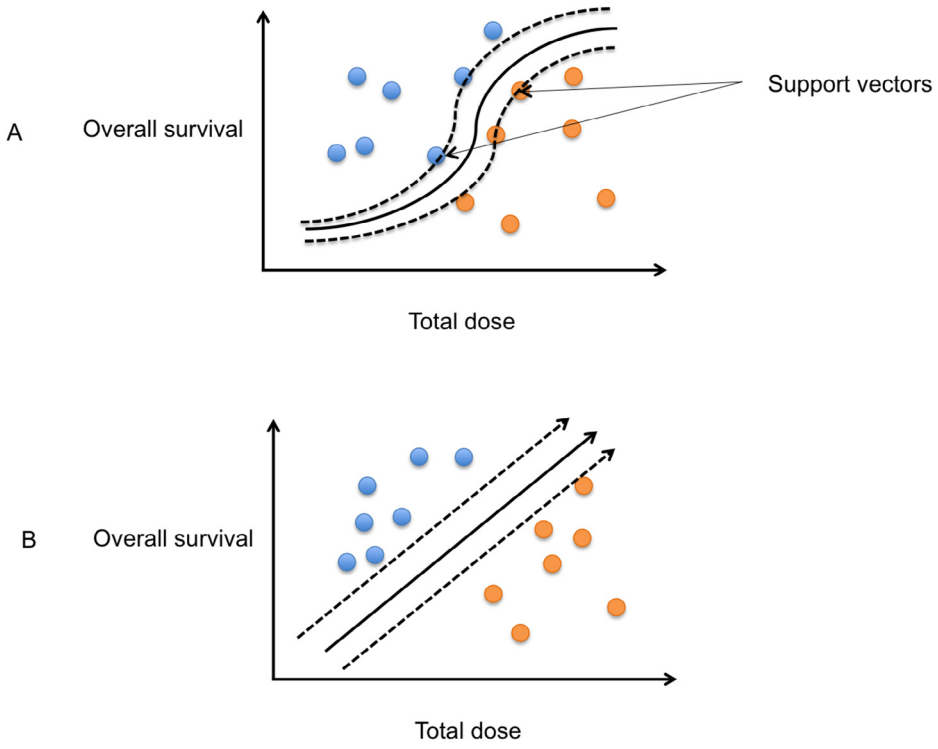

Fig. 4. Support Vector Machine transforms data points (A) with the help of support vectors in order to classify patients (B).

All of them except FEV1 and whether or not the patient underwent chemotherapy prior to radiotherapy were required for optimal prediction. In another study, ANNs have been used to predict survival in uterine cervical cancer treated with irradiation [45]. In that study, the predictive model used only seven parameters (age, performance status, hemoglobin, total protein, FIGO stage, histological and grading of radiation effect determined by periodic biopsy examination).

Deep learning

Deep learning is a variant of ANN. While ANN commonly features one or two hidden layers and is considered as supervised machine learning, DL differentiates itself with a higher number of hidden layers and is able to perform supervised or unsupervised learning. While DL is gaining interest in medical imaging [46,47] for classification or segmentation, it has not been used to predict the outcome after radiotherapy yet.

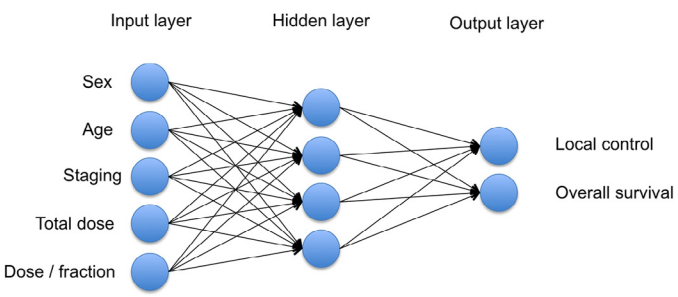

Fig. 5. Artificial Neural Network – Each neuron belongs to a layer and has a weight. Data are passed from layer to layer from the input (factors explored) to the output layer (outcome).

Difference between supervised learning (SL) and unsupervised learning (UL)

The goal of supervised learning is to predict a known output. It is commonly used to recognize images of objects or types of documents. A supervised algorithm will analyze a training dataset (where each example is a pair including an input feature and the desired output value) in order to create a function that best matches these training examples. The machine will generalize this function to pairs with unknown output value to predict them. In UL, the data provided are unlabeled and the algorithm will try to find natural patterns or groups within the data. In medicine, this will consist of characterizing each patient with vectors with values given to clinical features. Higher level features can be detected that would not have been seen as potential predictive or prognostic factors by a human intervention. UL could also identify new physiopathology by highlighting new groups of patients. With the increasing role of Graphics Processors Units (GPU) [48], UL could see more interest from researcher. If unsupervised machine learning can find salient correlations and connections between data points that no human would have thought to look for, a significant drawback could be that it does not necessarily provide any insight into what those correlations and connections mean. This could lead in some cases to a highly significant correlation found by UL that no human can understand. To this day, almost all machine learning algorithm studies within the context of predictive oncology are using supervised learning [49,50].

Discussion

Why should we use Big Data in radiation oncology?

The fours Vs of Big Data are Volume, Variety, Velocity and Veracity [51]. A comprehensive Electronic Health Record for any cancer patient will be around 8 GB, with genomic data being much larger

than all other data combined (volume). Creating a predictive model in radiation oncology requires a significant variety and heterogeneity of the data types that need to be included (variety). This represents in itself a significant challenge. In line with this, if we develop a decision support system based on such models, we will need fast data processing to support decision-making (velocity). Finally, data quality is exceptionally high in radiation oncology, as all departments use record-and-verify systems that prospectively store all information regarding the treatment prescribed, how the treatments have actually been performed and the potential deviations (veracity). All this contributes to believe that exploiting big data in radiation oncology is truly a match made in heaven.

Artificial intelligence and machine learning for cancer research

Studies using machine learning have not all been conducted with rigorous experimental designs. Data size is not always large enough to be partitioned into a training and a test set, let alone validate the algorithm on another dataset. Testing a model will need initially taking out 10–20% of the data for later internal validation [49]. In order to remove, or at least minimize, any bias coming from the data used, an external validation using a different (large enough) dataset will be required. The patient-per-feature ratio is also very challenging when you want to integrate thousands of information (especially genomics) into a model, because it is considered that this ratio should be at least 5 to 10 [52]. A small ratio will result in overtraining (or overfitting), i.e. creating a model that describes random errors or noise, specific to the dataset used to train the ANN that is not reliable on another population. Multiple predictor models based on different machine learning techniques should be used to assess the performance of the model. Ideally, a new model should outperform older classifications. It is estimated that only about 17% of the published ML studies in oncology tested more than one ML method [49].

ANN is the most used method in oncology predictive modeling, but Deep Learning (DL) is gaining interest in many fields [35]. With the release of open source software library, such as Google's TensorFlow [53], we could see more DL studies in the coming years.

A vision of the future: the learning health system

The task of creating and validating a truly integrative model in radiation oncology to guide treatment will require multicentric sharing of data and scientists. However, these models and the methodology used to create them can be used in all tumor localization. They will underpin decision support system that will use big data in every RO department in 10–15 years. These systems will need to be updated almost in real-time with dynamic programming and reinforcement learning techniques. They will guide decisions at the time of initial consultation for the best treatment options according to the patient's feature and state of knowledge. Optimal dose distribution, treatment time, associated chemotherapy, targeted therapy or immunotherapy will be chosen not by the physician, but by an algorithm. Private initiative, such as IBM's Watson, is already used in some Institutions, such as the Memorial Sloan Kettering Cancer Center in New York [54,55]. The same system could also guide decisions treatment for adverse event management and after the treatments for follow-up and early detection of any relapse. This "learning health system" will certainly be a game-changer in oncology, if it can actually be achieved. Follow-up will have to integrate all the data collected by wearable devices and connected objects that are being adopted by a large proportion of the population [56,57]. Continuous, real-time monitoring of abnormal events will lead to earlier detection of relapse, optimization of salvage treatment's efficiency and cost. Eventually, overall survival will be impacted by such approaches [16].

Implications for clinical research

Precision medicine has given birth to new clinical trial designs. For example, the SHIVA clinical trial compares targeted therapy based on tumor molecular profiling versus conventional therapy in patients with refractory cancer. Similarly, personalized radiation treatment based on data driven algorithms could be compared to conventional radiation therapy [58].

Big data in radiation oncology means studying large cohorts of patients and integrating heterogeneous types of data. Using these types of data through unsupervised machine learning holds great promises for identifying patterns beyond human comprehension. Oncology is already moving away from therapies based on anatomical and histological features and focusing on molecular abnormalities that define new groups of patients and diseases. This evolution induces an increasingly complex and changing base of knowledge that ultimately will not be usable by physicians. The other consequence of this is that, as we individualize molecular traits, designing clinical trials will become more and more difficult to the point where it will become statistically impossible to achieve sufficient power. The financial and methodological burdens of designing these clinical trials will eventually become unsustainable. Electronic Health Record (EHR) use in most institutions is an elegant and easy way to digitally capture large amount of data on patient characteristics, treatment features, adverse events and follow-up. The wealth of information should be used to generate new knowledge. The quality and nature of the data captured are important, as poor data will generate poor results ("garbage in, garbage out") and Big Data should not be seen as a magical box able to answer any question with ease and trust. Clinical trials are designed to avoid confounding factors and gather detailed data, not always available in EHR [59].

Several SEER studies have generated fast results on important questions [60–64]. However, when studying radiation treatments, a major limitation of big data is the lack of detailed information on treatment characteristics. Integrating these features straight out of the record and verify systems will provide faithful dosimetric and temporal data.

Several teams have already published studies using prediction to better adapt radiation treatments [65–70]. However, none of these approaches have reached clinical daily use. A simple, easy-to-use system would need to be directly implemented into the treatment planning system to provide decision support. The best achievable treatment plan based on a patient's medical history and anatomy would be given to the dosimetrist or physicist. The same system would be used to monitor patients during treatment and notify physicians whenever an adverse event outside of the predicted norm would happen. The data generated by each patient and treatment would be integrated into the model. We are however very far from this vision and in order to achieve it several methodological challenges will need to be addressed (capture core RO data into EHR, integrate clinical, dosimetric and biologic data into a single model, validate this model on a prospective cohort of patients).

Conclusion

A significant trend of big data analytics and machine learning will hopefully create high quality evidence in radiation oncology. Decrease in computer power cost, generalization of EHR, and advances in machine learning algorithms (mostly ANNs and DL) will drive innovations in this field. The improvement of the performance of predictive models will result in their use in learning health system that will help in personalizing radiation treatments with safety and efficiency. The physician must take ownership of these algorithms in order to remain at the center of healthcare.

Search strategy and selection criteria

Information for this Review was compiled by searching the PubMed and MEDLINE databases for articles published between January 1980 and April 2016, including early release publications. Search terms included "radiation therapy," "bioinformatics," "big data", "genomics", "electronic health records", "decision support systems" and "machine learning". Only articles published in English were considered and references were chosen based on suitability for inclusion. Full articles were obtained and the reference lists were checked for additional material, when appropriate.

Authors' contributions

J.E.B. and A.B. researched data for article; J.E.B., P.G. and A.B. contributed substantially to discussion of content; J.E.B. wrote the article; and P.G. and A.B. reviewed/edited the manuscript before submission.

Conflict of interest

The authors declare no competing interests.

References

[1] C. Chen, M. He, Y. Zhu, L. Shi, X. Wang, Five critical elements to ensure the precision medicine, Cancer Metastasis Rev. 34 (2015) 313–318, doi:10.1007/s10555-015-9555-3.

[2] J.G. Paez, P.A. Jänne, J.C. Lee, S. Tracy, H. Greulich, S. Gabriel, et al., EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy, Science 304 (2004) 1497–1500, doi:10.1126/science.1099314.

[3] F.S. Hodi, S.J. O'Day, D.F. McDermott, R.W. Weber, J.A. Sosman, J.B. Haanen, et al., Improved survival with ipilimumab in patients with metastatic melanoma, N. Engl. J. Med. 363 (2010) 711–723, doi:10.1056/NEJMoa1003466.

[4] A.G. Georgakilas, A. Pavlopoulou, M. Louka, Z. Nikitaki, C.E. Vorgias, P.G. Bagos, et al., Emerging molecular networks common in ionizing radiation, immune and inflammatory responses by employing bioinformatics approaches, Cancer Lett. 368 (2015) 164–172, doi:10.1016/j.canlet.2015.03.021.

[5] A.P. Abernethy, L.M. Etheredge, P.A. Ganz, P. Wallace, R.R. German, C. Neti, et al., Rapid-learning system for cancer care, J. Clin. Oncol. 28 (2010) 4268–4274, doi:10.1200/JCO.2010.28.5478.

[6] E.R. Mardis, A decade's perspective on DNA sequencing technology, Nature 470 (2011) 198–203, doi:10.1038/nature09796.

[7] M.L. Metzker, Sequencing technologies – the next generation, Nat. Rev. Genet. 11 (2010) 31–46, doi:10.1038/nrg2626.

[8] DNA sequencing costs. n.d. <http://www.genome.gov/sequencingcosts/> (accessed 12.03.16).

[9] N. Geifman, A.J. Butte, Do cancer clinical trial populations truly represent cancer patients? A comparison of open clinical trials to the cancer genome atlas, Pac. Symp. Biocomput. 21 (2016) 309–320.

[10] I.S. Kohane, J.M. Drazen, E.W. Campion, A glimpse of the next 100 years in medicine, N. Engl. J. Med. 367 (2012) 2538–2539, doi:10.1056/NEJMe1213371.

[11] P. Lambin, R.G.P.M. van Stiphout, M.H.W. Starmans, E. Rios-Velazquez, G. Nalbantov, H.J.W.L. Aerts, et al., Predicting outcomes in radiation oncology–multifactorial decision support systems, Nat. Rev. Clin. Oncol. 10 (2013) 27–40, doi:10.1038/nrclinonc.2012.196.

[12] J.-E. Bibault, I. Fumagalli, C. Ferté, C. Chargari, J.-C. Soria, E. Deutsch, Personalized radiation therapy and biomarker-driven treatment strategies: a systematic review, Cancer Metastasis Rev. 32 (2013) 479–492, doi:10.1007/s10555-013-9419-7.

[13] Q.-T. Le, D. Courter, Clinical biomarkers for hypoxia targeting, Cancer Metastasis Rev. 27 (2008) 351–362, doi:10.1007/s10555-008-9144-9.

[14] P. Okunieff, Y. Chen, D.J. Maguire, A.K. Huser, Molecular markers of radiation-related normal tissue toxicity, Cancer Metastasis Rev. 27 (2008) 363–374, doi:10.1007/s10555-008-9138-7.

[15] J. Kang, R. Schwartz, J. Flickinger, S. Beriwal, Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective, Int. J. Radiat. Oncol. Biol. Phys. 93 (2015) 1127–1135, doi:10.1016/j.ijrobp.2015.07.2286.

[16] F. Denis, S. Yossi, A.-L. Septans, A. Charron, E. Voog, O. Dupuis, et al., Improving survival in patients treated for a lung cancer using self-evaluated symptoms reported through a web application, Am. J. Clin. Oncol. (2015) doi:10.1097/COC.0000000000000189.

[17] A.D. Falchook, G. Tracton, L. Stravers, M.E. Fleming, A.C. Snavely, J.F. Noe, et al., Use of mobile device technology to continuously collect patient-reported symptoms during radiotherapy for head and neck cancer: a prospective feasibility study, Adv, Radiat. Oncol. (2016) doi:10.1016/j.adro.2016.02.001.

[18] M. Li, S. Yu, K. Ren, W. Lou, Securing personal health records in cloud computing: patient-centric and fine-grained data access control in multi-owner settings, in: S. Jajodia, J. Zhou (Eds.), Security and Privacy in Communication Networks, Springer Berlin, Heidelberg, 2010, pp. 89–106. <http://link.springer.com/chapter/10.1007/978-3-642-16161-2_6> (accessed 21.05.16).

[19] V. Canuel, B. Rance, P. Avillach, P. Degoulet, A. Burgun, Translational research platforms integrating clinical and omics data: a review of publicly available solutions, Brief. Bioinform. 16 (2015) 280–290, doi:10.1093/bib/bbu006.

[20] V. Huser, J.J. Cimino, Impending challenges for the use of Big Data, Int. J. Radiat. Oncol. Biol. Phys. (2015) doi:10.1016/j.ijrobp.2015.10.060.

[21] Systematized nomenclature of medicine – clinical terms – summary | NCBO BioPortal. n.d. <https://bioportal.bioontology.org/ontologies/SNOMEDCT> (accessed 07.03.16).

[22] National cancer institute thesaurus – summary | NCBO BioPortal. n.d. <https://bioportal.bioontology.org/ontologies/NCIT> (accessed 07.03.16).

[23] Common terminology criteria for adverse events – summary | NCBO BioPortal. n.d. <https://bioportal.bioontology.org/ontologies/CTCAE> (accessed 07.03.16).

[24] Fact SheetUMLS® Metathesaurus®. n.d. <https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html> (accessed 07.03.16).

[25] Radiation oncology ontology – summary | NCBO BioPortal. n.d. <http://bioportal.bioontology.org/ontologies/ROO> (accessed 07.03.16).

[26] I. El Naqa, J. Bradley, A.I. Blanco, P.E. Lindsay, M. Vicic, A. Hope, et al., Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors, Int. J. Radiat. Oncol. Biol. Phys. 64 (2006) 1275–1286, doi:10.1016/j.ijrobp.2005.11.022.

[27] T.-F. Lee, P.-J. Chao, H.-M. Ting, L. Chang, Y.-J. Huang, J.-M. Wu, et al., Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of Xerostomia after intensity-modulated radiotherapy for head and neck cancer, PLoS ONE 9 (2014) e89700, doi:10.1371/journal.pone.0089700.

[28] T.-F. Lee, M.-H. Liou, Y.-J. Huang, P.-J. Chao, H.-M. Ting, H.-Y. Lee, et al., LASSO NTCP predictors for the incidence of xerostomia in patients with head and neck squamous cell carcinoma and nasopharyngeal carcinoma, Sci. Rep. 4 (2014) 6217, doi:10.1038/srep06217.

[29] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.

[30] P. Langley, W. Iba, K. Thompson, An analysis of Bayesian classifiers, in: AAAI, 1992, pp. 223–228.

[31] P. Langley, S. Sage, Induction of selective Bayesian classifiers, in: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1994, pp. 399–406. <http://dl.acm.org/citation.cfm?id=2074445> (accessed 12.03.16).

[32] E.A. Patrick, F.P. Fischer III, A generalized k-nearest neighbor rule, Inf. Control 16 (1970) 128–152, doi:10.1016/S0019-9958(70)90081-1.

[33] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer Verlag, New York, 1982.

[34] D.E. Rumelhart, J. McClelland, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, Cambridge, 1986.

[35] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, doi:10.1038/nature14539.

[36] S. Chen, S. Zhou, F.-F. Yin, L.B. Marks, S.K. Das, Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis, Med. Phys. 34 (2007) 3808–3814.

[37] R.J. Klement, M. Allgäuer, S. Appold, K. Dieckmann, I. Ernst, U. Ganswindt, et al., Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer, Int. J. Radiat. Oncol. Biol. Phys. 88 (2014) 732–738, doi:10.1016/j.ijrobp.2013.11.216.

[38] Y. Hayashida, K. Honda, Y. Osaka, T. Hara, T. Umaki, A. Tsuchida, et al., Possible prediction of chemoradiosensitivity of esophageal cancer by serum protein profiling, Clin. Cancer Res. 11 (2005) 8042–8047, doi:10.1158/1078-0432.CCR-05-0656.

[39] T.J. Bryce, M.W. Dewhirst, C.E. Floyd, V. Hars, D.M. Brizel, Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck, Int. J. Radiat. Oncol. Biol. Phys. 41 (1998) 339–345.

[40] S.L. Gulliford, S. Webb, C.G. Rowbottom, D.W. Corne, D.P. Dearnaley, Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate, Radiother. Oncol. 71 (2004) 3–12, doi:10.1016/j.radonc.2003.03.001.

[41] A. Pella, R. Cambria, M. Riboldi, B.A. Jereczek-Fossa, C. Fodor, D. Zerini, et al., Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy, Med. Phys. 38 (2011) 2859–2867.

[42] S. Tomatis, T. Rancati, C. Fiorino, V. Vavassori, G. Fellin, E. Cagna, et al., Late rectal bleeding after 3D-CRT for prostate cancer: development of a neural-network-based predictive model, Phys. Med. Biol. 57 (2012) 1399–1412, doi:10.1088/0031-9155/57/5/1399.

[43] S. Chen, S. Zhou, J. Zhang, F.-F. Yin, L.B. Marks, S.K. Das, A neural network model to predict lung radiation-induced pneumonitis, Med. Phys. 34 (2007) 3420–3427.

[44] M. Su, M. Miften, C. Whiddon, X. Sun, K. Light, L. Marks, An artificial neural network for predicting the incidence of radiation pneumonitis, Med. Phys. 32 (2005) 318–325.

[45] T. Ochi, K. Murase, T. Fujii, M. Kawamura, J. Ikezoe, Survival prediction using artificial neural networks in patients with uterine cervical cancer treated by radiation therapy alone, Int. J. Clin. Oncol. 7 (2002) 294–300, doi:10.1007/s101470200043.

[46] K.-L. Hua, C.-H. Hsu, S.C. Hidayati, W.-H. Cheng, Y.-J. Chen, Computer-aided classification of lung nodules on computed tomography images via deep

learning technique, Onco Targets Ther. 8 (2015) 2015–2022, doi:10.2147/OTT.S80733.
[47] Y. Guo, Y. Gao, D. Shen, Deformable MR prostate segmentation via deep feature learning and sparse patch matching, IEEE Trans. Med. Imaging (2015) doi:10.1109/TMI.2015.2508280.
[48] R. Raina, A. Madhavan, A.Y. Ng, Large-scale deep unsupervised learning using graphics processors, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA, 2009, pp. 873–880, doi:10.1145/1553374.1553486.
[49] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Cancer Inform. 2 (2006) 59–77.
[50] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17, doi:10.1016/j.csbj.2014.11.005.
[51] C.U. Lehmann, B. Séroussi, M.-C. Jaulent, Big^3. Editorial, Yearb. Med. Inform. 9 (2014) 6–7, doi:10.15265/IY-2014-0030.
[52] R.L. Somorjai, B. Dolenko, R. Baumgartner, Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions, Bioinformatics 19 (2003) 1484–1491.
[53] TensorFlow – an open source software library for machine intelligence. n.d. <https://www.tensorflow.org/> (accessed 12.03.16).
[54] S. Parodi, G. Riccardi, N. Castagnino, L. Tortolina, M. Maffei, G. Zoppoli, et al., Systems medicine in oncology: signaling network modeling and new-generation decision-support systems, Methods Mol. Biol. 1386 (2016) 181–219, doi:10.1007/978-1-4939-3283-2_10.
[55] Memorial Sloan Kettering Cancer Center, Watson oncology. n.d. <https://www.mskcc.org/about/innovative-collaborations/watson-oncology> (accessed 10.03.16).
[56] The coming era of human phenotyping, Nat. Biotechnol. 33 (2015) 567, doi:10.1038/nbt.3266.
[57] N. Savage, Mobile data: made to measure, Nature 527 (2015) S12–S13, doi:10.1038/527S12a.
[58] N. Servant, J. Roméjon, P. Gestraud, P. La Rosa, G. Lucotte, S. Lair, et al., Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial, Front. Genet. 5 (2014) 152, doi:10.3389/fgene.2014.00152.
[59] R.C. Chen, P.E. Gabriel, B.D. Kavanagh, T.R. McNutt, How will big data impact clinical decision making and precision medicine in radiation therapy, Int. J. Radiat. Oncol. Biol. Phys. (2015) doi:10.1016/j.ijrobp.2015.10.052.
[60] A. Berrington de Gonzalez, R.E. Curtis, S.F. Kry, E. Gilbert, S. Lamart, C.D. Berg, et al., Proportion of second cancers attributable to radiotherapy treatment in adults: a cohort study in the US SEER cancer registries, Lancet Oncol. 12 (2011) 353–360, doi:10.1016/S1470-2045(11)70061-4.
[61] B.A. Virnig, J.L. Warren, G.S. Cooper, C.N. Klabunde, N. Schussler, J. Freeman, Studying radiation therapy using SEER-Medicare-linked data, Med. Care 40 (2002) IV–49–54, doi:10.1097/01.MLR.0000020940.90270.4D.
[62] S.C. Darby, P. McGale, C.W. Taylor, R. Peto, Long-term mortality from heart disease and lung cancer after radiotherapy for early breast cancer: prospective cohort study of about 300,000 women in US SEER cancer registries, Lancet Oncol. 6 (2005) 557–565, doi:10.1016/S1470-2045(05)70251-5.
[63] X. Du, J.L. Freeman, J.S. Goodwin, Information on radiation treatment in patients with breast cancer: the advantages of the linked medicare and SEER data. Surveillance, epidemiology and end results, J. Clin. Epidemiol. 52 (1999) 463–470.
[64] Y. Song, W. Wang, G. Tao, W. Zhu, X. Zhou, P. Pan, Survival benefit of radiotherapy to patients with small cell esophagus carcinoma – an analysis of Surveillance Epidemiology and End Results (SEER) data, Oncotarget (2015) doi:10.18632/oncotarget.6764.
[65] B. Wu, T. McNutt, M. Zahurak, P. Simari, D. Pang, R. Taylor, et al., Fully automated simultaneous integrated boosted-intensity modulated radiation therapy treatment planning is feasible for head-and-neck cancer: a prospective clinical study, Int. J. Radiat. Oncol. Biol. Phys. 84 (2012) e647–e653, doi:10.1016/j.ijrobp.2012.06.047.
[66] B. Wu, F. Ricchetti, G. Sanguineti, M. Kazhdan, P. Simari, R. Jacques, et al., Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning, Int. J. Radiat. Oncol. Biol. Phys. 79 (2011) 1241–1247, doi:10.1016/j.ijrobp.2010.05.026.
[67] S.F. Petit, B. Wu, M. Kazhdan, A. Dekker, P. Simari, R. Kumar, et al., Increased organ sparing using shape-based treatment plan optimization for intensity modulated radiation therapy of pancreatic adenocarcinoma, Radiother. Oncol. 102 (2012) 38–44, doi:10.1016/j.radonc.2011.05.025.
[68] L.M. Appenzoller, J.M. Michalski, W.L. Thorstad, S. Mutic, K.L. Moore, Predicting dose-volume histograms for organs-at-risk in IMRT planning, Med. Phys. 39 (2012) 7446–7461, doi:10.1118/1.4761864.
[69] X. Zhu, Y. Ge, T. Li, D. Thongphiew, F.-F. Yin, Q.J. Wu, A planning quality evaluation tool for prostate adaptive IMRT based on machine learning, Med. Phys. 38 (2011) 719–726.
[70] S.P. Robertson, H. Quon, A.P. Kiess, J.A. Moore, W. Yang, Z. Cheng, et al., A data-mining framework for large scale analysis of dose-outcome relationships in a database of irradiated head and neck cancer patients, Med. Phys. 42 (2015) 4329–4337, doi:10.1118/1.4922686.