Methods xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

# Perspectives on making big data analytics work for oncology

Issam El Naqa MS, MA, PhD, DABR

*University of Michigan, Department of Radiation Oncology, Ann Arbor, MI, United States*

ARTICLE INFO

*Article history:*
Received 1 May 2016
Received in revised form 19 August 2016
Accepted 25 August 2016
Available online xxxx

*Keywords:*
Big data
Oncology
Machine learning
Clinical decision support

ABSTRACT

Oncology, with its unique combination of clinical, physical, technological, and biological data provides an ideal case study for applying big data analytics to improve cancer treatment safety and outcomes. An oncology treatment course such as chemoradiotherapy can generate a large pool of information carrying the 5 Vs hallmarks of big data. This data is comprised of a heterogeneous mixture of patient demographics, radiation/chemo dosimetry, multimodality imaging features, and biological markers generated over a treatment period that can span few days to several weeks. Efforts using commercial and in-house tools are underway to facilitate data aggregation, ontology creation, sharing, visualization and varying analytics in a secure environment. However, open questions related to proper data structure representation and effective analytics tools to support oncology decision-making need to be addressed. It is recognized that oncology data constitutes a mix of structured (tabulated) and unstructured (electronic documents) that need to be processed to facilitate searching and subsequent knowledge discovery from relational or NoSQL databases. In this context, methods based on advanced analytics and image feature extraction for oncology applications will be discussed. On the other hand, the classical p (variables) $\gg$ n (samples) inference problem of statistical learning is challenged in the Big data realm and this is particularly true for oncology applications where p-omics is witnessing exponential growth while the number of cancer incidences has generally plateaued over the past 5-years leading to a quasi-linear growth in samples per patient. Within the Big data paradigm, this kind of phenomenon may yield undesirable effects such as echo chamber anomalies, Yule–Simpson reversal paradox, or misleading ghost analytics. In this work, we will present these effects as they pertain to oncology and engage small thinking methodologies to counter these effects ranging from incorporating prior knowledge, using information-theoretic techniques to modern ensemble machine learning approaches or combination of these. We will particularly discuss the pros and cons of different approaches to improve mining of big data in oncology.



## 1. Introduction

Cancer is a leading cause of death worldwide in both men and women, with about a yearly 14 million new cases and 8 million cancer related deaths (15% of all deaths) according to the latest global statistics [1]. It is a diverse family of diseases (about 100 known cancers) that is characterized by an abnormal growth of cells with acquired ability of uncontrolled progression and invasion of surrounding tissues disrupting their vital functions and leading to patient death. Biologically, cancer cells comprise of six biological capabilities that underlie their genetic characteristics: sustained proliferative signaling, evasion of growth suppressors, resistance to cell death, replicated immortality, induction of angiogenesis, and activation of invasion and metastasis [2].

Patients diagnosed with cancer have multiple treatment options to choose from depending on their disease type and its status. Typically, patient with localized and early stage disease would receive surgery, benefit from watchful waiting or active surveillance as in prostate cancer [3], or the use of medication for management purpose, rather than curative purpose, as in certain types of leukemia [4]. However, patients at more advanced stages of disease would receive chemoradiotherapy, molecular targeted therapy, immunotherapy, or a combination of these treatment modalities. In this work, we will primarily draw examples from radiation oncology; however, the ideas could be directly generalized to other treatment modalities.

Radiotherapy constitute of the use of high-energy irradiation to eradicate the tumor cells delivered from external linear accelerators (Linacs) or internally sealed radioisotopes (Brachytherapy). Radiotherapy data enjoys a unique combination of clinical patient demographics, physical use of high-energy irradiation, application of image-guidance (radiomics), and biological markers (genomics, proteomics, metabolomics), generated over a treatment period that can span few days to several weeks [5]. This data carries the

*E-mail address:* ielnaqa@med.umich.edu

http://dx.doi.org/10.1016/j.ymeth.2016.08.010
1046-2023/

five Vs (volume, velocity, variety, veracity, and value) hallmarks of big data as depicted in Fig. 1. Thus, it can provide an ideal case study for applying big data (also known as pan-Omics) analytics to improve its safe delivery and patient treatment outcomes. However, the classical statistical inference problem of large number of variables with relatively small number of samples is challenged in the context of oncology big pan-omics realm [6,7]. Moreover, it is noted that the selection of clinical or biological endpoints can also be defined in many different ways: disease-free, disease-specific and overall survival, local or loco-regional tumor control, etc. This endpoint selection would further influence the inter-relationship with pan-Omics variables differently adding to the problem complexity. In the following, we will discuss the effect of these challenges and highlight methods and techniques to overcome these impending challenges.

## 2. Material and methods

### 2.1. The big panomics of oncology

Oncology pan-Omics data could be divided based on their nature into four categories: Clinical, dosimetric, imaging, and biological, which are briefly described in the following [6].

#### 2.1.1. Clinical data

Clinical data in oncology and particularly in chemoradiotherapy typically refers to cancer diagnostic information (e.g., site, histology, stage, grade, etc.), patient-related characteristics (e.g., age, gender, co-morbidities, etc.), and physiological metrics (e.g., pulmonary function measurements, heart/pulse rates, blood cell counts, body mass index (BMI), etc.). Prior to the era of genetic profiling, these clinical variables were considered the only gold standard for clinical management and decision-making in oncology. From an informatics perspective, the mining of such data could be challenging particularly if the data is unstructured as typically the case, however, there are good opportunities for applying natural language processing (NLP) techniques to assist in the organization of such data [8].

#### 2.1.2. Dosimetric data

This type of data is related to the treatment planning process in radiotherapy or the chemical agent in chemotherapy. In radiotherapy, this data is composed of radiation dose delivery virtual simulation using computed tomography (CT) imaging; specifically, the number of beams, irradiation angles, energies, monitor units, and most importantly dose-volume metrics derived from dose-volume histograms (DVHs) graphs. Dose-volume metrics have been extensively studied in the radiation oncology literature for outcomes modeling [9–14]. These metrics are extracted from the DVH such as the volume receiving at least certain dose x (Vx), minimum dose to x% volume (Dx), mean, maximum and minimum dose, etc. [15]. Moreover, software dedicated tools such as 'DREES" have been used for deriving these metrics and modeling of radiotherapy response [16].

There are different categories of chemical agents that aim on eradicating tumor cancers[17]. Among the most common ones are alkylating agents, which substitute an alkyl groups (hydrocarbon) for hydrogen atom of organic compound including DNA (e.g., Temozolomide). There are also antibiotics (e.g., Doxorubicin, Blemoycin). Another common one are antimetabolites (e.g., Methotrexate, 5-Fluorouracil,Taxanes,vinca alkaloids). Another agents that do not fall into any of these classes include: Platinum compounds (Cisplatin) and topoisomerases DNA winding enzymes inhibitors). Recently, more signaling pathway targeted agents have been developed such anti-EGFR such as Cetuximab or Erbitux[18]. In addition to the agent type and dosage, the timing of the administration of the agent influence treatment response. Chemotherapy could be administrated after the completion of the local treatment such as radiation and is called adjuvant chemotherapy, before the local treatment and is called induction chemotherapy, or given during local treatment and is called concurrent chemotherapy. In particular, concurrent chemoradiation has been demonstrated to effective in the treatment of several cancers, in which the chemotherapy agent can act as a radiosensitizer by aiding the destruction of radiation resistant clones or act systematically and potentially eradicate distant metastases [19].

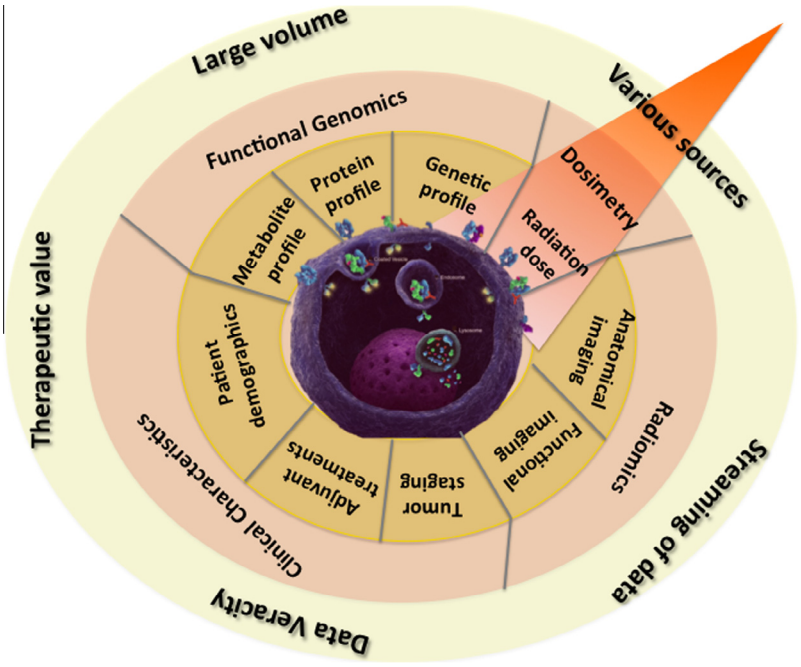



**Fig. 1.** The Big data (pan-Omics) of radiotherapy highlighting its data categories and its 5 Vs characteristics.

*2.1.3. Radiomics (Imaging features)*

Cancer patients are treated based on observational assessment from diagnostic imaging particularly computed tomography (CT) in combination with other clinical factors [20]. Information from multiple imaging modalities could be used to improve treatment monitoring and prognosis in different cancer types. For example, physiological information (tumor metabolism, proliferation, necrosis, hypoxic regions, etc.) can be collected directly from nuclear imaging modalities such as single-photon emission computed tomography (SPECT) and positron emission tomography (PET) or indirectly from magnetic resonance imaging (MRI) [21,22]. The complementary nature of these different imaging modalities has led to efforts toward combining information to achieve better treatment outcomes. For instance, PET/CT has been utilized for staging, planning, and assessment of response to chemoradiation therapy [23,24]. Similarly, MRI has been applied in tumor delineation and assessing toxicities in head and neck cancers [25,26]. Moreover, quantitative information from hybrid-imaging modalities could be related to biological and clinical endpoints, a new emerging field referred to as 'radiomics' [27,28]. Potential of this new field to monitor and predict response to chemoradiotherapy has been demonstrated in esophageal [29], head and neck [30,31], cervix [30,32], lung [33] [34], sarcoma [35] cancers, in turn allowing for adapting and individualizing treatment.

*2.1.4. Biological markers*

A biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention" [36]. Biomarkers can be categorized based on the biochemical source of the marker into exogenous or endogenous.

Exogenous biomarkers are based on introducing a foreign substance into the patient's body such as those used in molecular imaging as discussed above. Conversely, endogenous biomarkers can further be classified as (1) 'expression biomarkers,' measuring changes in gene expression or protein levels or (2) 'genetic biomarkers,' based on variations, for tumors or normal tissues, in the underlying DNA genetic code. Measurements are typically based on tissue or fluid specimens, which are analyzed using molecular biology laboratory techniques [37]. Aggregation of large-scale genetic biomarkers have been the subject of large national efforts such as The Cancer Genome Atlas (TCGA) Data Portal, which provides a very useful platform for researchers to analyze datasets generated by TCGA. It contains clinical information, genomic characterization data, and high-level sequence analysis of the tumor genomes in different cancer types [38–41].

*2.2. Database technologies for oncology*

Traditionally, relational database management systems (RDBMS) have been the technology of choice for storing and query oncology information. RDBMS are based on organizing the data relation schema in a tabular format (sets of rows (tuples) and columns (attributes)) in accordance with Codd's 12 rules [42]. SQL (structured query language) is a 4th generational programming language that is used to process the data in an RDBMS. RDBMS and SQL have been the driving technology for Electronic Health Record (EHR) management software including that of oncology. Several governmental, commercial, and open source resources for EHR exist. For instance, in the United States, more than 50% of patient records are stored in the Epic systems (Verona, WI), privately held software, which employs an object-oriented RDBMS. In addition, there are open source EHR systems, however, they did not receive traction.

Recently, there has been resurgence in NoSQL (not only SQL) database technologies. NoSQL allows for a blend of structured and unstructured data with no commitment to a schema unless needed and enjoys a remarkable horizontal scalability for aggregating and querying massive datasets. Interestingly, the VistA EHR system developed by department of Veterans affairs in the 1960s is based on the MUMPS (Massachusetts General Hospital Utility Multi-Programming System), which is a key-value NoSQL database system. Today, the NoSQL open source Hadoop architecture is considered the platform of choice for processing big data and potentially oncology data. The enabling technology that sprung its big data analytics potential is called MapReduce, which is a new parallel programming paradigm that involves two steps a *Map* function for filtering and sorting and a *Reduce* function for grouping and aggregation of data. However, an issue that may impact NoSQL adoption that in some instances the common so called ACID properties of (Atomicity, Consistency, Isolation, and Durability) of a reliable transactional processing may need to be compromised to achieve higher analytical performance. As a compromise this is created market for NewSQL that rely on storing large data in memory, which is advocated by M. Stonebraker (VoltDB, Inc., Bedford, MA).

*2.3. Pan- vs. p-OMICs*

Due to advances imaging and biotechnology radiotherapy data has witnessed tremendous exponential growth in the past decade, however, number of cancer incidences has generally plateaued as depicted in Fig. 2. The fact that $p$ (variables) $\gg n$ (samples) constitute a serious challenge for class inference methods of statistical learning. This $p$-omics phenomenon may yield undesirable effects such as spurious correlations, echo chamber anomalies, Yule–Simpson reversal paradox, or misleading ghost analytics as discussed in the following.

*2.3.1. Spurious relationship*

This pitfall commonly emerges in big data analysis when two variables have no true relationship, however, such one may be wrongly inferred due to confounding effects [43]. This is an important process when attempting to identify a biomarker of chemoradiation response, for instance. The famous example of such a case is the association of the number of ice cream sold and increased risk of drowning; the confounding effect or lurking explanatory variable is simply warm seasons. Understanding of the problem setup and possible prior knowledge of potential confounding effects is helpful in mitigating such effect.

*2.3.2. Echo chamber effect*

This happens when a relationship in the data is magnified by the data aggregation process itself in a cyclic manner [44]. This is typically a sampling problem (selection bias), in which the analyzed sample is not representative of the intended population. A common example is encountered in meta-analyses of previous oncology biomarker study findings, where negative results are typically less likely to be published [45].

*2.3.3. Yule–Simpson paradox*

This is reverse effect of the echo chamber, where a true association is found in small datasets but lost or even reversed when larger data is aggregated. This paradox was reported by Simpson in the analysis of contingency tables in interpreting second order interactions [46]. A lauded example in cancer research of this paradox is noted in the backlash generated by the paper by Tomasetti and Vogelstein that implied that variations in cancer risk are mainly explained by the number of stem cell divisions [47], a mere 'bad luck' issue irrespective of the environment. This has been

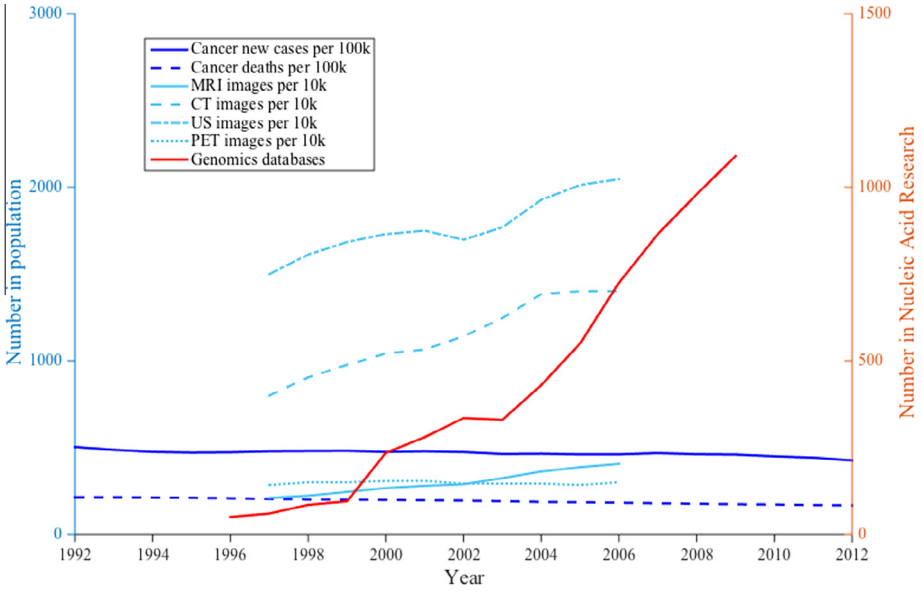



**Fig. 2.** The p-omics vs. pan-omics in cancer outcome modeling problem, where the number of variables growing rapidly in imaging and genetics while the number of samples have largely plateaued. The data is compiled querying cancer information from the SEER 2015 statistics, imaging information from Health Affairs 2008, and genomics from the Nucleic Acids Research archive.

debated vigorously demonstrating selection bias and Yule-Simpson effects in the performed data analysis [48]. A common pitfall that could results in this paradox can happen when not accounting for known patient characteristics such as age and gender when conducting population studies. For instance, it is known that there exists a negative relationship between medicine dosages and recovery in both males and females, however, when grouped together in a larger, a surprising positive relationship emerges as shown in Fig. 3 [49]. A possible remedy for this effect is using stratification by variables or more systematically performing unsupervised clustering to uncover such sub-population effects.

*2.3.4. Ghost analytics*

This refers to erroneous (mis-) using of statistical tests or learning algorithms when analyzing large datasets. For instance, this problem arises when not accounting for assumptions embedded in a statistical test before applying it. A classical example is encountered when conducting multiple comparisons and reporting a "significant" p-value of the null hypothesis testing yielding misleading results by not adjusting the level of Type 1 error. Interestingly, when statistician R. Fisher introduced the notion of p-values in the 1920s, he did not intend to have it as a definitive test rather simply as an informal way to judge whether association evidence was worthy of a second look. Therefore, it is necessary to understand the assumptions made in a statistical test before attempting to apply it in order to achieve meaningful results.

## 3. Modeling methods

Modeling techniques in oncology in general and radiation oncology in particular could be divided generally into bottom-up and top-down approaches as depicted in Fig. 4. The focus of this review will be on top-down approaches, while bottom-up methods are described for completeness and as a way to constrain the search space when conducting datamining exercises.

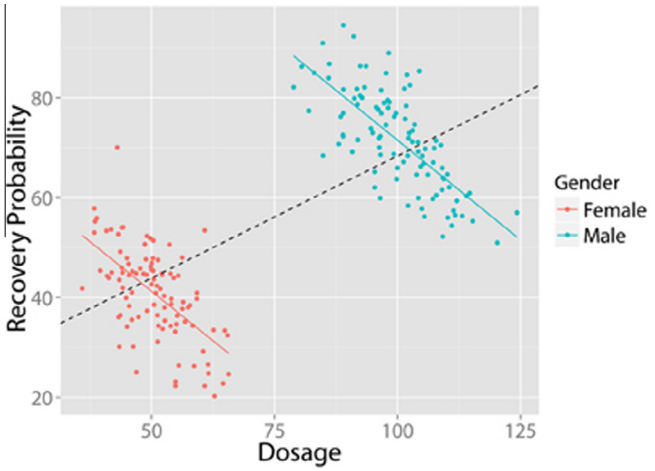



**Fig. 3.** Simpson paradox illustration where population and subgroups give contradictory results when analyzing the association between medical dosage and recovery in males and females (from [49]).

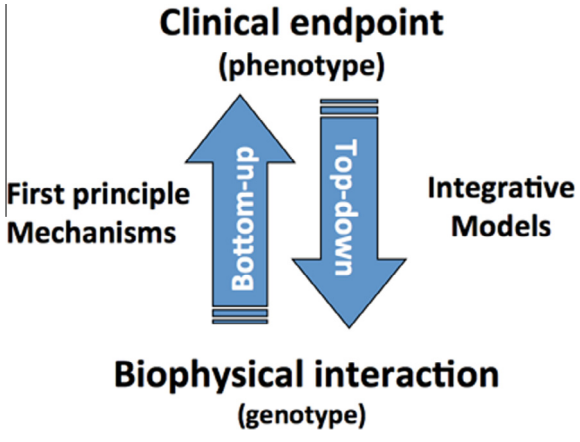



**Fig. 4.** Outcomes modeling schemes in oncology could be divided into: *top-down* (starting from the observed clinical outcome and attempting to identify the relevant variables that could explain the phenomena) or *bottom-up* (starting from basic principles to the observed clinical outcome in a multi-scale fashion).

### 3.1. Bottom-up approaches for modeling oncology response

These approaches utilize first principles of physics, chemistry and biology to model cellular damage temporally and spatially in response to treatment. Typically, they would apply advanced numerical methods such as Monte-Carlo (MC) techniques to estimate the molecular spectrum of damage in clustered and not-clustered DNA lesions ($Gbp^{-1}\,Gy^{-1}$) [50]. For instance in the case of radiotherapy, the temporal and spatial evolution of the effects from ionizing radiation can be divided into three phases: physical, chemical, and biological in a multi-scale fashion [51]. This information, however, could be used to guide incorporating prior knowledge or imposing constraints on a datamining approach narrowing its search space for optimal answers.

### 3.2. Top-down approaches for modeling oncology response

These are typically phenomenological (non-mechanistic) models and depend on parameters available from the collected clinical, dosimetric and/or biological data [15]. In the context of data-driven and multi-variable modeling of outcomes, the observed treatment outcome is considered as the result of functional mapping of several input variables [52]. Mathematically, this is expressed as $f(\mathbf{x}; \mathbf{w}^*)$: $X \rightarrow Y$ where $x_i \in R^N$ is composed of the input metrics (patient disease-specific prognostic factors, dosimetric metrics or biological markers). The expression $y_i \in Y$ is the corresponding observed treatment outcome. The variable $\mathbf{w}^*$ includes the optimal parameters of the model $f(\cdot)$ obtained by learning a certain objective functional. Learning is defined in this context of outcome modeling as estimating dependencies from data [53]. Based on the human-machine interaction, there is two common types of learning: supervised and unsupervised. Supervised learning is used when the endpoints of the treatments such as tumor control or toxicity grades are known; these endpoints are provided by experienced oncologists following institutional or National Cancer Institute (NCI) criteria and it is the most commonly used learning method in outcomes modeling. Nevertheless, unsupervised methods such as clustering methods or the use of principal component analysis (PCA) as a mean to reduce the learning problem dimensionality, feature extraction, and to aid in the visualization of multivariate data and the selection of the optimal learning method parameters for supervised learning methods [54].

It is noted that the selection of the functional form of the model $f(\cdot)$ is closely related to the prior knowledge of the problem. In mechanistic models, the shape of the functional form is selected based on the clinical or biological process at hand, however, in data-driven models; the objective is usually to find a functional form that best fits the data [55]. Below we will highlight this approach using logistic regression and artificial intelligence methods in cases where the clinical endpoints is expressed as a binary dichotomy (failed/did not fail) as commonly practiced. However, the methods could be extended in cases with more than two classes or the endpoint is a continuous variable.

#### 3.2.1. Logistic regression

In oncology outcomes modeling, the response will usually follow an S-shaped curve. This suggests that models with sigmoidal shapes are the most appropriate to use [9,10,12,14,56–59]. A commonly used sigmoidal form is the logistic regression model, which also has nice numerical stability properties. The results of this type of approach are expressed in the model parameters, which are chosen in a stepwise fashion to define the abscissa of the regression model $f(\cdot)$. However, it is the user's responsibility to determine whether interaction terms or higher order variables should be added. Penalty techniques based on ridge (L2-norm) or Lasso (L1-norm) methods could aid in the process by eliminating least relevant variables and imposing sparsity conditions [60]. An alternative solution to ameliorate this problem is offered by applying machine learning methods.

#### 3.2.2. Machine learning methods

Machine learning techniques are class of artificial intelligence (e.g., neural networks, decision trees, support vector machines), which are able to emulate human intelligence by learning the surrounding environment from the given input data and can detect nonlinear complex patterns in such data. In particular, neural networks were extensively investigated to model post-radiation treatment outcomes for cases of lung injury [61,62] and biochemical failure and rectal bleeding in prostate cancer [63,64]. A rather more robust approach of machine learning methods is support vector machines (SVMs), which are universal constructive learning procedures based on the statistical learning theory [65]. For discrimination between patients who are at low risk versus patients who are at high risk of treatment, the main idea of SVM would be to separate these two classes with 'hyper-planes' that maximize the margin between them in the nonlinear feature space defined by an implicit kernel mapping [66–68]. However, these methods have been stigmatized as black boxes, hindering their application in practical clinical contexts.

In an effort, to alleviate the black box stigma of generic machine learning methods and incorporate more system-like approaches methods based on graphical approaches such as Bayesian networks (BNs) have witnessed increased used in outcome modeling of cancer [69–71]. A BN provides graphical representation of the relationships between the variables represented as nodes in a directed acyclic graph (DAG), which encodes the presence and direction of relationship influence among the variables themselves and the clinical endpoint of interest. The relationship between parent and child nodes is modeled by conditional probabilities using Bayes chain rule. These methods are also robust for variable uncertainties and missing data, which would make them excellent candidates for clinical applications [72,73].

## 4. Results

Some of the issues encountered in the application of big data analytics could be caught by using visualization by clustering or by controlling the false discovery rate using multiple comparison adjustment as discussed earlier, however, many culprits remain at-large. Here, we present examples to overcome these problems drawn from radiation oncology by invoking small thinking to dwarf big data challenges.

### 4.1. Incorporating prior knowledge

Problems that suffer from the curse of high dimensionality [74] could be regularized by incorporating prior knowledge into the problem design. This is typically the case in situations where $p \gg n$ or nonlinear mapping approaches into higher dimensions are employed such as commonly the case in machine learning methods and kernel-based approaches. In the following, we present two examples from the problems of radio-proteomics and radio-genomics.

#### 4.1.1. Graph-based approach for identifying robust biomarkers from proteomics

To identify robust biomarkers to predict a treatment limiting side effect of thoracic irradiation injury known as radiation pneumonitis (RP) in lung cancer patients who received radiotherapy as part of their treatment, mass spectrometry was performed on peripheral blood samples from a longitudinal $3 \times 3$

matched-control cohort. To compensate for the large number variables to samples, a graph-based scoring function was developed to rank and identify the most robust biomarkers [69]. The proposed method measures the proximity between candidate proteins identified by mass spectrometry analysis utilizing prior reported knowledge in the literature of known markers of RP as shown in

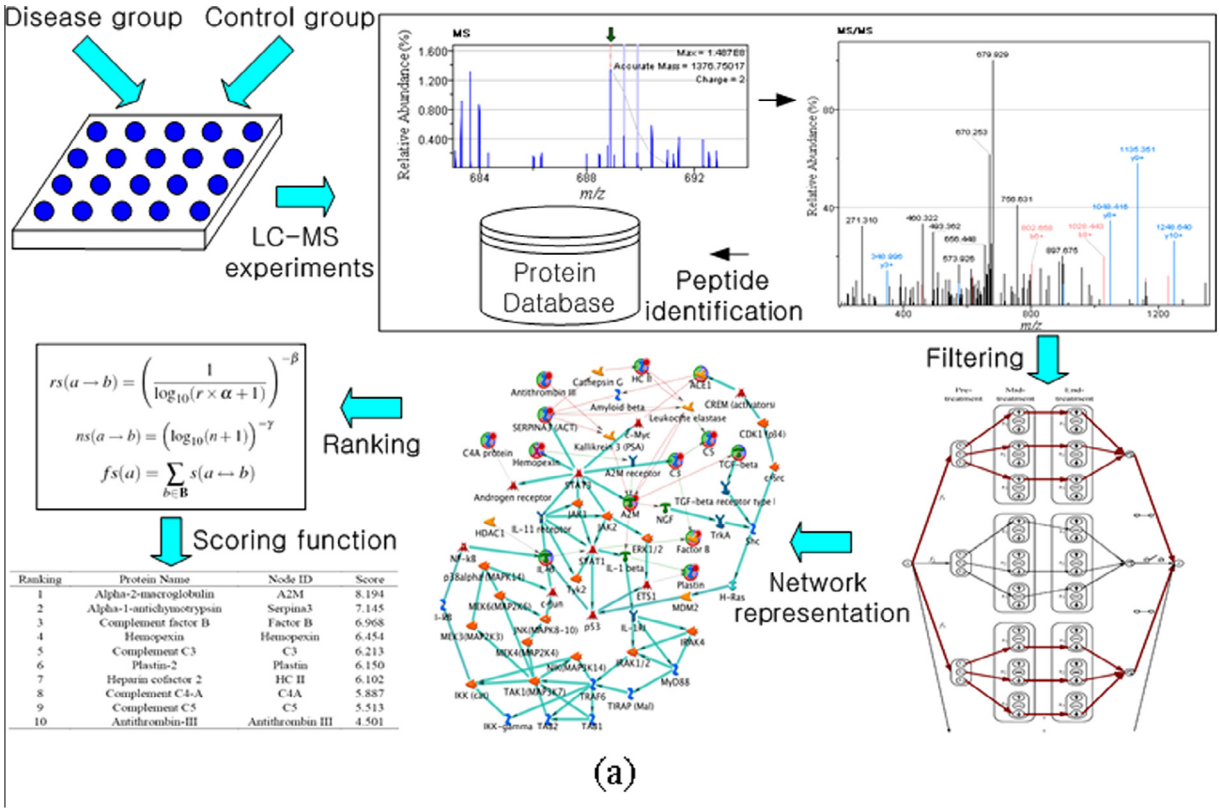



**Fig. 5.** Incorporation of prior knowledge in proteomics analysis. (a) Graph-based proteomics analysis to incorporate prior knowledge in which mass spectrometry data is analyzed in conjunction with known biomarkers of RP using filtering and network analysis. The approach identified α2M as the top ranked candidate. (b) Independent validation using ELISA analysis. Interestingly, it is noted that α2M acts as a radioprotector (higher expression leads to less incidences of RP) and as a biomarker (Patients likely to develop RP experience large increase during therapy).

Please cite this article in press as: I. El Naqa, Methods (2016), http://dx.doi.org/10.1016/j.ymeth.2016.08.010

Fig. 5a. The α-2-macroglobulin (α2M) protein was ranked as the top candidate biomarker. As an independent validation of this candidate protein, an enzyme-linked immunosorbent assay (ELISA) was performed on independent cohort of 20 patients' samples resulting in early significant discrimination between RP and non-RP patients ($p = 0.002$) as shown in Fig. 5b.

*4.1.2. Genetic maps for identifying candidate copy number variations*

Different characteristics of the genetic code influence response to cancer treatment. The most frequently studied type of genetic variations in outcome modeling is single-nucleotide polymorphisms (SNPs). SNPs may reflect mis-repair of previously acquired base damage and are thought to hinder functional gene products, for example, by inducing conformational changes in a protein reducing its catalytic activity. In addition to SNPs, copy number variations (CNVs) have been just recently incorporated into outcome models of radiotherapy toxicity [75]. CNVs are large-scale structural changes and may reflect the number of copies of genes contained within a given genome. Depending on the gene, CNVs can span orders of magnitude more nucleotides than SNPs and are therefore thought to play an increased role in influencing response to treatment. Given the rich literature on SNPs and their association with outcomes, neighboring CNVs particularly in SNP-rich areas could be selected using genetic maps (Fig. 6a). Then, so called radiogenomics model building process that involves other clinical and dosimetric metrics using regression and statistical resampling could be invoked. For instance, a radiogenomics model of rectal bleeding (RB) in prostate cancer was constructed using a CNV of XRCC1 identified from genetics mapping of reported SNPs mixed with dosimetric variables in logistic regression framework (Fig. 6b) [75,76]. Principle component analysis (PCA) could be used for visualization of the model. It is noted from the PCA plot (Fig. 6c) that the dosimetric metrics and the genetic variable have orthogonal relationship suggesting an interesting complementary effect.

*4.2. Information theory approach*

We will demonstrate the process of resolving the p-omics problem in the case of radiomics. The extraction of quantitative information from imaging modalities and relating information to biological and clinical endpoints is a new emerging field referred to as 'radiomics' as mentioned earlier [27,28,30]. Radiomics could be thought of as consisting of two main steps: (1) extraction of relevant static and dynamic imaging features and (2) incorporating these features into mathematical models to predict outcomes [7]. This process of feature extraction could involve hundreds to thousands of imaging features with different parameterizations, however, the number of patients remain limited. To resolve this problem, methods based on information theory could be utilized.

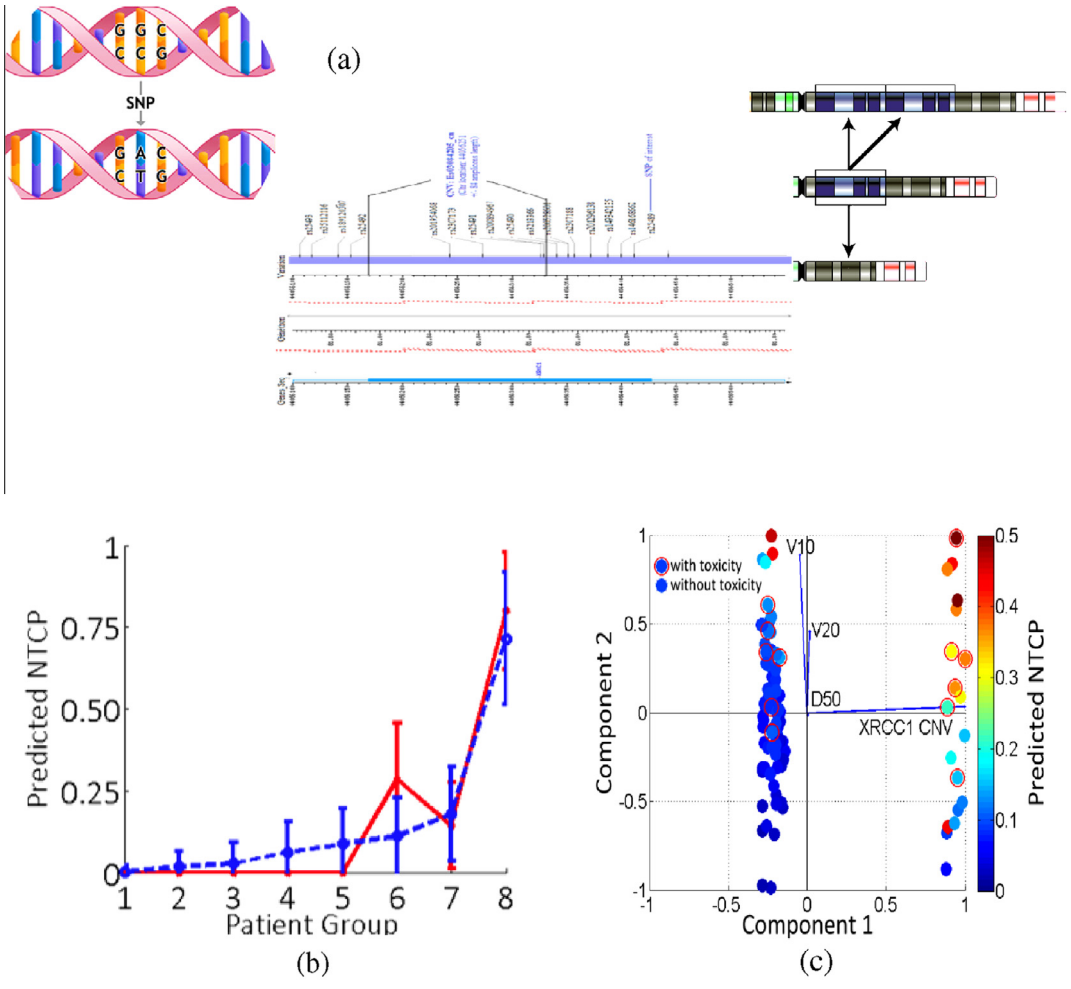
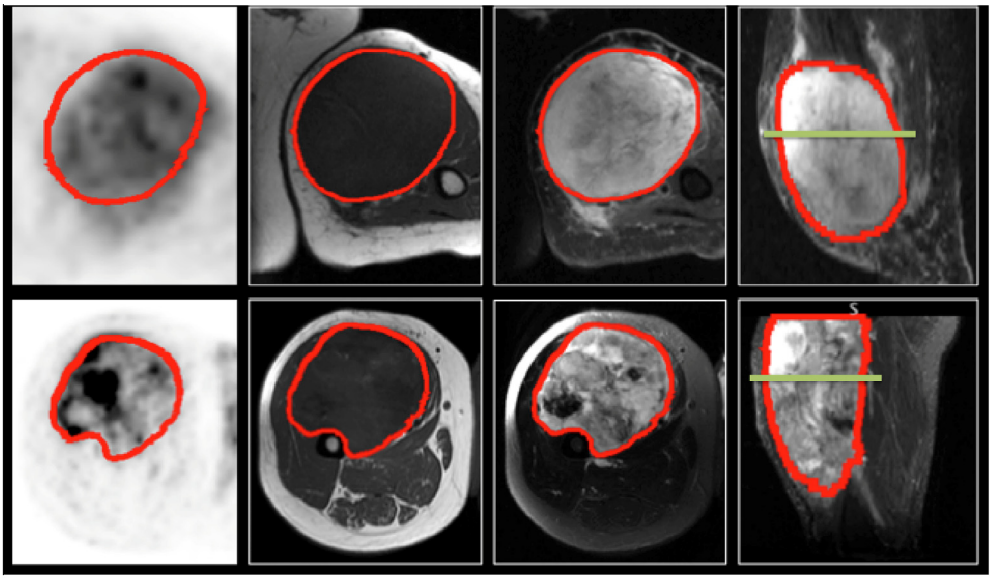
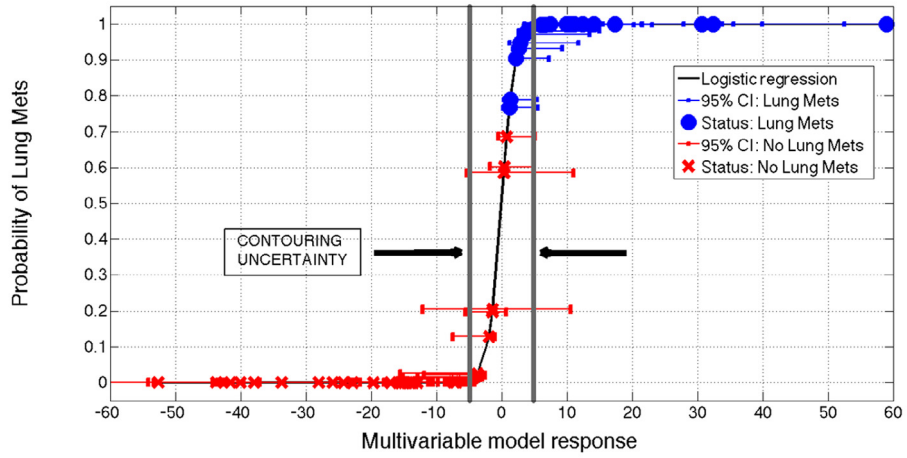



**Fig. 6.** Radiogenomics modeling of rectal bleeding (RB) post-radiotherapy in prostate cancer. (a) Selection of candidate CNVs by proximity of known SNPs in the literature of RB. (b) Logistic regression model (blue) of RB incidences (red). (c) PCA visualization of the RB model. Note orthogonal relationship between the CNV and dosimetric variables.



(a)




(b)

**Fig. 7.** (a) FDG-PET and MR diagnostic images of two patients with soft-tissue sarcomas of the extremities. Top row: patient that did not develop lung metastases. Bottom row: patient that eventually developed lung metastases. Fro left to right: FDG-PET, MR T1w, T2FS, and STIR (sagittal). The lines in the images of the 4th column correspond to the plane shown in the previous images were taken. (b) Probability of developing lung metastases as a function of the response of the final multivariable model identified in this work, for all patients of the retrospective cohort [35].

For instance, a dataset of 51 patients with histologically proven STS was retrospectively analyzed. All patients had pre-treatment FDG-PET and MR scans. MR data comprised of: T1-weighted (T1w), T2 fat-saturated (T2FS) and T2 short tau inversion recovery (STIR) sequences as shown in Fig. 7a.

A volume fusion process was carried out to combine information from two different volumes (PET and MR) into a single composite volume that is potentially more informative for texture feature analysis. Fusion of the scans was performed using the discrete wavelet transform (DWT) and a band-pass frequencies enhancement technique [35]. In total, 41 different features were extracted out of the tumor regions of 5 different types of scans: FDG-PET, T1w and T2FS, fused FDG-PET/T1 and fused FDG-PET/T2FS scans. The features were primarily texture (heterogeneity) features of 41 original metrics textures by 240 extraction parameter combinations yielding about ~10,000 total variables. The idea to select these features was based on defining the following information theoretic metric, where for any feature ($j$), its value in the model is defined by the following information gain functional ($IG_j$):

$$IG_j = \gamma \cdot |r_s(\mathbf{x}_j, \mathbf{y})| + \delta_a \cdot \left[ \sum_{k=1}^{f} \left( \frac{2(f-k+1)}{f(f+1)} \right) PIC(\mathbf{x}_k, \mathbf{x}_j) \right] + \delta_b \cdot \left[ \frac{1}{F} \sum_{l=1}^{F} PIC(\mathbf{x}_l, \mathbf{x}_j) \right],$$

where $rs$ is the Spearman's rank correlation coefficient, PIC is the potential information coefficient defined as PIC = 1 − MIC, MIC is the maximal information coefficient [77], and $\gamma$, $\delta_a$ and $\delta_b$ are predetermined parameters that balance association of the feature with the clinical endpoint (first term), association with current chosen features in the model (second term), and association with features that have not been chosen yet (third term).

The final optimal features were found using texture optimization based on imbalance-adjusted 0.632 + bootstrap resampling method [78]. The resulting model consisted of four texture features representing variations in size and intensity of the different tumor sub-regions. It yielded a performance estimate in bootstrapping evaluations, with an area under the receiver-operating

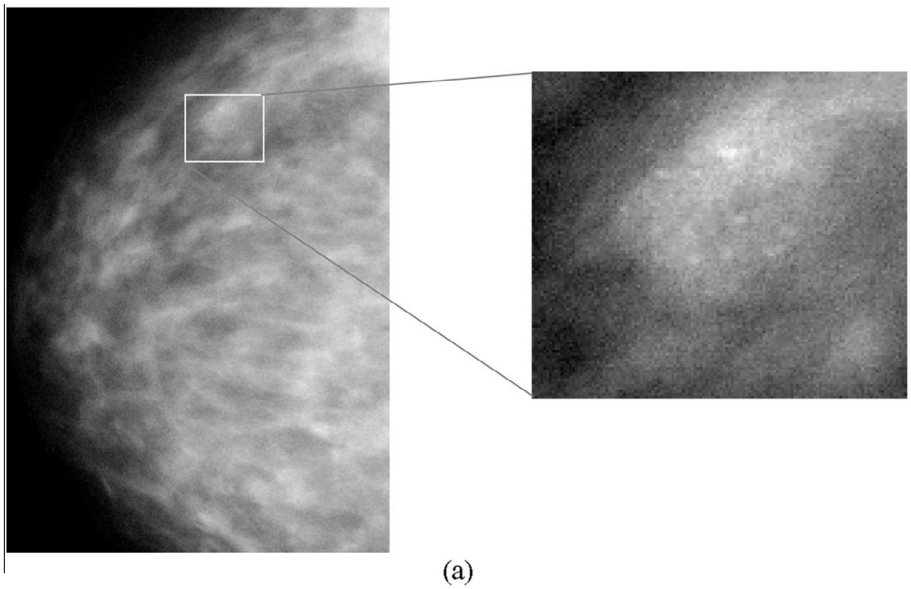
characteristic curve (AUC) of 0.984 ± 0.002, a sensitivity of 0.955 ± 0.006, a specificity of 0.926 ± 0.004 and an accuracy of 0.934 ± 0.003 as shown in Fig. 7b.

### 4.3. Ensemble of machine learning

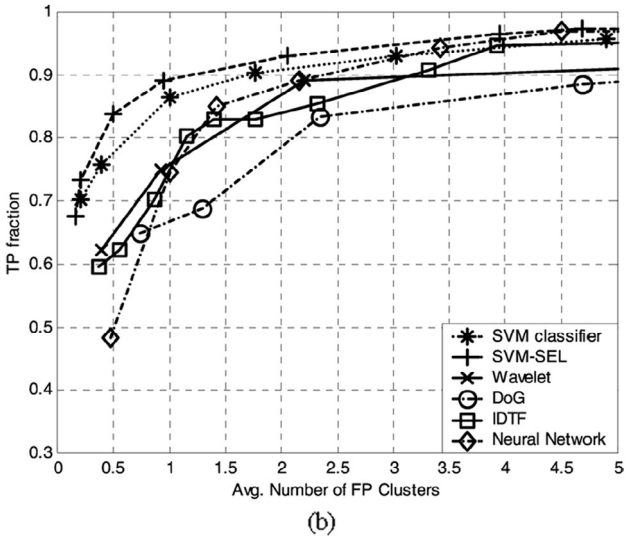
In order to avoid the limited learning ability of single classifiers and problems of data under- over-fitting pitfalls, meta-algorithms called ensemble machine learning algorithms were developed [54]. The main idea is to train a group of 'weak' learners with a given dataset and combine their output in order to compensate for increased variance of any individual learner. For instance, in building decision tree classifiers, a bag of trees is combined into so called a random forest [79]. After a bag of trees is trained, prediction is made for all the individual trees and the most frequent class selected by the trees is taken as the final result. Boosting is another ensemble meta-algorithm [80]. In this setting, the weak learners are trained sequentially such that incorrectly classified training examples (false positives or negatives) are re-assigned with larger weights and the subsequent classifier is learned with the reweighed training set. The final classification result is taken as an average output of the group of the classifiers. This technique is typically applied in the context of decision trees. Extension of this principle was demonstrated in the case of SVM using a successive enhancement learning (SEL) approach and applied in the detection of microcalcifications in mammogram images as shown in Fig. 8 [81].

### 4.4. Combined methods approach

A combination of the different previously described methods could be used to build complex system models, this is particularly of importance in the context of applying systems biology to outcome modeling. For instance, an ensemble of Bayesian networks (BNs) was developed for predicting radiation pneumonitis (RP) as shown in Fig. 9a [70]. Fifty-four NSCLC patients who received curative 3D-conformal radiotherapy were used to train a BN. Serum concentration of the following four candidate biomarkers were measured at baseline and mid-treatment: alpha-2-macroglobulin (α2M), angiotensin converting enzyme (ACE), transforming growth factor (TGF-1β), interleukin-6 (IL-6). Dose-volumetric and clinical parameters were also included as covariates. Feature selection



(a)




(b)

**Fig. 8.** (a) *Left* Mammogram in craniocaudal view. *Right* Expanded view showing MCs. (b) FROC curves of different microcalcification detection methods. The best performance was obtained by a successive learning SVM classifier, which achieves around 90% detection rate at a cost of one false positive (RP) cluster per image.

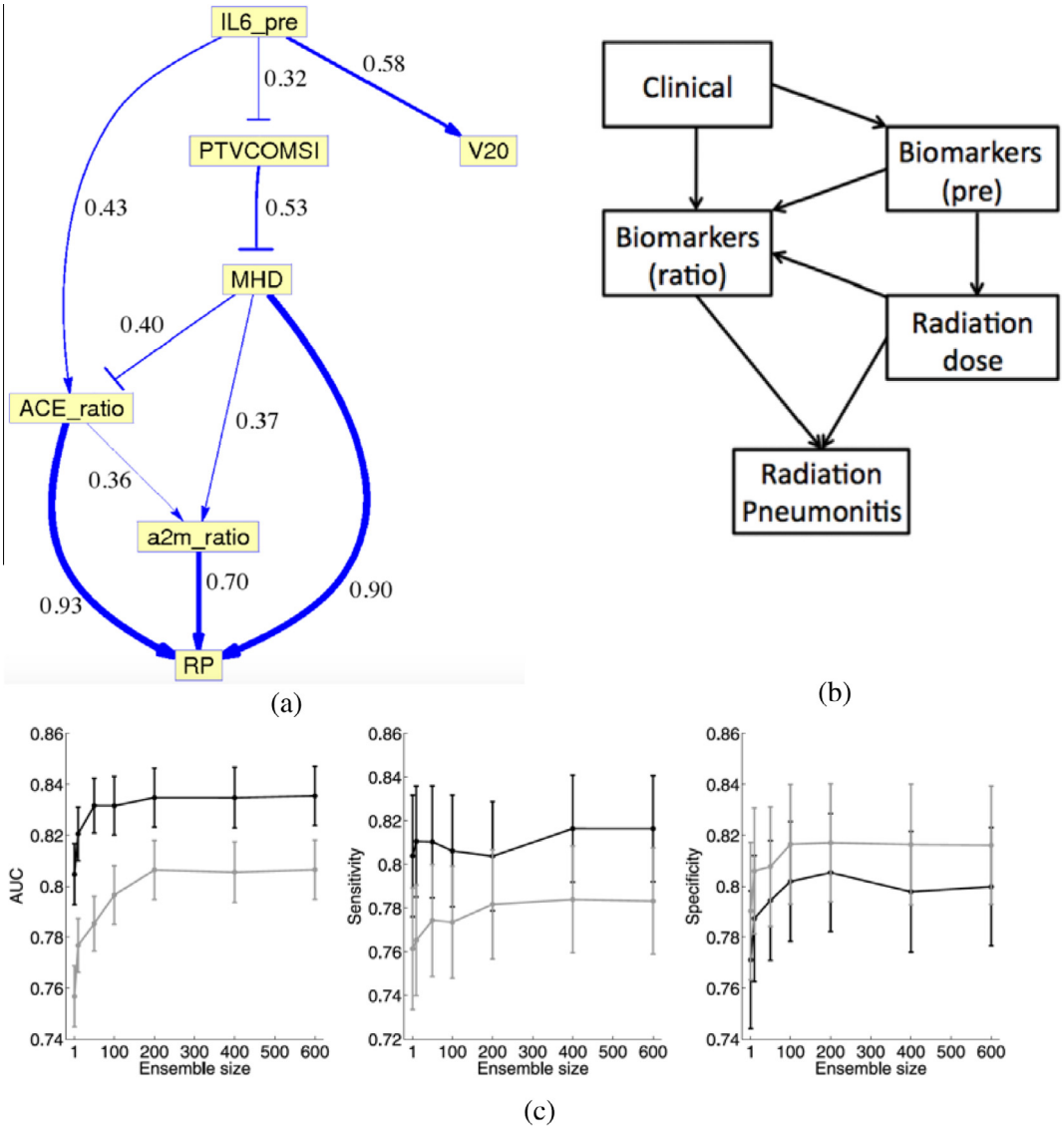



**Fig. 9.** (a) Variables connected by directed edges with a confidence level higher than random. Edge thickness is proportional to its confidence level. Arrow-headed and bar-headed edges are assigned to positive and negative correlations, respectively. (b) Diagram of allowed causal links between variable categories used for accepting/rejecting graph samples during MCMC simulation. (c) ROC metrics using the Bayesian network model ensemble with a varying sizes. Black: prediction with a complete dataset, gray: prediction without intra-treatment biomarker measurements. Error bars: bootstrap-estimated 95% confidence intervals.

was performed using a Markov blanket approach based on the Koller-Sahami filter. The Markov Chain Monte Carlo (MCMC) technique estimated the posterior distribution of BN graphs built from the observed data of the selected variables with prior causality constraints based on known knowledge of some of these variables biophysical interactions as shown in Fig. 9b. A resampling method based on bootstrapping was applied to model training and validation in order to control under- and over fitting pitfalls. RP prediction power of the BN ensemble approach reached its optimum at a size of 200. The optimized performance of the BN model recorded an area under the ROC curve (AUC) of 0.83 as shown in Fig. 9c, which was significantly higher than multivariate logistic regression (0.77).

## 5. Discussion

In the era of personalized medicine, oncology, with its multimodality multidisciplinary approach provides a unique combination of clinical, physical, technological, and biological data that could be evaluated as an ideal case study for employing big data analytics to improve treatment effectiveness and outcomes in medicine. Oncology data is comprised of clinical patient characteristics, varying imaging acquisitions, laboratory and biochemical measurements, etc. carrying all the hallmarks of big data. It is believed that Big data analytics hold great promise to improve safe treatment delivery and enable development of better clinical decision support systems for personalized medicine as lauded by the NIH Personalized Medicine Initiative (PMI). Furthermore, Big data analytics has been highlighted in the American Society of Clinical Oncology (ASCO) progress report as one of the promising opportunity in the fight against cancer as envisioned in the development of its data aggregation portal known as CancerLinQ [82]. The same sentiment has been echoed in Radiation Oncology [83].

The path for data collection and aggregation in oncology has been traditionally to develop a hypothesis based on a clinical or experimental observation then test this hypothesis in a controlled clinical trial institutionally, then multi-institutionally if it deemed promising. This path generally can account for about 5% of all patients' data with 95% of the clinical data, termed "dark data," is generated and stored during regular clinical processes. However,

this dark data is generally unstructured, untrusted, and fails to be useful for improving research, quality assessment, or clinical care. It is this invisible data that oncology Big Data initiatives such as CancerLinQ aim to bring to light. However, to make such data visible would require both cultural changes that would respect standardized lexicons and proper curation of this data on a routine basis. This would necessitate procedures that facilitate the data aggregation process, and local and national data champions within the oncology community. Moreover, making this data more visible would also need collaboration between all stakeholders to develop infrastructures and rigorous procedures to maintain its security and eliminate lingering patient privacy concerns.

New database technologies such as NoSQL or NewSQL whether terrestrial or in the cloud will yield better storage and query of oncology data while allowing application of more advanced analytics in real-time as part of a clinical quality assurance or improvement program. The MapReduce framework allows embedding of machine learning algorithms as part of its architecture. This technology would work well with parallelizable algorithms. However, many oncology modeling schemes particularly ones that involve iterative or gradient descent optimization techniques do not lend themselves naturally to this framework. This would require further investigation to overcome this limitation and to exploit such technologies for oncology real-time analytics.

One of main challenge to big data analytics in oncology remains the inherent p-omics versus pan-omics problem. In the presented examples using primarily typical applications in oncology, we demonstrated different methods to mitigate this effect such as using prior knowledge, information theory techniques, ensemble of machine learning, or different combinations of all these methods. Issues related to echo chamber or Yule-Simpson paradox need also to be carefully tested in the context of big data in oncology. However, the role of big data and its challenges is expected to grow as more current dark data being brought into light with many missing or poorly curated information and the pool of applications is ever expanding. This problem is further exacerbated when dealing with multiple clinical endpoints each may lead to different relationships with the input data. Moreover, despite decades of research many issues in dealing with multiple clinical or biological endpoints remain open [84]. The typical practice in oncology has been to optimize each point independently or to use heuristics to combine multiple endpoints in utility functions in order to account for competing risk effects and o quantify their subjective desirabilities [85]. Alternatively, such utilities could be presented as a multi-output system that would jointly optimize the prediction of the competing endpoints, of course, on the expense of increased sample size requirements posing further challenges to big data in oncology. Therefore, it is paramount to develop oncology-specific approaches that exploit bottom-up biological knowledge in cancer combined with advanced information theoretic and machine-learning methodologies to mitigate current challenges of noisy analytic pitfalls and achieve the big data promise in cancer research.

## 6. Conclusions

Big data hold great promise for oncology research and clinical care. There are several challenges related to data aggregation and analytics. The oncology data is largely unstructured and untabulated, however, emerging NoSQL and NewSQL database technologies will help improve aggregation and retrieval compared to classical methods. The process of combining data suffers from statistical limitations that could be addressed by applying prior knowledge and utilization of advances in machine learning techniques. The fulfilling of the big data dream in oncology would require collaboration between all stakeholders (clinicians, statisticians, bioinformaticians, physicists and biologists) to build necessary infrastructures to maintain data integrity and eliminate patient privacy concerns while continue to develop oncology specific computational approaches to leverage big data and potentially making computer-aided personalized clinical decision-making a practical reality.

## Acknowledgements

This work was by part supported by NIH P01 CA59827 and the University of Michigan Cancer Center fund G017459.

## References

[1] W.H. Organization, World Cancer Report 2014, 2014.
[2] D. Hanahan, Robert A. Weinberg, Hallmarks of cancer: the next generation, Cell 144 (5) (2011) 646–674.
[3] A. Bill-Axelson, L. Holmberg, M. Ruutu, M. Häggman, S.-O. Andersson, S. Bratell, A. Spångberg, C. Busch, S. Nordling, H. Garmo, J. Palmgren, H.-O. Adami, B.J. Norlén, J.-E. Johansson, Radical prostatectomy versus watchful waiting in early prostate cancer, N. Engl. J. Med. 352 (19) (2005) 1977–1984.
[4] J. Evans, S. Ziebland, A.R. Pettitt, Incurable, invisible and inconclusive: watchful waiting for chronic lymphocytic leukaemia and implications for doctor-patient communication, Eur. J. Cancer Care (Engl) 21 (1) (2012) 67–77.
[5] E.C. Halperin, C.A. Perez, L.W. Brady, Perez and Brady's Principles and Practice of Radiation Oncology, fifth ed., Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, 2008.
[6] I. El Naqa, Biomedical informatics and panomics for evidence-based radiation therapy, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 4 (4) (2014) 327–340.
[7] I. El Naqa, The role of big data in radiation oncology: challenges and potentials, in: B. Wang, R. Li, W. Perrizo (Eds.), Big Data Analytics in Bioinformatics and Healthcare, IGI Global, Hershey, PA, 2014, pp. 163–185.
[8] C. Shivade, P. Raghavan, E. Fosler-Lussier, P.J. Embi, N. Elhadad, S.B. Johnson, A. M. Lai, A review of approaches to identifying patient phenotype cohorts using electronic health records, J. Am. Med. Inform. Assoc. (2013).
[9] A.I. Blanco, K.S. Chao, I. El Naqa, G.E. Franklin, K. Zakarian, M. Vicic, J.O. Deasy, Dose-volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy, Int. J. Radiat. Oncol. Biol. Phys. 62 (4) (2005) 1055–1069.
[10] J. Bradley, J.O. Deasy, S. Bentzen, I. El-Naqa, Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma, Int. J. Radiat. Oncol. Biol. Phys. 58 (4) (2004) 1106–1113.
[11] A.J. Hope, P.E. Lindsay, I. El Naqa, J.R. Alaly, M. Vicic, J.D. Bradley, J.O. Deasy, Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters, Int. J. Radiat. Oncol. Biol. Phys. 65 (1) (2006) 112–124.
[12] A.J. Hope, P.E. Lindsay, I. El Naqa, J.D. Bradley, M. Vicic, J.O. Deasy, Clinical, dosimetric, and location-related factors to predict local control in non-small cell lung cancer, in: ASTRO 47th Annual Meeting, Denver, CO, 2005, p. S231.
[13] S. Levegrun, A. Jackson, M.J. Zelefsky, M.W. Skwarchuk, E.S. Venkatraman, W. Schlegel, Z. Fuks, S.A. Leibel, C.C. Ling, Fitting tumor control probability models to biopsy outcome after three-dimensional conformal radiation therapy of prostate cancer: pitfalls in deducing radiobiologic parameters for tumors from clinical data, Int. J. Radiat. Oncol. Biol. Phys. 51 (4) (2001) 1064–1080.
[14] L.B. Marks, Dosimetric predictors of radiation-induced lung injury, Int. J. Radiat. Oncol. Biol. Phys. 54 (2) (2002) 313–316.
[15] J.O. Deasy, I. El Naqa, Image-based modeling of normal tissue complication probability for radiation therapy, Cancer Treat. Res. 139 (2008) 215–256.
[16] I. El Naqa, G. Suneja, P.E. Lindsay, A.J. Hope, J.R. Alaly, M. Vicic, J.D. Bradley, A. Apte, J.O. Deasy, Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships, Phys. Med. Biol. 51 (22) (2006) 5719–5735.
[17] B.A. Chabner, T.G. Roberts, Chemotherapy and the war on cancer, Nat. Rev. Cancer 5 (1) (2005) 65–72.
[18] T. Khoukaz, Administration of anti-EGFR therapy: a practical review, Semin. Oncol. Nurs. 22 (2006) 20–27.
[19] T.Y. Seiwert, J.K. Salama, E.E. Vokes, The concurrent chemoradiation paradigm—general principles, Nat. Clin. Pract. Oncol. 4 (2) (2007) 86–100.
[20] P.Y. Wen, D.R. Macdonald, D.A. Reardon, T.F. Cloughesy, A.G. Sorensen, E. Galanis, J. DeGroot, W. Wick, M.R. Gilbert, A.B. Lassman, C. Tsien, T. Mikkelsen, E.T. Wong, M.C. Chamberlain, R. Stupp, K.R. Lamborn, M.A. Vogelbaum, M.J. van den Bent, S.M. Chang, Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group, J. Clin. Oncol. 28 (11) (2010) 1963–1972.
[21] J. Condeelis, R. Weissleder, In vivo imaging in cancer, Cold Spring Harb. Perspect. Biol. 2 (12) (2010) a003848.
[22] J.K. Willmann, N. van Bruggen, L.M. Dinkelborg, S.S. Gambhir, Molecular imaging in drug development, Nat. Rev. Drug Discov. 7 (7) (2008) 591–607.
[23] J. Bussink, J.H.A.M. Kaanders, W.T.A. van der Graaf, W.J.G. Oyen, PET-CT for radiotherapy treatment planning and response monitoring in solid tumors, Nat. Rev. Clin. Oncol. 8 (4) (2011) 233–242.

[24] H. Zaidi, I. El Naqa, PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques, Eur. J. Nucl. Med. Mol. Imaging 37 (11) (2010) 2165–2187.
[25] K. Newbold, M. Partridge, G. Cook, S.A. Sohaib, E. Charles-Edwards, P. Rhys-Evans, K. Harrington, C. Nutting, Advanced imaging applied to radiotherapy planning in head and neck cancer: a clinical review, Br. J. Radiol. 79 (943) (2006) 554–561.
[26] D. Piet, K. Frederik De, V. Vincent, S. Sigrid, H. Robert, N. Sandra, Diffusion-weighted magnetic resonance imaging to evaluate major salivary gland function before and after radiotherapy, Int. J. Radiat. Oncol. Biol. Phys. (2008).
[27] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G. van Stiphout, P. Granton, C.M. Zegers, R. Gillies, R. Boellard, A. Dekker, H.J. Aerts, Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (4) (2012) 441–446.
[28] V. Kumar, Y. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, D.B. Goldgof, L.O. Hall, P. Lambin, Y. Balagurunathan, R.A. Gatenby, R.J. Gillies, Radiomics: the process and the challenges, Magn. Reson. Imaging 30 (9) (2012) 1234–1248.
[29] F. Tixier, C.C. Le Rest, M. Hatt, N. Albarghach, O. Pradier, J.P. Metges, L. Corcos, D. Visvikis, Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer, J. Nucl. Med. 52 (3) (2011) 369–378.
[30] I. El Naqa, P. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, W. Thorstad, J.O. Deasy, Exploring feature-based approaches in PET images for predicting cancer treatment outcomes, Pattern Recognit. 42 (6) (2009) 1162–1171.
[31] N.M. Cheng, Y.H. Fang, J.T. Chang, C.G. Huang, D.L. Tsan, S.H. Ng, H.M. Wang, C. Y. Lin, C.T. Liao, T.C. Yen, Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma, J. Nucl. Med. 54 (10) (2013) 1703–1709.
[32] E.A. Kidd, I. El Naqa, B.A. Siegel, F. Dehdashti, P.W. Grigsby, FDG-PET-based prognostic nomograms for locally advanced cervical cancer, Gynecol. Oncol. 127 (1) (2012) 136–140.
[33] M. Vaidya, K.M. Creach, J. Frye, F. Dehdashti, J.D. Bradley, I. El Naqa, Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer, Radiother. Oncol. 102 (2) (2012) 239–245.
[34] G.J. Cook, C. Yip, M. Siddique, V. Goh, S. Chicklore, A. Roy, P. Marsden, S. Ahmad, D. Landau, Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy?, J Nucl. Med. 54 (1) (2013) 19–26.
[35] M. Vallieres, C.R. Freeman, S.R. Skamene, I. El Naqa, A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities, Phys. Med. Biol. 60 (14) (2015) 5471–5496.
[36] B.D.W. Group, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework, Clin. Pharmacol. Ther. 69 (3) (2001) 89–95.
[37] I. El Naqa, J. Craft, J. Oh, J. Deasy, Biomarkers for early radiation response for adaptive radiation therapy, in: X.A. Li (Ed.), Adaptive Radiation Therapy, Taylor & Francis, Boca Baton, FL, 2011, pp. 53–68.
[38] N. The Cancer Genome Atlas, Comprehensive genomic characterization of head and neck squamous cell carcinomas, Nature 517 (7536) (2015) 576–582.
[39] A. Abeshouse, J. Ahn, R. Akbani, A. Ally, S. Amin, Christopher D. Andry, M. Annala, A. Aprikian, J. Armenia, A. Arora, J.T. Auman, M. Balasundaram, S. Balu, Christopher E. Barbieri, T. Bauer, Christopher C. Benz, A. Bergeron, R. Beroukhim, M. Berrios, A. Bivol, T. Bodenheimer, L. Boice, Moiz S. Bootwalla, R. Borges dos Reis, Paul C. Boutros, J. Bowen, R. Bowlby, J. Boyd, Robert K. Bradley, A. Breggia, F. Brimo, Christopher A. Bristow, D. Brooks, Bradley M. Broom, Alan H. Bryce, G. Bubley, E. Burks, Yaron S.N. Butterfield, M. Button, D. Canes, Carlos G. Carlotti, R. Carlsen, M. Carmel, Peter R. Carroll, Scott L. Carter, R. Cartun, Brett S. Carver, June M. Chan, Matthew T. Chang, Y. Chen, Andrew D. Cherniack, S. Chevalier, L. Chin, J. Cho, A. Chu, E. Chuah, S. Chudamani, K. Cibulskis, G. Ciriello, A. Clarke, Matthew R. Cooperberg, Niall M. Corcoran, Anthony J. Costello, J. Cowan, D. Crain, E. Curley, K. David, John A. Demchok, F. Demichelis, N. Dhalla, R. Dhir, A. Doueik, B. Drake, H. Dvinge, N. Dyakova, I. Felau, Martin L. Ferguson, S. Frazer, S. Freedland, Y. Fu, Stacey B. Gabriel, J. Gao, J. Gardner, Julie M. Gastier-Foster, N. Gehlenborg, M. Gerken, Mark B. Gerstein, G. Getz, Andrew K. Godwin, A. Gopalan, M. Graefen, K. Graim, T. Gribbin, R. Guin, M. Gupta, A. Hadjipanayis, S. Haider, L. Hamel, D.N. Hayes, David I. Heiman, J. Hess, Katherine A. Hoadley, Andrea H. Holbrook, Robert A. Holt, A. Holway, Christopher M. Hovens, Alan P. Hoyle, M. Huang, Carolyn M. Hutter, M. Ittmann, L. Iype, Stuart R. Jefferys, Corbin D. Jones, Steven J.M. Jones, H. Juhl, A. Kahles, Christopher J. Kane, K. Kasaian, M. Kerger, E. Khurana, J. Kim, Robert J. Klein, R. Kucherlapati, L. Lacombe, M. Ladanyi, Phillip H. Lai, Peter W. Laird, Eric S. Lander, M. Latour, Michael S. Lawrence, K. Lau, T. LeBien, D. Lee, S. Lee, K.-V. Lehmann, Kristen M. Leraas, I. Leshchiner, R. Leung, John A. Libertino, Tara M. Lichtenberg, P. Lin, W.M. Linehan, S. Ling, Scott M. Lippman, J. Liu, W. Liu, L. Lochovsky, M. Loda, C. Logothetis, L. Lolla, T. Longacre, Y. Lu, J. Luo, Y. Ma, Harshad S. Mahadeshwar, D. Mallery, A. Mariamidze, Marco A. Marra, M. Mayo, S. McCall, G. McKercher, S. Meng, A.-M. Mes-Masson, Maria J. Merino, M. Meyerson, Piotr A. Mieczkowski, Gordon B. Mills, Kenna R.M. Shaw, S. Minner, A. Moinzadeh, Richard A. Moore, S. Morris, C. Morrison, Lisle E. Mose, Andrew J. Mungall, Bradley A. Murray, Jerome B. Myers, R. Naresh, J. Nelson, Mark A. Nelson, Peter S. Nelson, Y. Newton, Michael S. Noble, H. Noushmehr, M. Nykter, A. Pantazi, M. Parfenov, Peter J. Park, Joel S. Parker, J. Paulauskis, R. Penny, Charles M. Perou, A. Piché, T. Pihl, Peter A. Pinto, D. Prandi, A. Protopopov, Nilsa C. Ramirez, A. Rao, W.K. Rathmell, G. Rätsch, X. Ren, Victor E. Reuter, Sheila M. Reynolds, Suhn K. Rhie, K. Rieger-Christ, J. Roach, A.G. Robertson, B. Robinson, Mark A. Rubin, F. Saad, S. Sadeghi, G. Saksena, C. Saller, A. Salner, F. Sanchez-Vega, C. Sander, G. Sandusky, G. Sauter, A. Sboner, Peter T. Scardino, E. Scarlata, Jacqueline E. Schein, T. Schlomm, Laura S. Schmidt, N. Schultz, Steven E. Schumacher, J. Seidman, L. Neder, S. Seth, A. Sharp, C. Shelton, T. Shelton, H. Shen, R. Shen, M. Sherman, M. Sheth, Y. Shi, J. Shih, I. Shmulevich, J. Simko, R. Simon, Janae V. Simons, P. Sipahimalani, T. Skelly, Heidi J. Sofia, Matthew G. Soloway, X. Song, A. Sorcini, C. Sougnez, S. Stepa, C. Stewart, J. Stewart, Joshua M. Stuart, Travis B. Sullivan, C. Sun, H. Sun, A. Tam, D. Tan, J. Tang, R. Tarnuzzer, K. Tarvin, Barry S. Taylor, P. Teebagy, I. Tenggara, B. Têtu, A. Tewari, N. Thiessen, T. Thompson, Leigh B. Thorne, Daniela P. Tirapelli, Scott A. Tomlins, Felipe A. Trevisan, P. Troncoso, Lawrence D. True, Maria C. Tsourlakis, S. Tyekucheva, E. Van Allen, David J. Van Den Berg, U. Veluvolu, R. Verhaak, Cathy D. Vocke, D. Voet, Y. Wan, Q. Wang, W. Wang, Z. Wang, N. Weinhold, John N. Weinstein, Daniel J. Weisenberger, Matthew D. Wilkerson, L. Wise, J. Witte, C.-C. Wu, J. Wu, Y. Wu, Andrew W. Xu, Shalini S. Yadav, L. Yang, L. Yang, C. Yau, H. Ye, P. Yena, T. Zeng, Jean C. Zenklusen, H. Zhang, J. Zhang, J. Zhang, W. Zhang, Y. Zhong, K. Zhu, E. Zmuda, The molecular taxonomy of primary prostate cancer, Cell 163 (4) (2015) 1011–1025.
[40] N. The Cancer Genome Atlas Research, Comprehensive molecular profiling of lung adenocarcinoma, Nature 511 (7511) (2014) 543–550.
[41] Cameron W. Brennan, Roel G.W. Verhaak, A. McKenna, B. Campos, H. Noushmehr, Sofie R. Salama, S. Zheng, D. Chakravarty, J.Z. Sanborn, Samuel H. Berman, R. Beroukhim, B. Bernard, C.-J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, Sachet A. Shukla, G. Ciriello, W.K. Yung, W. Zhang, C. Sougnez, T. Mikkelsen, K. Aldape, Darell D. Bigner, Erwin G. Van Meir, M. Prados, A. Sloan, Keith L. Black, J. Eschbacher, G. Finocchiaro, W. Friedman, David W. Andrews, A. Guha, M. Iacocca, Brian P. O'Neill, G. Foltz, J. Myers, Daniel J. Weisenberger, R. Penny, R. Kucherlapati, Charles M. Perou, D.N. Hayes, R. Gibbs, M. Marra, Gordon B. Mills, E. Lander, P. Spellman, R. Wilson, C. Sander, J. Weinstein, M. Meyerson, S. Gabriel, Peter W. Laird, D. Haussler, G. Getz, L. Chin, C. Benz, J. Barnholtz-Sloan, W. Barrett, Q. Ostrom, Y. Wolinsky, Keith L. Black, B. Bose, Paul T. Boulos, M. Boulos, J. Brown, C. Czerinski, M. Eppley, M. Iacocca, T. Kempista, T. Kitko, Y. Koyfman, B. Rabeno, P. Rastogi, M. Sugarman, P. Swanson, K. Yalamanchii, Ilana P. Otey, Yingchun S. Liu, Y. Xiao, J.T. Auman, P.-C. Chen, A. Hadjipanayis, E. Lee, S. Lee, Peter J. Park, J. Seidman, L. Yang, R. Kucherlapati, S. Kalkanis, T. Mikkelsen, Laila M. Poisson, A. Raghunathan, L. Scarpace, B. Bernard, R. Bressler, A. Eakin, L. Iype, Richard B. Kreisberg, K. Leinonen, S. Reynolds, H. Rovira, V. Thorsson, I. Shmulevich, Matti J. Annala, R. Penny, J. Paulauskis, E. Curley, M. Hatfield, D. Mallery, S. Morris, T. Shelton, C. Shelton, M. Sherman, P. Yena, L. Cuppini, F. DiMeco, M. Eoli, G. Finocchiaro, E. Maderna, B. Pollo, M. Saini, S. Balu, Katherine A. Hoadley, L. Li, C.R. Miller, Y. Shi, Michael D. Topal, J. Wu, G. Dunn, C. Giannini, Brian P. O'Neill, B.A. Aksoy, Y. Antipin, L. Borsu, Samuel H. Berman, Cameron W. Brennan, E. Cerami, D. Chakravarty, G. Ciriello, J. Gao, B. Gross, A. Jacobsen, M. Ladanyi, A. Lash, Y. Liang, B. Reva, C. Sander, N. Schultz, R. Shen, Nicholas D. Socci, A. Viale, Martin L. Ferguson, Q.-R. Chen, John A. Demchok, Laura A.L. Dillon, Kenna R.M. Shaw, M. Sheth, R. Tarnuzzer, Z. Wang, L. Yang, T. Davidsen, Mark S. Guyer, Bradley A. Ozenberger, Heidi J. Sofia, J. Bergsten, J. Eckman, J. Harr, J. Myers, C. Smith, K. Tucker, C. Winemiller, Leigh A. Zach, Julia Y. Ljubimova, G. Eley, B. Ayala, Mark A. Jensen, A. Kahn, Todd D. Pihl, David A. Pot, Y. Wan, J. Eschbacher, G. Foltz, N. Hansen, P. Hothi, B. Lin, N. Shah, J.-g. Yoon, C. Lau, M. Berens, K. Ardlie, R. Beroukhim, Scott L. Carter, Andrew D. Cherniack, M. Noble, J. Cho, K. Cibulskis, D. DiCara, S. Frazer, Stacey B. Gabriel, N. Gehlenborg, J. Gentry, D. Heiman, J. Kim, R. Jing, Eric S. Lander, M. Lawrence, P. Lin, W. Mallard, M. Meyerson, Robert C. Onofrio, G. Saksena, S. Schumacher, C. Sougnez, P. Stojanov, B. Tabak, D. Voet, H. Zhang, L. Zou, G. Getz, Nathan N. Dees, L. Ding, Lucinda L. Fulton, Robert S. Fulton, K.-L. Kanchi, Elaine R. Mardis, Richard K. Wilson, Stephen B. Baylin, David W. Andrews, L. Harshyne, Mark L. Cohen, K. Devine, Andrew E. Sloan, Scott R. VandenBerg, Mitchel S. Berger, M. Prados, D. Carlin, B. Craft, K. Ellrott, M. Goldman, T. Goldstein, M. Grifford, D. Haussler, S. Ma, S. Ng, Sofie R. Salama, J.Z. Sanborn, J. Stuart, T. Swatloski, P. Waltman, J. Zhu, R. Foss, B. Frentzen, W. Friedman, R. McTiernan, A. Yachnis, D.N. Hayes, Charles M. Perou, S. Zheng, R. Vegesna, Y. Mao, R. Akbani, K. Aldape, O. Bogler, Gregory N. Fuller, W. Liu, Y. Liu, Y. Lu, G. Mills, A. Protopopov, X. Ren, Y. Sun, C.-J. Wu, W.K.A. Yung, W. Zhang, J. Zhang, K. Chen, John N. Weinstein, L. Chin, Roel G.W. Verhaak, H. Noushmehr, Daniel J. Weisenberger, Moiz S. Bootwalla, Phillip H. Lai, Timothy J. Triche Jr., David J. Van Den Berg, Peter W. Laird, David H. Gutmann, Norman L. Lehman, Erwin G. VanMeir, D. Brat, Jeffrey J. Olson, Gena M. Mastrogianakis, Narra S. Devi, Z. Zhang, D. Bigner, E. Lipp, R. McLendon, The somatic genomic landscape of glioblastoma, Cell 155 (2) (2013) 462–477.
[42] E.F. Codd, The Relational Model for Database Management: Version 2, Addison-Wesley, Reading, Mass., 1990.
[43] J. Pearl, M. Glymour, N.P. Jewell, Causal Inference in Statistics: A Primer, John Wiley & Sons, Chichester, UK; Hoboken, NJ, 2016.
[44] P. Lake, R. Drake, Information Systems Management in the Big Data Era, Springer Berlin Heidelberg, New York, NY, 2015.
[45] F. Andre, L.M. McShane, S. Michiels, D.F. Ransohoff, D.G. Altman, J.S. Reis-Filho, D.F. Hayes, L. Pusztai, Biomarker studies: a call for a comprehensive biomarker study registry, Nat. Rev. Clin. Oncol. 8 (3) (2011) 171–176.
[46] E.H. Simpson, The Interpretation of Interaction in Contingency Tables, J. R. Stat. Soc. Ser. B (Methodological) 13 (2) (1951) 238–241.
[47] C. Tomasetti, B. Vogelstein, Variation in cancer risk among tissues can be explained by the number of stem cell divisions, Science 347 (6217) (2015) 78–81.

Please cite this article in press as: I. El Naqa, Methods (2016), http://dx.doi.org/10.1016/j.ymeth.2016.08.010

[48] A. Albini, S. Cavuto, G. Apolone, D.M. Noonan, Strategies to prevent "Bad Luck" in cancer, J. Natl. Cancer Inst. 107 (10) (2015).
[49] R. Kievit, W.E. Frankenhuis, L. Waldorp, D. Borsboom, Simpson's paradox in psychological science: a practical guide, Front. Psychol. 4 (2013).
[50] H. Nikjoo, S. Uehara, D. Emfietzoglou, F.A. Cucinotta, Track-structure codes in radiation research, Radiat. Meas. 41 (9–10) (2006) 1052–1074.
[51] I. El Naqa, P. Pater, J. Seuntjens, Monte Carlo role in radiobiological modelling of radiotherapy outcomes, Phys. Med. Biol. 57 (11) (2012) R75–R97.
[52] I. El Naqa, J.D. Bradley, P.E. Lindsay, A.I. Blanco, M. Vicic, A.J. Hope, J.O. Deasy, Multi-variable modeling of radiotherapy outcomes including dose-volume and clinical factors, Int. J. Radiat. Oncol. Biol. Phys. 64 (4) (2006) 1275–1286.
[53] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations, Springer, New York, 2001.
[54] I. El Naqa, R. Li, M.J. Murphy, Machine Learning in Radiation Oncology: Theory and Application, Springer International Publishing, Switzerland, 2015.
[55] I. El Naqa, J. Bradley, P.E. Lindsay, A. Hope, J.O. Deasy, Predicting radiotherapy outcomes using statistical learning techniques, Phys. Med. Biol. 54 (2009) S9–S30.
[56] S.L. Tucker, R. Cheung, L. Dong, H.H. Liu, H.D. Thames, E.H. Huang, D. Kuban, R. Mohan, Dose-volume response analyses of late rectal bleeding after radiotherapy for prostate cancer, Int. J. Radiat. Oncol. Biol. Phys. 59 (2) (2004) 353–365.
[57] J.D. Bradley, A. Hope, I. El Naqa, A. Apte, P.E. Lindsay, W. Bosch, J. Matthews, W. Sause, M.V. Graham, J.O. Deasy, Rtog, A nomogram to predict radiation pneumonitis, derived from a combined analysis of RTOG 9311 and institutional data, Int. J. Radiat. Oncol. Biol. Phys. 69 (4) (2007) 985–992.
[58] E.X. Huang, J.D. Bradley, I.E. Naqa, A.J. Hope, P.E. Lindsay, W.R. Bosch, J.W. Matthews, W.T. Sause, M.V. Graham, J.O. Deasy, Modeling the risk of radiation-induced acute esophagitis for combined Washington University and RTOG trial 93-11 lung cancer patients, Int. J. Radiat. Oncol. Biol. Phys. (2011).
[59] E.X. Huang, A.J. Hope, P.E. Lindsay, M. Trovo, I. El Naqa, J.O. Deasy, J.D. Bradley, Heart irradiation as a risk factor for radiation pneumonitis, Acta Oncol. 50 (1) (2011) 51–60.
[60] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, CRC Press, Taylor & Francis Group, Boca Raton, 2015.
[61] M.T. Munley, J.Y. Lo, G.S. Sibley, G.C. Bentel, M.S. Anscher, L.B. Marks, A neural network to predict symptomatic lung injury, Phys. Med. Biol. 44 (1999) 2241–2249.
[62] M. Su, M. Miftena, C. Whiddon, X. Sun, K. Light, L. Marks, An artificial neural network for predicting the incidence of radiation pneumonitis, Med. Phys. 32 (2) (2005) 318–325.
[63] 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 (1) (2004) 3–12.
[64] S. Tomatis, T. Rancati, C. Fiorino, V. Vavassori, G. Fellin, E. Cagna, F.A. Mauro, G. Girelli, A. Monti, M. Baccolini, G. Naldi, C. Bianchi, L. Menegotti, M. Pasquino, M. Stasi, R. Valdagni, Late rectal bleeding after 3D-CRT for prostate cancer: development of a neural-network-based predictive model, Phys. Med. Biol. 57 (5) (2012) 1399.
[65] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[66] I. El Naqa, J.D. Bradley, P.E. Lindsay, A.J. Hope, J.O. Deasy, Predicting radiotherapy outcomes using statistical learning techniques, Phys. Med. Biol. 54 (18) (2009) S9–S30.
[67] I. El Naqa, J.O. Deasy, Y. Mu, E. Huang, A.J. Hope, P.E. Lindsay, A. Apte, J. Alaly, J. D. Bradley, Datamining approaches for modeling tumor control probability, Acta Oncol. 49 (8) (2010) 1363–1373.
[68] I. El Naqa, Machine learning methods for predicting tumor response in lung cancer, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2 (2) (2012) 173–181.
[69] J.H. Oh, J. Craft, R. Al Lozi, M. Vaidya, Y. Meng, J.O. Deasy, J.D. Bradley, I. El Naqa, A Bayesian network approach for modeling local failure in lung cancer, Phys. Med. Biol. 56 (6) (2011) 1635–1651.
[70] S. Lee, N. Ybarra, K. Jeyaseelan, S. Faria, N. Kopek, P. Brisebois, J.D. Bradley, C. Robinson, J. Seuntjens, I. El Naqa, Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk, Med. Phys. 42 (5) (2015) 2421–2430.
[71] K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin, A.L. Dekker, Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy, Med. Phys. 37 (4) (2010) 1401–1407.
[72] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, Cambridge, MA, 2009.
[73] C. Sinoquet, R.L. Mourad, Probabilistic Graphical Models for Genetics, Genomics and Postgenomics, first ed., Oxford University Press, Oxford, 2014.
[74] E. Keogh, A. Mueen, Curse of dimensionality, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning, Springer, US, Boston, MA, 2010, pp. 257–258.
[75] J. Coates, A.K. Jeyaseelan, N. Ybarra, M. David, S. Faria, L. Souhami, F. Cury, M. Duclos, I. El Naqa, Contrasting analytical and data-driven frameworks for radiogenomic modeling of normal tissue toxicities in prostate cancer, Radiother. Oncol. 115 (1) (2015) 107–113.
[76] J. Coates, I. El Naqa, Outcome modeling techniques for prostate cancer radiotherapy: data, models, and validation, Phys. Med. (2016).
[77] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518–1524.
[78] B. Sahiner, H.-P. Chan, L. Hadjiiski, Classifier performance prediction for computer-aided diagnosis using a limited dataset, Med. Phys. 35 (4) (2008) 1559–1570.
[79] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[80] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (2) (1990) 197–227.
[81] I. El-Naqa, Y. Yang, M.N. Wernick, N.P. Galatsanos, R.M. Nishikawa, A support vector machine approach for detection of microcalcifications, IEEE Trans. Med. Imaging 21 (12) (2002) 1552–1563.
[82] D.S. Dizon, L. Krilov, E. Cohen, T. Gangadhar, P.A. Ganz, T.A. Hensing, S. Hunger, S.S. Krishnamurthi, A.B. Lassman, M.J. Markham, E. Mayer, M. Neuss, S.K. Pal, L. C. Richardson, R. Schilsky, G.K. Schwartz, D.R. Spriggs, M.A. Villalona-Calero, G. Villani, G. Masters, Clinical cancer advances 2016: annual report on progress against cancer from the American Society of Clinical Oncology, J. Clin. Oncol. (2016).
[83] S.H. Benedict, I. El Naqa, E.E. Klein, Introduction to big data in radiation oncology: exploring opportunities for research, quality assessment, and clinical care, Int. J. Radiat. Oncol. Biol. Phys. 95 (3) (2016) 871–872.
[84] M. Gail, A review and critique of some models used in competing risk analysis, Biometrics 31 (1) (1975) 209–222.
[85] T.A. Murray, P.F. Thall, Y. Yuan, Utility-based designs for randomized comparative trials with categorical outcomes, Stat. Med. (2016).

Please cite this article in press as: I. El Naqa, Methods (2016), http://dx.doi.org/10.1016/j.ymeth.2016.08.010